Coarse-Grained Molecular Dynamics
Parametrizing Non-Bonded Interactions in Water via Force Matching

EYM-272: Modeling of Physical Systems and Machine Learning Methods

Vourvachakis S. Georgios (ID: mse354)[a]

[a]*Department of Materials Science and Engineering, University of Crete, 70013 Heraklion, Greece.*

March, 2025

Abstract

This project details the development and application of the Force Matching (FM) method for parametrizing non-bonded interactions in a coarse-grained (CG) model of liquid water, approximating the Potential of Mean Force with effective pairwise potentials represented by cubic B-splines. A comprehensive computational pipeline was implemented for a system of 1000 TIP3P water molecules, encompassing all-atom simulation, center-of-mass mapping, and Adam optimization of spline parameters against mapped atomistic forces from 10,000 configurations. Investigating two distinct initialization schemes (Lennard-Jones-like vs. SPC/E-inspired) and learning rates (10^{-3} vs. 10^{-2}), the FM optimization yielded final Mean Squared Error values of **89.41** and **1439.47**, respectively, demonstrating significant sensitivity to these choices. Subsequent CG simulations using the optimized potential qualitatively reproduced the liquid structure of water, as evidenced by Radial Distribution Function and Voronoi tessellation analyses, though quantitative deviations from atomistic references were observed. This work establishes a functional FM pipeline for *data-driven bottom-up CG potential development* and underscores the critical impact of optimization strategy on model fidelity.

1. Introduction

Molecular dynamics (MD) simulations have become an indispensable tool in understanding the structural and dynamical properties of molecular systems since the pioneering work of Alder and Wainwright in the late 1960s, who first demonstrated the long-time behavior of hard-sphere fluids through computer simulations [1]. Their early algorithmic advances laid the foundation for simulating systems of interacting particles, providing direct access to microscopic trajectories that complement experimental observations. Over the subsequent decades, improvements in potential energy functions, integration schemes, and computational power enabled the study of increasingly complex systems, from simple liquids to biomolecules, offering detailed insight into equilibrium and non-equilibrium phenomena [31],[28],[48],[18].

Despite these successes, full atomistic MD simulations remain computationally expensive for large systems or long timescales, motivating the development of coarse-grained (CG) models [44],[15]. In CG molecular dynamics, groups of atoms are represented by single interaction sites—or beads—thereby reducing the number of degrees of freedom and smoothing the potential energy landscape (see Fig.). This reduction allows exploration of larger length and time scales, albeit at the cost of fine structural detail. The CG methodology can be broadly categorized into bottom-up approaches, which derive effective interactions from all-atom reference data, and top-down approaches, which fit parameters to reproduce macroscopic observables [21],[46].

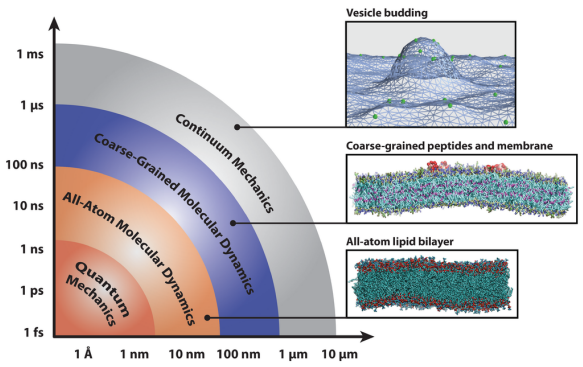


Figure 1: Diagram of computational methods for studying biophysical systems across a range of time- and length-scales. Representative snapshots depict an all-atom lipid bilayer, peptides embedded in a coarse-grained bilayer and proteins remodeling a continuum mechanics membrane model. Bilayers were simulated with the CHARMM36 [30] and Martini [41] force fields and rendered with Visual Molecular Dynamics [19].

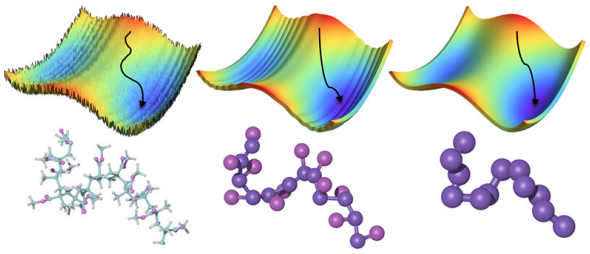


Figure 2: Schematic illustration of coarse-grained free energy surface and corresponding representative coarse-grained molecular models.[55]

1.1. Bottom-Up Coarse-Graining and Parametrization

Kalligiannaki and coworkers (2016) frame CG as a problem in dimensionality reduction, akin to well-established techniques in scientific computing and applied statistics, such as principal component analysis and diffusion maps [25]. They categorize bottom-up parametrization methods according to the chosen observables and minimization principles:

1. **Structure-based methods** minimize discrepancies in radial distribution functions, , or bonded degree-of-freedom distributions, using direct or iterative Boltzmann inversion [51],[27] and inverse Monte Carlo techniques [37],[38].

2. **Force matching (FM)**, also known as multiscale coarse-graining (MS-CG), formulates the parametrization as a least-squares problem matching the total forces on CG sites to reference all-atom forces [11].

3. **Relative entropy (RE)** methods minimize the Kullback–Leibler divergence [32] between atomistic and CG Gibbs measures, yielding models with broad applicability across observables via information-theoretic bounds [43],[53].

These bottom-up strategies approximate the potential of mean force (PMF) over CG degrees of freedom, either through numerical optimization or analytical theories rooted in liquid-state physics. Complementarily, top-down or reverse-mapping approaches seek to reconstruct atomistic detail from CG configurations (e.g., Fig.3), an area seeing growing synergy with machine learning techniques, such as graph neural networks and generative models, which enhance back-mapping accuracy [34],[61].

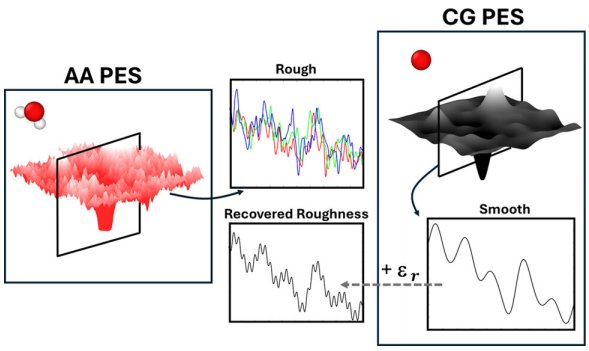


Figure 3: Visualization of an idealized schematic of a rugged all-atom landscape (top-left) and a smoothened CG landscape (top-right). The sliced potential energy surface at the bottom provides a qualitative representation of high-frequency "rough" features, which can be emulated by introducing perturbation to the CG potential.[45]

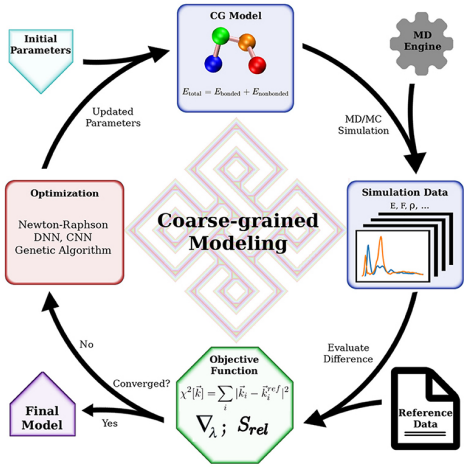


Figure 4: General workflow of coarse-grained modeling. A software that performs simulation, an objective function to evaluate simulation data and an optimization algorithm to improve the CG model. The final model, which will reproduce the reference data, is obtained when the objective function is satisfied. When the reference data is extracted from fine-grained model simulations, this flow chart represents a general bottom-up CG algorithm. When the reference data is macroscopic quantities obtained in experiments, this flow chart indicates a top-down CG procedure. Hybrid CG modeling is achieved with reference data from both lower-level models and experiments, with a corresponding objective function.[58]

In this study, we apply the Force Matching optimization method to derive an estimator for the coarse-grained PMF of liquid water. We construct an effective pair potential using a cubic B-spline basis set, mapping each water molecule (modeled with TIP3P [40]) to its center of mass and eliminating intramolecular degrees of freedom. We perform simulations of 1000 CG water beads to evaluate the

accuracy of our model in reproducing structural and dynamical properties.

First, we review atomistic and CG frameworks, detailing the Force Matching formalism and its interpretation as a projection onto coarse observable spaces.

Then, inspired by Kalligiannaki et al.'s workflow [25], we introduce efficient neighbor-search via the KD-tree data structure [14] and implement a data-size-aware optimization pipeline offering both L-BFGS-B [63] and Adam [29] solvers. We explore spline parameter initialization schemes to accelerate convergence (e.g., for CAD data representation, using genetic algorithms refer to [33]).

Afterwards, we compare the CG and all-atom representations, demonstrating that purely pairwise, single-site models cannot capture many-body effects such as hydrogen bonding and intramolecular stiffness. We quantify force variability underestimation and report root-mean-square errors in force predictions.

Lastly, we summarize the limitations of single-site CG models and outline directions for incorporating directional interactions and many-body terms.

2. Paradigm of Coarse-Graining

In general, the governing equations of all-atomistic (AA) molecular dynamics (md) is a set of classical Newtonian equations for N atoms,

$$\dot{\mathbf{q}}_i = \mathbf{v}_i + \mathbf{f}_{\text{baro}} \tag{1}$$

$$m_i \dot{\mathbf{v}}_i = -\nabla_{\mathbf{q}_i} U(\mathbf{q}) + \mathbf{f}_{\text{thermo}} \ , \tag{2}$$

where $\mathbf{q} = (\mathbf{q}_1, \cdots, \mathbf{q}_N) \in \mathbb{R}^{3N}$ describes the position of the N particles. The auxiliary $f_{\text{thermo/baro}}$ represent the terms for thermostats (e.g., Nosé-Hoover [12], Berendsen [5], Langevin [9]) and barostats (e.g., Nosé-Hoover [10], Parrinello-Rahman [49], Andersen [2]). The chemical identities of a given system are determined by the AA interactions $U(\mathbf{q})$. Thus, U determines both structure and dynamics of a given system (e.g., polymeric solution).

A CG site/bead is often comprised by several atoms, for example, a monomer or even a Kuhn segment (see Kuhn scale matching in Kremer-Grest models [13]),

$$\mathbf{Q}_i = \frac{\sum_{j \in \text{CG}_i} m_j \mathbf{q}_j}{\sum_{j \in \text{CG}_i} m_j (\equiv M_i)} \ ,$$

where M_i is the inertial mass of the i^{th} CG bead.

Thus, the equations of motion (EoM) of the CG system are:

$$\dot{\mathbf{Q}}_i = \mathbf{V}_i + \mathbf{C} \tag{3}$$

$$M_i \dot{\mathbf{V}}_i = -\nabla_{\mathbf{Q}_i} U_{\text{CG}}(\mathbf{Q}) + \mathbf{D} \ , \tag{4}$$

where $U_{\text{CG}}(\mathbf{Q})$ is the conservative CG force field, and by construction, it determines the unique static structure of the system as stated by Henderson's theorem (1974)[1][17]. The auxiliary terms $\mathbf{C}$ and $\mathbf{D}$ are induced by the CG procedure, which determines the manifold of the dynamics.

Upon coarse-graining, the aforementioned terms may become stochastic due to the absence of *high-frequency dof*. For example, in dilute polymeric solutions, the solvents are much smaller compared to the CG particles, allowing them to be *abstracted into thermal perturbations*.

Moreover, the solutes can interact with each other via the solvents[2], making hydrodynamic interactions (HI) important.

The essence of coarse-graining is to find an appropriate set of parameters that describe U_{CG}, $\mathbf{C}$ and $\mathbf{D}$ from AA models, reproducing the same structures and dynamics generated from eq.1.

Therefore, the CG procedure can be divided into two parts:

(i) identifying the conservative force field and

(ii) establishing the EoM necessary to describe the dynamics and morphology of the system under study.

3. Theoretical Description of Coarse-Graining

Coarse-graining (CG) aims to simplify complex atomistic systems by reducing the number of degrees of freedom while preserving the essential physics relevant at larger length and time scales. This is achieved by grouping atoms into "CG sites" or "beads" and deriving an effective interaction potential, known as the Potential of Mean Force (PMF), that governs the behavior of these CG sites.

Let us consider an atomistic system comprising N atoms. The configuration of this system is described by the vector of atomic positions $\mathbf{q} = (\mathbf{q}_1, \cdots, \mathbf{q}_N) \in \mathbb{R}^{3N}$. The interactions between these atoms are governed by an atomistic potential energy function, $U(\mathbf{q})$ (see eq.1).

In a coarse-grained representation, these N atoms are mapped to M CG sites, where $M < N$. The configuration of the CG system is described by the vector of CG site positions $\mathbf{Q} = (\mathbf{Q}_1, \cdots, \mathbf{Q}_M) \in \mathbb{R}^{3M}$.

[1] Henderson's theorem states that two potential energy functions that produce the same radial distribution function (RDF) can differ only by a constant. In the realm of coarse-graining, though, we usually infer that remarkably different potentials give rise to RDF's which are indistinguishable within the available numerical accuracy.[50]

[2] Especially for "bad" solvents, where the Flory-Huggins interaction parameter, χ, is below 1/2 (θ-solvent regime) [59].

The mapping from the atomistic to the CG representation is defined by a linear contraction mapping function (or operator) $\mathbf{\Pi} : \mathbb{R}^{3N} \mapsto \mathbb{R}^{3M}$:

$$\mathbf{q} \mapsto \mathbf{Q} \equiv \mathbf{\Pi}(\mathbf{q}) \tag{5}$$

on the all-atomistic state space, forming $M(< N)$ sites as a function of the microscopic configuration $\mathbf{q}$.

For a reference I^{th} CG site in $\mathbf{Q}$, one has

$$\mathbf{Q}_I = \mathbf{\Pi}_I(\mathbf{q}) = \sum_{i=1}^{N} c_{I_i} \mathbf{q}_i \ , \ \forall \ I \in [M] \ , \tag{6}$$

where c_{I_i} are mapping coefficients. Typically, each CG bead $\mathbf{Q}_I$ represents a specific group of atoms, and its position is often chosen as the center of mass (CoM) of that group[3]. In that setting, $c_{I_i} \coloneqq m_i / \sum_{j \in \text{CG}_I} m_j$ if atom i belong to group I, and $c_{I_i} = 0$ otherwise, where m_i is the mass of atom i.

This mapping is a projection from a higher-dimensional space to a lower-dimensional one, meaning multiple atomistic configurations $\mathbf{q}$ can map to the same CG configuration $\mathbf{Q}$.

The set of all atomistic coordinates that map to a specific CG configuration is given by

$$\Omega(\mathbf{Q}) = \{\mathbf{q} \in \mathbb{R}^{3N} : \mathbf{\Pi}(\mathbf{q}) = \mathbf{Q}\} \ .$$

We use the "CG" subscript notation for quantities on the CG space.

The first goal of coarse-graining is to find an effective CG potential $U_{\text{CG}}(\mathbf{Q})$ such that the statistical properties of the CG system, particularly the probability distribution of CG configurations $P_{\text{CG}}(\mathbf{Q})$, match those derived from the underlying atomistic system.

In the canonical (NVT) ensemble, the probability density of observing an atomistic configuration $\mathbf{q}$ is given by:

$$P(\mathbf{q}) \coloneqq \frac{e^{-\beta U(\mathbf{q})}}{Z} \ , \tag{7}$$

where $\beta = 1/k_B T$, k_B is Boltzmann's constant, T is the absolute temperature, and $Z \coloneqq \int e^{-\beta U(\mathbf{q})} d\mathbf{q}$ is the microscopic partition function.

The probability density of observing a CG configuration $\mathbf{Q}$ is obtained by integrating $P(\mathbf{q})$ over all configurations $\mathbf{q} = \{\mathbf{q}_i\}_{i \in [N]}$ that are consistent with $\mathbf{Q}$:

$$P_{\text{CG}}(\mathbf{Q}) \coloneqq \int_{\mathbb{R}^{3N}} P(\mathbf{q}) \delta(\mathbf{Q} - \mathbf{\Pi}(\mathbf{q})) d\mathbf{q} \equiv \int_{\Omega(\mathbf{Q})} P(\mathbf{q}) d\mathbf{q} \ , \tag{8}$$

where $\delta(\cdot)$ is the Dirac delta function ensuring the mapping condition is met.

The (many-body) Potential of Mean Force (PMF), $U_{\text{CG}}(\mathbf{Q})$, is formally defined as the potential that reproduces this CG probability distribution [36]:

$$P_{\text{CG}}(\mathbf{Q}) \coloneqq \frac{e^{-\beta U_{\text{CG}}(\mathbf{Q})}}{Z_{\text{CG}}} \ , \tag{9}$$

where $Z_{\text{CG}} \coloneqq \int e^{-\beta U_{\text{CG}}(\mathbf{Q})} d\mathbf{Q}$, is the CG partition function.

Combining these, the PMF can be expressed (up to an arbitrary constant, often chosen such that $\int e^{-\beta U_{\text{CG}}(\mathbf{Q})} d\mathbf{Q} = \int_{\Omega(\mathbf{Q})} e^{-\beta U(\mathbf{q})} d\mathbf{q}$ as

$$U_{\text{CG}} = -\beta^{-1} \log \left(\int_{\Omega(\mathbf{Q})} e^{-\beta U(\mathbf{q})} d\mathbf{q} \right) + C \ . \tag{10}$$

Alternatively, as presented in Kalligiannaki et al. (reformulation of eq.6 in [25]), it is often expressed as:

$$e^{-\beta U_{\text{CG}}(\mathbf{Q})} = \frac{1}{\mathcal{N}_{\mathbf{Q}}} \int_{\mathbb{R}^{3N}} e^{-\beta U(\mathbf{q})} \delta(\mathbf{Q} - \mathbf{\Pi}(\mathbf{q})) d\mathbf{q} \ , \tag{11}$$

where $\mathcal{N}_{\mathbf{Q}} = \int \delta(\mathbf{Q} - \mathbf{\Pi}(\mathbf{q})) d\mathbf{q}$ is a normalization factor representing the volume of atomistic phase space compatible with the CG configuration $\mathbf{Q}$. This $\mathcal{N}_{\mathbf{Q}}$ term can depend on $\mathbf{Q}$ and contributes to the complexity of the PMF.

The exact PMF, $U_{\text{PMF}} \equiv U_{\text{CG}}(\mathbf{Q})$, os a highly complex many-body potential that depends on the positions of all M CG sites simultaneously. It implicitly accounts for all averaged-out atomistic degrees of freedom. Directly calculating or representing this many-body PMF is generally intractable due to the high-dimensional integral and its complex functional form. Therefore, approximations are necessary.

3.1. Approximation of the Potential of Mean Force

A common and practical approximation is to assume that the many-body PMF can be decomposed into a sum of simpler terms, such as bonded terms (bond stretching, angles, dihedrals) and non-bonded terms (typically pairwise interactions) [6],[8]:

$$\begin{aligned} U_{\text{CG}}(\mathbf{Q}) \approx & \sum_{\text{bonds}} u_{\text{bond}}(R_{IJ}) \\ & + \sum_{\text{angles}} u_{\text{angle}}(\phi_{IJK}) \\ & + \sum_{\text{dihedrals}} u_{\text{dihedral}}(\psi_{IJKL}) \\ & + \sum_{I<J, \text{ non-bonded}} u_{\text{pair}}(R_{IJ}) \ . \end{aligned}$$

[3] In our case, we CG map on the CoM of the water molecules, neglecting the bonded interactions.

Here, $R_{IJ} \equiv \|\mathbf{Q}_I - \mathbf{Q}_J\|$ is the distance between CG sites I and J, ϕ_{IJK} is the angle formed by sites I, J, K, and ψ_{IJKL} is the dihedral angle formed by sites I, J, K, L. Each of these terms $u(\cdot)$ is a simpler function (e.g., pairwise, three-body, four-body) than the full $U_{\mathrm{CG}}(\mathbf{Q})$ In Fig.5 we schematically summarize the above *additive force field*.

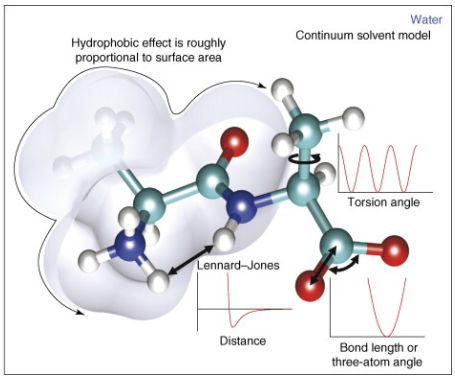


Figure 5: Molecular mechanics potential energy function with continuum solvent.[6]

For many systems, especially simpler ones or those where long-range electrostatics are screened or absent, the non-bonded interactions are often dominated by pairwise additive potentials. We consider an approximate/effective CG potential $U_{\mathrm{CG}}^{\mathrm{approx}}(\mathbf{Q};\boldsymbol{\theta})$ that includes a sum over non-bonded pair interactions:

$$U_{\mathrm{CG}}^{\mathrm{approx}}(\mathbf{Q};\boldsymbol{\theta}) = \sum_{I<J} u(R_{IJ};\boldsymbol{\theta}_{\mathrm{pair}}) + \cdots (\text{other terms if present}) \ ,$$

where $u(R_{IJ};\boldsymbol{\theta}_{\mathrm{pair}})$ is the parametrized pairwise interaction potential between CG sites I and J, separated by a distance R_{IJ}, The vector $\boldsymbol{\theta}_{\mathrm{pair}}$ contains the set of parameters that define the shape of this pair potential. The Force Matching method is then used to determine the optimal values for these parameters $\boldsymbol{\theta}$.

3.2. Spline Basis Expansion of the Pair CG Interaction Potential

To allow for flexibility in capturing the true form of the effective pair potential $u(R)$, which can be non-trivial, it is common to represent it as a linear combination of basis functions. A particularly powerful and widely used choice is a basis set of splines[4].

[4]Similar to Bézier curves in computer graphics, although the latter curves are defined by all control points globally, whereas B-Splines offer local control [16] (see Fig.6).

Let $u(R;\boldsymbol{\theta})$ be the pair potential as a function of the inter-particle distance R. We can approximate $u(R)$ using a set of N_b basis functions $\mathbf{B}(R) = \{B_k(R)\}_{k\in[N_b]}$:

$$u(R;\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{B}(R) = \sum_{k=1}^{N_b} \theta_k B_k(R) \ . \qquad (12)$$

Here $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{N_b})^\top$ is the vector of coefficients (parameters) to be determined.

In our workflow, we employ cubic splines as the basis set. Cubic splines are piecewise cubic polynomials defined over a set of intervals (see Fig.7 for a comparison with different spline interpolation methods).

1. **Knots:** A sequence of $N_k + 1$ knot points $r_0 < r_1 < \cdots < r_{N_k}$ is defined over the domain of interest for the pair potential, typically from $R = 0$ (or a small minimum distance) up to a cutoff distance $R = r_{cut}$ (so $r_{N_k} = r_{cut}$). Beyond r_{cut}, $u(R)$ is set to zero (e.g., by applying the indicator function $\mathbb{1}\{r < r_{cut}\}$ on $u(R;\boldsymbol{\theta})$).

2. **Piecewise Polynomials:** In each interval $[r_j, r_{j+1}]$, the spline function is a cubic polynomial.

3. **Continuity:** At each interior knot r_j ($j \in [N_k - 1]$), the spline function and its first and second derivatives are continuous. This smoothness is crucial for generating stable molecular dynamics simulations with well-defined forces.

4. **Basis Functions $B_k(R)$:** The individual $B_k(R)$ functions are themselves cubic splines[5]. The number of basis functions N_b is related to the number of knots N_k and the boundary conditions imposed on the splines (e.g., natural splines, clamped splines).

5. **Parameters θ_k:** The coefficients θ_k are the parameters that are optimized during the Force Matching procedure. They effectively control the amplitude of each basis spline, and their linear combination constructs the overall shape of $u(R)$.

[5]Often $B_k(R)$ are *B-splines* which have *local support*, meaning each $B_k(R)$ is non-zero only over a small number of adjacent knot intervals.

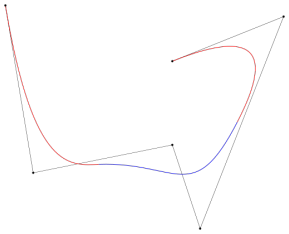


Figure 6: B-spline with control points/control polygon, and marked component curves.[by Wojciech mula]

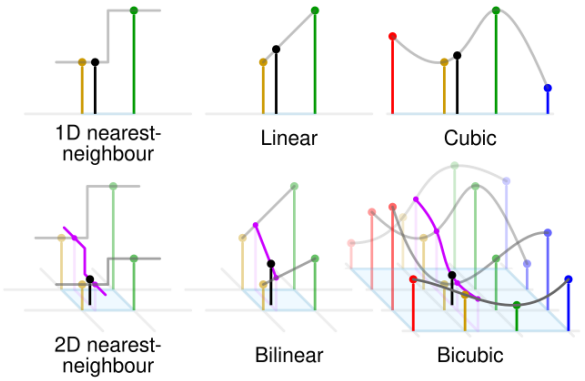


Figure 7: Comparison of nearest-neighbor, linear, cubic, bilinear and bicubic interpolation methods. The black dots correspond to the point being interpolated, and the red, yellow, green and blue dots correspond to the neighboring samples. Their heights above the ground correspond to their values.[by Cmglee]

The spline-based parametrization of the pair interaction potential provides a robust and adaptable framework for representing non-bonded pair potentials in CG models. They can approximate a wide variety of functional forms without prior assumptions about the potential's analytical shape (e.g., Lennard-Jones, Morse) The parameters $\boldsymbol{\theta} \in \mathbb{R}^{N_b}$ are then determined by minimizing the difference between forces calculated from this approximate CG potential and reference forces obtained from atomistic simulations, as dictated by the Force Matching method (described in the next section).

4. The Force Matching Optimization Method

The Force Matching (FM) method, pioneered by Ercolessi and Adams [11] and later generalized by Izvekov and Voth [20], provides a robust route to parametrize the approximate CG potential $U_{\text{CG}}^{\text{approx}}$. The core idea is to optimize the parameters $\boldsymbol{\theta}$ such that the forces derived from $U_{\text{CG}}^{\text{approx}}(\mathbf{Q};\boldsymbol{\theta})$ best reproduce the "true" mean forces acting on the CG sites, as obtained from the underlying atomistic simulation.

The exact mean force acting on the J^{th} CG site for a given CG configuration $\mathbf{Q} \in \mathbb{R}^{3M}$ is defined as the conditional expectation of the sum of atomistic forces acting on the atoms that constitute CG site J, given that the system is in CG configuration $\mathbf{Q}$:

$$\begin{aligned}\mathbf{F}_J^{\text{PMF}}(\mathbf{Q}) &:= \mathbb{E}\left[\sum_{i\in CG_J} \mathbf{f}_i(\mathbf{q}) \middle| \boldsymbol{\Pi}(\mathbf{q}) = \mathbf{Q}\right] \\ &= \mathbb{E}\left[\mathbf{f}_J(\mathbf{q}) | \boldsymbol{\Pi}(\mathbf{q}) = \mathbf{Q}\right] \ , \ \forall\ J \in [M] \ ,\end{aligned}$$

where $\mathbf{f}_i(\mathbf{q}) = -\nabla_{\mathbf{q}_i} U(\mathbf{q})$ is the instantaneous force on atom i in the atomistic system, and $\mathbf{f}_J(\mathbf{q}) = \sum_{i\in CG_J} \mathbf{f}_i(\mathbf{q})$ is the total instantaneous atomistic force on the group of atoms mapped to J^{th} CG particle. The expectation $\mathbb{E}[\cdot|\boldsymbol{\Pi}(\mathbf{q}) = \mathbf{Q}]$is taken over all atomistic configurations $\mathbf{q}$ that map to the specific CG configuration $\mathbf{Q}$. This $\mathbf{F}_J^{\text{PMF}}(\mathbf{Q})$ is also the negative gradient of the exact many-body PMF: $\mathbf{F}_J^{\text{PMF}}(\mathbf{Q}) = -\nabla_{\mathbf{Q}_J} U_{\text{CG}}(\mathbf{Q})$ (eq.7 in [25]).

The force on the J^{th} CG bead derived from our approximate CG potential $U_{\text{CG}}^{\text{approx}}(\mathbf{Q};\boldsymbol{\theta})$ is given by:

$$\mathbf{F}_{\text{CG},J}(\mathbf{Q},\boldsymbol{\theta}) = -\nabla_{\mathbf{Q}_J} U_{\text{CG}}^{\text{approx}}(\mathbf{Q};\boldsymbol{\theta}) \ . \tag{13}$$

If we consider, as established previously, that $U_{\text{CG}}^{\text{approx}}$ is primarily composed of a sum of pairwise non-bonded interactions $u()$ (and potentially other terms not explicitly dependent on $\boldsymbol{\theta}_{\text{pair}}$ for simplicity here, or whose parameters are determined separately):

$$U_{\text{CG}}^{\text{approx}}(\mathbf{Q};\boldsymbol{\theta}) = \sum_{I<K} u(R_{IK};\boldsymbol{\theta}) \tag{14}$$

(where we now use $\boldsymbol{\theta}$ generally for the parameters of the pair potential u). Then the force on site J is (similar to eq.10 in [25]):

$$\mathbf{F}_{\text{CG},J}(\mathbf{Q};\boldsymbol{\theta}) = -\sum_{K\neq J} \frac{\partial u(R_{JK};\boldsymbol{\theta})}{\partial R_{JK}} \hat{\mathbf{R}}_{JK} \ , \tag{15}$$

where $R_{JK} = \|\mathbf{Q}_J - \mathbf{Q}_K\|$ and $\hat{\mathbf{R}}_{JK} \equiv \frac{\mathbf{Q}_J - \mathbf{Q}_K}{R_{JK}}$ is the unit vector along the line connecting sites J and K.

Now, incorporating our cubic spline basis set representation for the pair potential $u(R,\boldsymbol{\theta}) = \sum_{k\in[N_b]} \theta_k B_k(R)$, the derivative term becomes:

$$\begin{aligned}\frac{\partial u(R_{JK};\boldsymbol{\theta})}{\partial R_{JK}} &= \nabla_{R_{JK}} \left(\sum_{k\in[N_b]} \theta_k B_k(R_{JK})\right) \\ &= \sum_{k\in[N_b]} \theta_k B_k'(R_{JK}) \ ,\end{aligned}$$

where $B_k'(R_{JK}$ is the derivative of k^{th} cubic spline basis function with respect to the inter-particle distance R_{JK}.

Thus, the contribution to the force $F_{\mathrm{CG},J}(\mathbf{Q},\boldsymbol{\theta})$ from pairwise interactions with parameter vector $\boldsymbol{\theta}$ is:

$$\mathbf{F}_{\mathrm{CG},J}(\mathbf{Q};\boldsymbol{\theta}) = -\sum_{K\neq J}\left(\sum_{l\in[N_b]}\theta_l B_l'(R_{JK})\right)\hat{\mathbf{R}}_{JK}\ . \tag{16}$$

Note the linearity of $\mathbf{F}_{\mathrm{CG},J}$ with respect to the parameters $\boldsymbol{\theta}$.

The Force Matching method seeks to find the parameters $\boldsymbol{\theta}$ by minimizing the sum of squared differences between the approximate CG forces and the true mean forces over a set of N_s configurations $\{\mathbf{Q}^{(n)}\}_{n=1}^{N_s}$ sampled from the atomistic simulation (e.g., via MD):

$$\chi^2(\boldsymbol{\theta}) = \sum_{n=1}^{N_s}\sum_{J=1}^{M}\|\mathbf{F}_{\mathrm{CG},J}(\mathbf{Q}^{(n)};\boldsymbol{\theta}) - \mathbf{F}_J^{\mathrm{PMF}}(\mathbf{Q}^{(n)})\|^2\ . \tag{17}$$

In practice, $\mathbf{F}_J^{\mathrm{PMF}}(\mathbf{Q}^{(n)})$ is approximated by the instantaneous atomistic force $\mathbf{f}_J(\mathbf{q}^{(n)})$ where $\boldsymbol{\Pi}(\mathbf{q}^{(n)}) = \mathbf{Q}^{(n)}$, under the assumption that for a sufficiently large and representative sample, the fluctuations average out.

This minimization problem is often a *linear least-squares problem* if $\mathbf{F}_{\mathrm{CG},J}$ is linear in $\boldsymbol{\theta}$, as is the case with our spline basis expansion.

5. Geometric Realization of Force Matching

The Force Matching procedure can be understood as a projection in a function space. Let *local mean force* $\mathbf{h}(\mathbf{q})$, representing the vector of instantaneous atomistic forces $(\mathbf{f}_1(\mathbf{q}),\cdots,\mathbf{f}_M(\mathbf{q}))$ projected onto the CG sites from the atomistic description (i.e., $\mathbf{h}_J(\mathbf{q}) = \sum_{i\in CG_J}\mathbf{f}_i(\mathbf{q})$). This force is a function of the atomistic coordinates $\mathbf{q}$ and belongs to a space $L^2(\mu)$ where $\mu(d\mathbf{q})$ is the canonical Gibbs (Boltzmann) measure $\mu(d\mathbf{q}) := Z^{-1}\exp\{-\beta U(\mathbf{q})\}d\mathbf{q}$.

We are seeking a CG force function $\mathbf{G}(\mathbf{Q})$ that best approximates $\mathbf{h}(\mathbf{q})$ when $\mathbf{Q} = \boldsymbol{\Pi}(\mathbf{q})$. That poses the minimization problem:

$$\mathcal{L}(\mathbf{G};\mathbf{h}) = \inf_{\mathbf{G}}\mathbb{E}_\mu\left[\|\mathbf{h}(\mathbf{q}) - \mathbf{G}\circ\boldsymbol{\Pi}(\mathbf{q})\|^2\right]\ , \tag{18}$$

where the infimum is taken over all functions $\mathbf{G}$ that depend only on the CG coordinates $\mathbf{Q}$. Namely, $\mathbf{G}\in L^2(\mu;\boldsymbol{\Pi})$, the subspace of functions measurable with respect to the CG coordinates and

$$L^2(\mu;\boldsymbol{\Pi}) = \{\mathbf{F}\in L^2(\mu)\mid \exists\ \mathbf{F}_{\mathrm{CG}}:\mathbb{R}^{3M}\mapsto\mathbb{R}^{3M}\ s.t.\ \mathbf{F}(\mathbf{q}) = \mathbf{F}_{\mathrm{CG}}(\boldsymbol{\Pi}(\mathbf{q}))\}\ .$$

The unique solution to this problem, as stated by Voth et al. [20], is the conditional expectation:

$$\mathbf{F}^*(\mathbf{Q}) = \mathbb{E}[\mathbf{h}(\mathbf{q})|\boldsymbol{\Pi}(\mathbf{q}) = \mathbf{Q}]\ , \tag{19}$$

which is precisely the true PMF force $\mathbf{F}^{\mathrm{PMF}}(\mathbf{Q})$ we defined earlier. It represents the best possible CG force function in the L^2 (square integrable) sense.

The Force Matching method then approximates this optimal (but generally intractable) $\mathbf{F}^*(\mathbf{Q})$ with a specific parametric form $\mathbf{F}_{\mathrm{CG}}(\mathbf{Q};\boldsymbol{\theta})$. The minimization of $\chi^2(\boldsymbol{\theta})$ finds the parameters $\boldsymbol{\theta}$ such that $\mathbf{F}_{\mathrm{CG}}(\mathbf{Q};\boldsymbol{\theta})$ is the "closest" function to $\mathbf{F}^*(\mathbf{Q})$ (or its sample estimate $\mathbf{h}(\mathbf{q}^{(n)})$) within the manifold of functions representable by the chosen parametric form (e.g., sums of spline-based pair potentials).

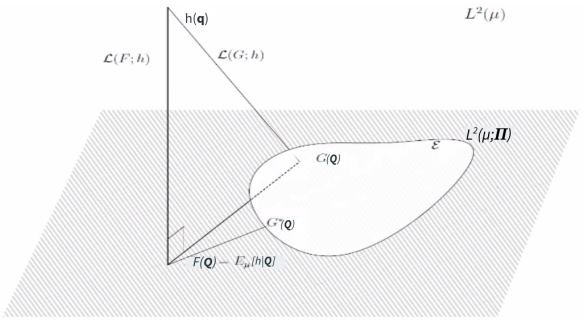


Figure 8: Geometrical interpretation of the Force Matching principle. The target quantity, such as the local mean force $\mathbf{h}(\mathbf{q})$, resides in the Hilbert space $L^2(\mu)$ of square-integrable functions with respect to the Boltzmann measure μ. The coarse-graining map $\boldsymbol{\Pi}$ projects atomistic configurations $\mathbf{q}$ to CG configurations $\mathbf{Q}$. The optimal CG representation of $\mathbf{h}(\mathbf{q})$, denoted as $\mathbf{F}^*(\mathbf{Q})$ (or $\mathbf{G}^*(\mathbf{Q})$), is its *orthogonal projection* onto the subspace $L^2(\mu;\boldsymbol{\Pi})$ of functions that depend only on $\mathbf{Q}$. This projection is given by the conditional expectation $\mathbf{F}^*(\mathbf{Q}) = \mathbb{E}_\mu[\mathbf{h}(\mathbf{q})|\boldsymbol{\Pi}(\mathbf{q}) = \mathbf{Q}]$. The Force Matching method aims to find a parametrically defined CG force $\mathbf{F}_{\mathrm{CG}}(\mathbf{Q};\boldsymbol{\theta})$ (represented by a point within the manifold $\mathcal{E}$) that is closest to this optimal target $\mathbf{F}^*(\mathbf{Q})$. The diagram illustrates the loss $\mathcal{L}(\mathbf{G};\mathbf{h})$ associated with an arbitrary CG function $\mathbf{G}(\mathbf{Q})$ and the minimized loss $\mathcal{L}(\mathbf{F}^*;\mathbf{h})$ achieved by the conditional expectation. The optimization process effectively seeks the best approximation within the constraints of the chosen functional form for the CG potential.[26]

6. Case Study on Water Liquid

To illustrate the application of the developed Force Matching methodology, a case study was performed on a system of 1000 water molecules, representing a *homogeneous liquid phase under ambient conditions* (300 K, 1 atm).

6.1. System Setup and Atomistic Simulation

The atomistic simulations employed the TIP3P (Transferable Intermolecular Potential 3 Point) water model [42]. The TIP3P model is a rigid three-site representation of water, with charges located

on each of the three atoms. It utilizes a Lennard-Jones potential acting solely between oxygen atoms to model van der Waals interactions, supplemented by Coulombic potentials between all charged sites (see documentation in LAMMPS[6]). Intramolecular geometry (O-H bond lengths and H-O-H angle) is typically maintained by constraint algorithms. The initial configuration, consisting of 1000 water molecules within a cubic simulation cell, was generated using the Packmol software package (see table 1), which facilitates the creation of dense, non-overlapping molecular systems suitable for initiating molecular dynamics simulations.

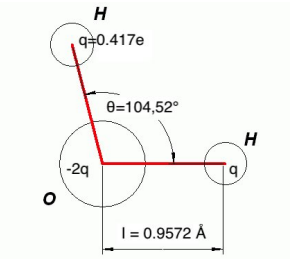


Figure 9: Parameters and geometry of the TIP3P water model. The three body system has a negative charge in the center that is two times that of its hydrogen point charges. Distance, angle and charges are empirically derived quantities.[52]

Table 1: Packmol TIP3P Water Configuration

Parameter	Value
Tolerance	2.0 Å
File Type	xyz
Output File	water_initial.xyz
Input Structure	water.xyz
Number of Molecules	1000
Box Dimensions	$40.0 \times 40.0 \times 40.0$ Å^3
Box Volume	64,000 Å^3
Estimated Density	$\sim$0.78 g/cm^3
Water Model	TIP3P
Packmol Results	
Final Objective Function	3.4687
Max Distance Violation	0.000000 Å
Max Constraint Violation	1.7757×10^{-3} Å

6.2. Atomistic Simulation Protocol

The generation of reference atomistic data followed a comprehensive workflow. Initially, requisite directories for output and logging were established. The Packmol-generated configuration was converted into a format suitable for the LAMMPS package, which was used for all molecular dynamics simulations.

The all-atom (AA) simulation was executed using LAMMPS. This simulation aimed to produce detailed trajectory and force information at the atomistic level. Key parameters and procedures for the AA simulation included:

- **Force Field and Interaction Potentials:** The simulation used `real` units and an `atom_style full`. Periodic boundary conditions were applied in all three dimensions. Interactions were modeled using a `lj/cut/coul/long` pair style with a 10.0Å cutoff for Lennard-Jones and short-range Coulombic terms; the Lennard-Jones potential was shifted to zero at the cutoff. Long-range electrostatics were handled via the Particle-Particle Particle-Mesh (PPPM) solver [62] with a specified accuracy of 1.0×10^{-4}. The TIP3P LJ parameters for oxygen-oxygen interactions were $\epsilon = 0.102$ kcal/mol and $\sigma = 3.188$ Å, with no LJ parameters assigned to hydrogen atoms. Intramolecular O-H bond lengths and H-O-H angles were constrained using the SHAKE algorithm [39]. Standard atomic masses for oxygen (15.9994 amu) and hydrogen (1.008 amu) were used. Intramolecular non-bonded interactions for 1-2 and 1-3 pairs were excluded, and 1-4 interactions were scaled by a factor of 0.5 for both LJ and Coulombic components.
- **Simulation Stages:** The protocol began with an energy minimization step. This was followed by a brief NVT equilibration phase of 5 ps, during which the system temperature was gently ramped from 50 K to 300 K using a Nosé-Hoover thermostat with a 100 fs damping parameter. Subsequently, the system underwent NPT equilibration at 300 K and 1.0 atm for 50 ps, employing a Nosé-Hoover thermostat and barostat (isotropic pressure coupling with a 2000 fs damping parameter), allowing the simulation cell volume to adjust. For the production phase, the ensemble was switched back to NVT at 300 K. A 1.0 fs timestep was used throughout all dynamics stages.
- **Production and Data Collection:** A production run of 1 ns was performed. During this phase, atomic coordinates and forces were recorded every 100 fs. These outputs served as the reference data for the subsequent CG procedure.

[6] LAMMPS: Large-scale Atomic/Molecular Massively Parallel Simulator

Following the AA simulation, a custom script processed the atomistic trajectories and forces to map them onto the CG representation. The optimized parameters from the force-matching algorithm were then used to parameterize a CG potential. This CG model was subsequently evaluated by performing a new MD simulation using LAMMPS with the derived CG interactions. Finally, results from both AA and CG simulations were analyzed, primarily through comparison of structural properties such as the radial distribution function.

6.3. Coarse-Grained Mapping

The transformation from the atomistic to the CG representation was achieved by mapping each water molecule to a single CG site (see Fig.10). This mapping was implemented via a custom Python script.

The position of the I^{th} CG particle, $\mathbf{Q}_I \in \mathbb{R}^3$, was defined as the center of mass (COM) of the I^{th} water molecule:

$$\mathbf{Q}_I = \frac{m_O \mathbf{q}_{I,O} + m_H \mathbf{q}_{I,H_1} + m_H \mathbf{q}_{I,H_2}}{m_O + 2m_H} \ , \qquad (20)$$

where $\mathbf{q}_{I,H_{1,2}}$ are the Cartesian coordinates of the oxygen and two hydrogen atoms, respectively, of the I^{th} water molecule, and $m_{O,H}$ are their corresponding atomic masses.

The total force, $\mathbf{F}_I$ acting on the I^{th} CG site was calculated as a mass-weighted sum of the instantaneous atomistic forces $\mathbf{f}$ acting on the constituent atoms of the I^{th} molecule:

$$\mathbf{F}_I = \frac{m_O \mathbf{f}_{I,O} + m_H \mathbf{f}_{I,H_1} + m_H \mathbf{f}_{I,H_2}}{m_O + 2m_H} \ . \qquad (21)$$

This mass-weighting convention is chosen to ensure consistency in momentum representation between the atomistic and CG levels, although direct summation of forces is also a common approach for PMF-based coarse-graining.

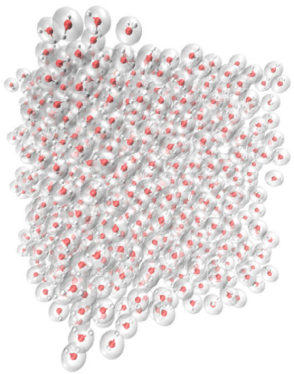


Figure 10: All-atom and center-of-mass coarse-grained representation of water liquid.[25]

6.4. Mapping Implementation Details

The mapping script performed several key operations. Firstly, it parsed the LAMMPS custom `dump` files containing atomistic positions and forces for each saved frame. A critical step involved the correct identification of atoms belonging to individual water molecules from the raw atomic coordinates. This molecular grouping was achieved based on geometric criteria: O-H bond distances were constrained to the range $[0.8, 1.2]$Å and the H-H intramolecular distance was required to be at least 1.4Å.

To efficiently perform the neighbor searches required for this grouping (i.e., finding hydrogen atoms in proximity to oxygen atoms) while correctly accounting for periodic boundary conditions (PBCs), we utilize the *k-d tree data structure*[7], specifically the `scipy.spatial.cKDTree` module.

Hydrogen atom positions were tiled into a $3 \times 3 \times 3$ supercell arrangement around the primary simulation cell prior to k-d tree construction to ensure accurate identification of bonds spanning PBCs. The k-d tree data structure allows for range and nearest-neighbor queries with an expected time complexity of $\mathcal{O}(\log(N))$ for N points in a fixed-dimensional space, offering a substantial computational advantage over brute-force $\mathcal{O}(N)$ searches per query [14]. In our case, we obtain an expected time complexity reduction:

$$\mathcal{O}(N_O N_H) \rightarrow \mathcal{O}(N_O \log(N_H)) \ ,$$

where $N_O = 1000$ and $N_H = 2 \times N_O$.

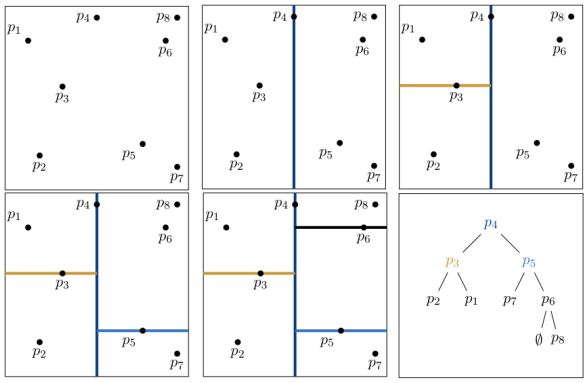


Figure 11: Recursively building a k-d tree on eight points. The hyperplanes are shown in the first five panels, while the whole tree is shown in the last subfigure on the lower right.[56]

Once atomistic configurations were mapped to CG sites and forces, the script also calculated the

[7] A K-Dimensional Tree (also known as K-D Tree) is a space-partitioning data structure for organizing points in a K-Dimensional space [56].

CG site-site *pair correlation* or *radial distribution function* (RDF), $g(R)$. The RDF provides a measure of the average structure of the CG liquid and is a crucial metric for validating the CG model. It was computed using a discretized histogram method ($g^{(\mathrm{AA})} \to g^{(\mathrm{CG})} \oplus$ discretization):

$$g^{(\mathrm{AA})}(r) = (\rho \cdot 4\pi r^2 \Delta r \cdot M)^{-1} \sum_{I<J} \delta^{(3)}(r - r_{IJ}) \implies$$

$$g(R_k) = \frac{\mathrm{count}(R_k)}{\rho_{\mathrm{CG}} \cdot V_{\mathrm{shell}}(R_k) \cdot M} , \qquad (22)$$

where $\mathrm{count}(R_k)$ is the number of CG site pairs found within the k^{th} radial distance bin $[R_k, R_k + \Delta R]$, $\rho_{\mathrm{CG}} = M/V_{\mathrm{box}}$ is the number density of the M CG beads in the simulation box of volume V_{box}, and $V_{\mathrm{shell}}(R_k)$ is the volume of the k^{th} spherical shell, approximately $4\pi R_k^2 \Delta R$.

The output of this mapping process for each atomistic frame consisted of arrays of CG site positions, CG site forces, the simulation box dimensions, and the calculated RDFs. These data formed the direct input for the subsequent force-matching parameter optimization procedure.

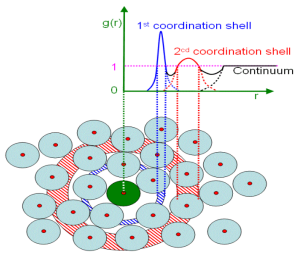


Figure 12: Schematic illustration of $g(r)$ dependence on Euclidean distance from reference particle, r.[35]

7. Force Matching Methodology

The core of the coarse-graining procedure lies in the optimization of the effective CG potential parameters. This is achieved by minimizing the discrepancy between forces derived from the parameterized CG potential and reference forces obtained from the atomistic simulation, mapped onto the CG sites. For the water system under study, where each molecule is represented by its center of mass and bonded interactions are eliminated, the effective CG potential is predominantly a non-bonded pairwise interaction. This pair potential is represented using a flexible cubic B-spline basis set.

7.1. Objective Function

The Force Matching algorithm seeks to find the optimal set of spline parameters $\boldsymbol{\theta}$ by minimizing a mean-squared error (MSE) objective function, $\mathcal{L}(\boldsymbol{\theta})$. This objective function quantifies the deviation between the reference forces $\mathbf{h}_I(\mathbf{q}_k)$ (atomistic forces summed onto the I^{th} CG site for configuration k) and the CG forces $\mathbf{F}_{\mathrm{CG},I}(\mathbf{Q}_k;\boldsymbol{\theta}) \in \mathbb{R}^{3M}$ derived from the approximate CG potential (see eq.13) $U_{\mathrm{CG}}^{\mathrm{approx}}(\mathbf{Q};\boldsymbol{\theta})$. The objective function is defined as (special case of eq.18):

$$\mathcal{L}(\boldsymbol{\theta};\mathbf{h}) = \frac{1}{N_{\mathrm{conf}}} \sum_{k=1}^{N_{\mathrm{conf}}} \sum_{I=1}^{M} \|\mathbf{h}_I(\mathbf{q}_k) - \mathbf{F}_{\mathrm{CG},I}(\mathbf{Q}_k;\boldsymbol{\theta})\|^2 , \qquad (23)$$

where N_{conf} is the total number of configurations sampled from the atomistic trajectory, M is the number of CG sites, $\mathbf{Q}_k$ is the vector of CG site positions for configuration k, and $\boldsymbol{\theta}$ represents the vector of parameters defining the CG potential (in this case, the values of the pair potential at the spline knots).

7.2. CG Force Calculation and Spline Parameterization

The CG force acting on site I is derived from the negative gradient of the approximate CG pair potential following eq.13. Assuming the CG potential is a sum of pairwise interactions $u(R_{IK};\boldsymbol{\theta})$ between I and K separated by distance R_{IK}, the force becomes the one in eq.15 (with $J \to I$). The pair interaction potential $u(R;\boldsymbol{\theta})$ is represented as a cubic spline. A set of $N_{\mathrm{knots}} (= N_b)$ knot points $\{r_j\}_{j\in[N_{\mathrm{knots}}]}$ is defined over the relevant range of inter-particle distances, typically from a minimum distance to a cutoff r_{cut}. The parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{N_{\mathrm{knots}}})$ are the values of the potential $u(r_j) = \theta_j$ at these knot points.

The `scipy.interpolate.CubicSpline` class is used to construct the piecewise cubic polynomial $u(R;\boldsymbol{\theta})$ that interpolates these knot values, and its analytical derivative $u'(R;\boldsymbol{\theta}) = \nabla_R u(R;\boldsymbol{\theta})$.

Thus, the force expression used in the implementation, based on the B-spline derivative formulation, is given as (similar with eq.16):

$$\mathbf{F}_{\mathrm{CG},I}(\mathbf{Q};\boldsymbol{\theta}) = -\sum_{K\neq J} u'(R_{IK};\boldsymbol{\theta})\hat{\mathbf{R}}_{IK} . \qquad (24)$$

The module `force_matching.py` specifically uses `CubicSpline(self.r_knots, spline_params)` to define the potential and `spline.derivative()` to obtain $u'(R_{IK};\boldsymbol{\theta})$.

7.3. Initialization Schemes of Spline Parameter

An important aspect for successful optimization is the choice of initial parameters $\boldsymbol{\theta}^{(0)}$. We employ an "SPC/E-Inspired Initialization" scheme, which is an improvement over a simpler Lennard-Jones-like

potential[8]. This scheme defines the initial potential $U(R_i)$ at each knot point R_i using physically informed parameters based on the SPC/E water model ($\epsilon = 0.65$ kcal/mol, $\sigma = 3.166$ Å)[7][Sklog-Wiki's SPC water model].

The potential is defined piecewise as:

$$U(R_i) = \begin{cases} 100\left(\frac{\sigma}{R_i}\right)^{12} & R_i < R_{\text{cut,hard}} \\ 4\epsilon\left[\left(\frac{\sigma}{R_i}\right)^{12} - \left(\frac{\sigma}{R_i}\right)^{6}\right] & R_{\text{cut,hard}} \le R_i \le R_{\text{cut}} \end{cases} \tag{25}$$

where $R_{\text{cut,hard}} = 2.5$ Å marks a transition from a very steep repulsive core to the Lennard-Jones form, and $R_{\text{cut}} = 10.0$ Å is the overall cutoff. The resulting knot values $\theta_i^{(0)} = U(R_i)$ are then smoothed using a *Savitzky-Golay filter*[54] to remove potential oscillations and finally shifted such that the potential is zero at the cutoff R_{cut} (i.e., $\theta_{N_{\text{knots}}}^{(0)} = 0$). This provides a physically reasonable starting point for the optimization.

7.4. Optimization Algorithms

The minimization of the objective function $\mathcal{L}(\boldsymbol{\theta})$ is performed using one of two approaches, depending on the dataset size:

1. **Deterministic Minimization (L-BFGS-B):** For smaller datasets (e.g. $N_{\text{config}} \lesssim 1000$ frames) the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B) [63], as implemented in `scipy.optimize.minimize`, is employed. This quasi-Newton method iteratively approximates the Hessian matrix to find a minimum. The objective function and its gradient (if provided analytically) are computed using the entire dataset of configurations.

2. **Stochastic Optimization (Adam):** For large datasets ($N_{\text{config}} \gg 1000$) a stochastic optimization approach is adopted to manage computational cost and memory. The Adaptive Moment Estimation (Adam) algorithm is used [29]. This method performs mini-batch updates:

   (i) A random subset (mini-batch) of N_{batch} configurations is sampled from the full dataset.

   (ii) The gradient of the batch MSE, $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{batch}}(\boldsymbol{\theta})$, is approximated using central finite differences:

$$\frac{\partial \mathcal{L}_{\text{batch}}}{\partial \theta_j} \approx \frac{1}{2\delta}[\mathcal{L}_{\text{batch}}(\dots, \theta_j + \delta, \dots) - \mathcal{L}_{\text{batch}}(\dots, \theta_j - \delta, \dots)]$$

   with a small perturbation δ (e.g., 10^{-6}).

   (iii) The spline parameters $\boldsymbol{\theta}$ are updated using Adam's rule[9], which incorporates adaptive learning rates based on estimates of the first and second moments of the gradients. This includes hyperparameters such as learning rate α, decay rates β_1, β_2, and ϵ. The implementation includes learning rate scheduling, gradient clipping, and parameter clipping to stabilize training.

7.5. Computational Multi-thread Optimizations

The `force_matching.py` module incorporates several computational strategies to enhance efficiency:

- **Vectorization and JIT Compilation:** Pairwise distance and vector calculations (`compute_pairs`) including the application of the minimum image convention (`min_image_vector`), are vectorized where possible and accelerated using Numba's Just-In-Time (JIT) compilation (`@njit`). This significantly speeds up these core computational loops.
- **Batch Processing:** For both objective function evaluation and stochastic optimization, data is loaded and processed in batches (`batch_size`). This reduces memory requirements, especially for large trajectory datasets.
- **Parallel Processing:** When using the L-BFGS-B optimizer, the calculation of the objective function across different batches of configurations can be parallelized using the `multiprocessing` module, leveraging multiple CPU cores.

[8] We define the potential $U(R)$ at knot points R_i using a Lennard-Jones–like formula, but with a constant repulsive core for $r < \sigma$:

$$U(R_i) = \begin{cases} \epsilon & R_i < \sigma \\ \epsilon\left[\left(\frac{\sigma}{R_i}\right)^{12} - 2\left(\frac{\sigma}{R_i}\right)^{6}\right] & R_i \ge \sigma \end{cases}.$$

This gives the initial values:

$$\theta_i^{(0)} = U(R_i)\ , \forall i \in [K]$$

[9]

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_{\boldsymbol{\theta}}\mathcal{L},\ v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_{\boldsymbol{\theta}}\mathcal{L})^2,$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}\ ,\ \hat{v}_t = \frac{v_t}{1-\beta_2^t}\ ,\ \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\frac{\hat{m}_t}{\sqrt{v_t}+\epsilon}\ .$$

with learning rate α, moments $\beta_{1,2}$, small constant ϵ to prevent division by zero.

- **Reduced Knot Count:** The number of spline knots (`n_spline_knots`, typically 20) is chosen as a balance between the flexibility of the potential representation and the dimensionality of the optimization problem. Fewer knots lead to faster optimization.
- **LAMMPS Table Generation:** After optimization, the resulting spline parameters are used to generate a tabulated potential file suitable for use in LAMMPS simulations, containing energies and forces at finely discretized distance points.
- **Logging and Visualization:** The optimization process, including parameter evolution and loss history, is logged. Visualizations of the potential and force curves, as well as comparisons between reference and CG force distributions, are generated to aid in assessing the quality of the derived CG model. A GIF illustrating the evolution of the spline function during training is also produced.

These methods collectively enable the robust and efficient parametrization of a cubic B-spline based pair potential for coarse-grained simulations of liquid water, starting from atomistic simulation data.

8. Coarse-Grained Simulation

Following the optimization of the pair potential parameters via Force Matching, a MD simulation of the CG water system was performed using LAMMPS. The purpose of this simulation was to assess the structural and dynamic properties of the CG model employing the derived effective pair potential.

The CG simulation workflow, illustrated in Figure 13, was implemented through the LAMMPS input script `in.cg_simulation`. The system employed `real` units with an `atomic` atom style under periodic boundary conditions, utilizing a cubic cell of approximately 40 Åfor 1000 CG beads (mass: 18.01528 amu each). Interactions were governed by a tabulated potential (`pair_style table linear 1000`) derived from the force-matching procedure.

Energy minimization proceeded through a two-stage conjugate gradient protocol: initial relaxation using a soft Lennard-Jones potential ($\epsilon = 1.0$ kcal/mol, $\sigma = 0.4$ Å), followed by hybrid optimization combining the tabulated CG potential with the soft LJ term. System equilibration involved temperature ramping from 100 K to 300 K over 50 ps (1.0 fs timestep), followed by 100 ps stabilization at 300 K (2.0 fs timestep) using Langevin thermostats with appropriate damping parameters.

The 5 ns production simulation utilized NVE integration with Langevin temperature regulation (300 K, 100.0 fs damping, 2.0 fs timestep). Concurrent structural analyses included *RDF calculations* and *Voronoi tessellation* (e.g., see Fig.14) for local environment characterization, with an additional 1 ns (default) extended sampling phase to enhance statistical accuracy of the Voronoi properties.

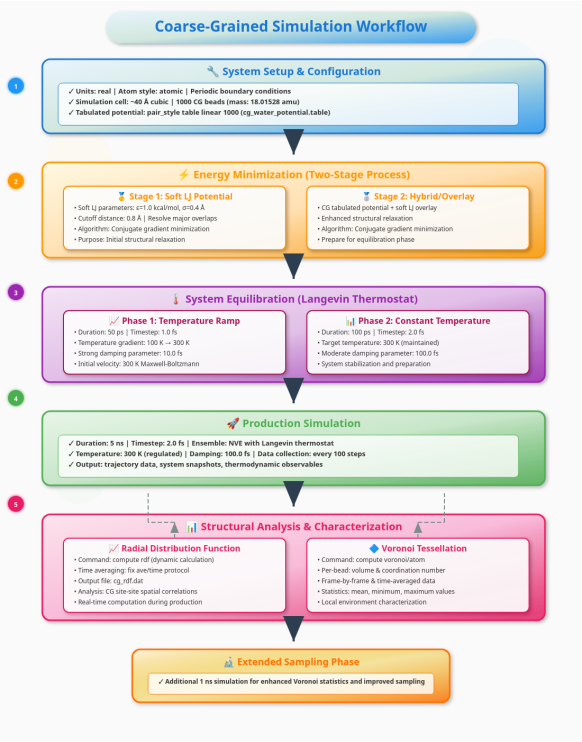


Figure 13: Solid arrows indicate sequential workflow progression. Dashed arrows show concurrent analysis processes. This CG simulation workflow enables comprehensive characterization of CG water behavior and structural properties.[created using the SVGViewer]

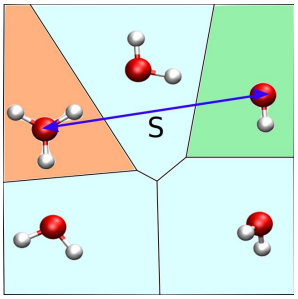


Figure 14: Illustration of the Voronoi tessellation used to assign a formal charge to oxygen atoms in the simulations by Andrade et al. (2023). Black lines correspond to the edges of the Voronoi polyhedron. Areas in light blue, orange, and green indicate regions in space pertaining to molecular water, hydronium, and hydroxide, respectively.[3]

8.1. Analysis and Validation

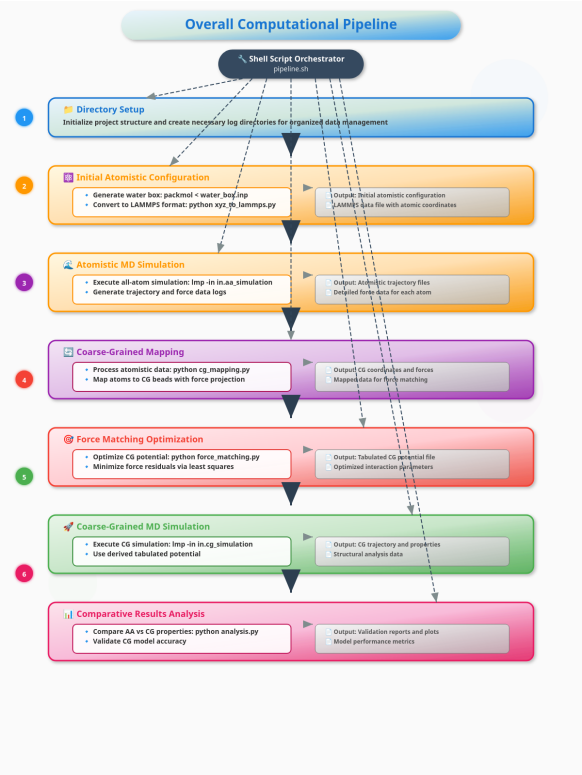
The `analysis.py` script was developed to systematically process output data from both the atomistic mapping stage and the CG simulations, thereby facilitating a quantitative comparison and rigorous validation of the derived CG model.

A primary component of this analysis involved the comparison of RDFs. We parse the RDF data generated by the LAMMPS CG simulation from the `cg_rdf.dat` file, extracting inter-particle distances (r), the $g(r)$ values, and coordination numbers. For comparative purposes, a reference atomistic RDF, typically the oxygen-oxygen RDF from the mapped atomistic data or a CG RDF derived directly from the CoM coordinates of water molecules in the AA simulation, was loaded.

Visualization of the CG potential and force characteristics formed another part of the analysis. The `plot_cg_potential` function utilized the optimized spline parameters and knot positions, along with the tabulated potential data, to generate plots depicting the interpolated spline potential against the discrete knot values, and a separate plot illustrating the derived force curve as a function of inter-particle distance.

We further employ extensive analysis of the Voronoi tessellation data obtained from the CG simulation. Functions were implemented to parse per-frame Voronoi `dump` files and to aggregate this data across multiple frames. Histograms of Voronoi cell volumes and neighbor counts were generated, offering insights into the local packing efficiency and coordination environment of the CG beads. A 2D histogram was also produced to explore correlations between Voronoi cell volume and neighbor count, providing a more nuanced characterization of local structural motifs. The temporal evolution of average Voronoi cell volumes and neighbor counts was plotted by processing global statistics files, which served to ascertain whether the CG system had achieved a structurally stable state during the simulation.

The automated computational pipeline for systematic coarse-grained model development was presented in the below diagram using the SVGViewer API. The workflow is orchestrated by a central shell script (`pipeline.sh`) that coordinates seven sequential phases: (1) directory setup and initialization, (2) atomistic configuration generation using `Packmol` and format conversion, (3) all-atom molecular dynamics simulation with trajectory logging, (4) coarse-grained mapping of atomistic data to CG representation, (5) force matching optimization to derive CG interaction parameters, (6) coarse-grained molecular dynamics simulation using the derived potential, and (7) comparative analysis and validation of atomistic versus coarse-grained results. Data flow arrows (dashed gray) indicate information transfer between processes, while orchestration lines (dark dashed) show script control over individual components. This automated pipeline ensures reproducibility and systematic validation of the coarse-graining methodology for liquid water systems.



9. Results & Remarks

This section presents the outcomes of the Force Matching (FM) optimization procedure applied to the coarse-grained (CG) water model. Two distinct experiments were conducted to investigate the impact of different initialization schemes and learning rates on the resulting CG potential and its ability to reproduce reference atomistic forces.

First we provide the RDF after CG mapping of the atomistic water system. The prominent first peak around 2.8 Å, the liquid-like structure with oscillations that decay with distance, and the technical context of the coarse-grained mapping approach used in the simulation is depicted in Fig.15.

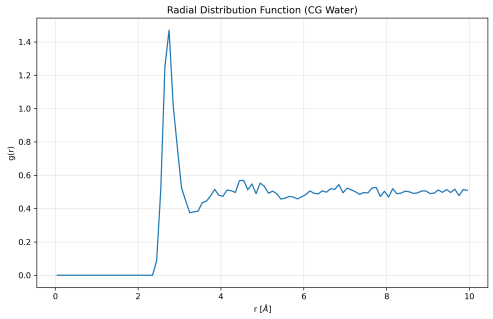


Figure 15: RDF $g(r)$ ($g(R_k)$ in eq. 22) for the CG water system containing 1000 molecules at liquid density. Each water molecule is represented by a single interaction site located at the CoM. The characteristic first peak at ~ 2.8 Å corresponds to the nearest-neighbor shell distance between water molecules in the liquid phase, while the subsequent oscillations indicate the liquid structure with partial ordering extending to ~ 6 Å before approaching the bulk value of unity at longer distances. The RDF was calculated from molecular dynamics trajectory data using periodic boundary conditions and the minimum image convention.

9.1. Experimental Settings

For both FM experiments, a consistent set of global parameters was maintained. The optimization was performed using a dataset of $N_{\text{config}} = 10000$ configurations, necessitating the use of the stochastic Adam optimizer. Computations were parallelized across 8 threads. During the Adam optimization, mini-batches of 100 frames (`frames_per_step`) were utilized. The CG pair potential was represented using cubic splines with $N_{\text{knots}} = 20$ knots, and the interaction cutoff was set to $r_{\text{cut}} = 10$ Å. The finite difference step for numerical gradient calculation was $\delta = 1 \times 10^{-6}$. Standard Adam hyperparameters were employed: momentum term $\beta_1 = 0.9$, squared gradient term $\beta_2 = 0.999$, and smoothing term $\epsilon = 1 \times 10^{-8}$. A learning rate decay factor of 0.98. was applied at each iteration. To ensure numerical stability, gradient clipping was set at 1×10^2 , and parameter values were clipped to a range of ± 10.0.The optimization was configured with a patience of 7 iterations for convergence based on improvement, though in practice, both experiments reached their maximum specified iterations.

9.2. Experiment 1: Lennard-Jones Initialization and Low Learning Rate

In the first experiment, the spline parameters were initialized using a naive Lennard-Jones (LJ) like potential form. The Adam optimizer was configured with a learning rate of $\alpha = 1 \times 10^{-3}$. The optimization was run for a maximum of 50 iterations, requiring approximately **2 days and 5 hours** of computation time. The process terminated upon reaching the maximum iteration count. This experiment yielded a final MSE of 89.41 between the CG forces and the reference atomistic forces.

Figure 16(a) illustrates the distribution of force magnitudes for the reference atomistic system and the CG system derived from this first experiment. Figure 16(b) provides a direct comparison of the magnitudes of CG forces versus their corresponding reference atomistic forces.

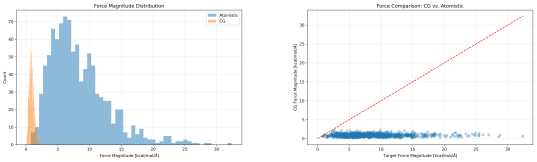


Figure 16: (Left) Histogram of force-magnitude distributions for reference atomistic forces (blue) and coarse-grained (CG) forces (orange), where the CG potential was initialized via the Lennard–Jones scheme and optimized with a learning rate of $\alpha = 1 \times 10^{-3}$. The CG forces display a much narrower distribution concentrated at lower magnitudes compared to the broader atomistic profile. (Right) Scatter plot of CG versus atomistic force magnitudes for the same optimized potential; the dashed red line denotes perfect agreement, highlighting the systematic underestimation of CG forces across the range.

9.3. Experiment 2: SPC/E-Based Initialization and Higher Learning Rate

The second experiment utilized the more physically informed SPC/E-based initialization scheme for the spline parameters. A higher learning rate of $\alpha = 1 \times 10^{-2}$ was employed with the Adam optimizer. This optimization was configured for a maximum of 30 iterations and completed in approximately **1 day and 39 minutes**, also terminating upon reaching the iteration limit. This approach resulted in a substantially higher final MSE of 1439.47.

Figure 17(a) presents the force magnitude distributions for the atomistic and CG systems from this second experiment, while Figure 17(b) shows the corresponding scatter plot of CG versus atomistic force magnitude.

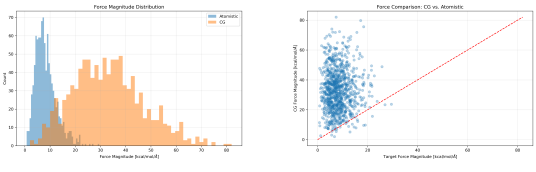


Figure 17: (Left) Histogram of force-magnitude distributions for reference atomistic forces (blue) and coarse-grained (CG) forces (orange), where the CG potential was initialized via the SPC/E-based scheme and optimized with a learning rate of $\alpha = 1 \times 10^{-2}$. The CG forces span a much broader dynamic range—more comparable to the atomistic distribution—though still exhibit some shifting of counts. (Right) Scatter plot of CG versus atomistic force magnitudes for the same optimized potential; the dashed red line denotes perfect agreement. Here the CG forces show considerable scatter and a tendency to overestimate at higher atomistic magnitudes, reflecting a closer but more variable match than in the first experiment.

The fact that one scheme underestimates and the other overestimates the variance suggests that neither is "perfect" on its own, but that a combined or adaptive strategy could probably leverage the strengths of both. By fusing or sequentially applying these initializations (or by adding a variance-matching term to our objective[10]), one should be able to drive both MSE and variance error down simultaneously.

[10] Adaptive weighting during training:

$$\mathcal{L} = \mathrm{MSE}(\mathbf{F}_{\mathrm{CG}}; \mathbf{h}) + \lambda |\mathrm{std}(\mathbf{F}_{\mathrm{CG}}) - \mathrm{std}(\mathbf{h})| \ ,$$

where we add a small regularization term that penalizes MSE and penalizes deviation in standard deviation from the reference force set.

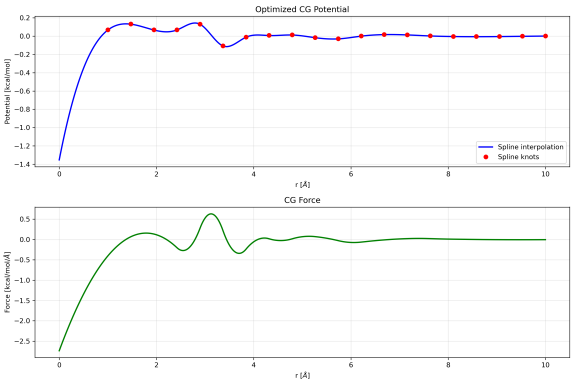


Figure 18: Optimized CG pair potential and corresponding force profile for the LJ–initialized spline parameters ($\alpha = 1 \times 10^{-3}$, 50 iterations). (Top) Spline-interpolated CG potential (blue line) with red markers indicating the optimized knot values at discrete r positions. The curve captures both the short-range repulsion and the damped oscillations beyond the first solvation shell. (Bottom) The negative derivative of the potential, showing the CG force as a function of interparticle separation. The force profile mirrors the potential features, with a strong repulsive core at small r and successive attractive/repulsive oscillations that decay at larger distances.

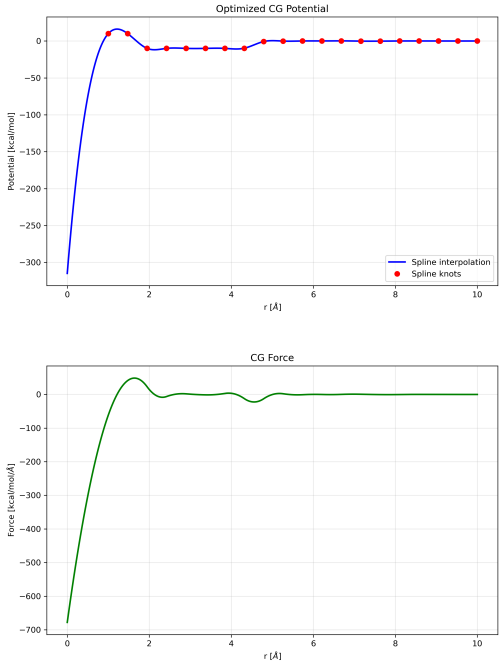


Figure 19: Optimized CG pair potential and corresponding force profile for the SPC/E–initialized spline parameters ($\alpha = 1 \times 10^{-2}$, 30 iterations).

Table 2: Comparison of spline parameter optimization results for two different initialization strategies in coarse-grained water force field development.

Parameter	LJ Initialization	SPC/E Initialization
Learning Rate	1×10^{-3}	1×10^{-2}
Number of Iterations	50	30
Total Optimization Time	2 days, 5 hours	1 day, 39 minutes
Average Step Time	–	2955.7 s
Final MSE	89.41	1439.47
Optimized Spline Parameters:		
c_1	6.816×10^{-2}	9.959
c_2	1.318×10^{-1}	1.000×10^{1}
c_3	6.816×10^{-2}	-1.000×10^{1}
c_4	6.874×10^{-2}	-9.999×10^{0}
c_5	1.314×10^{-1}	-9.987×10^{0}
c_6	-1.059×10^{-1}	-1.000×10^{1}
c_7	-1.071×10^{-2}	-1.000×10^{1}
c_8	7.645×10^{-3}	-9.783×10^{0}
c_9	1.348×10^{-2}	-7.561×10^{-1}
c_{10}	-1.580×10^{-2}	-1.007×10^{-1}
c_{11}	-2.881×10^{-2}	4.087×10^{-2}
c_{12}	1.059×10^{-3}	5.559×10^{-2}
c_{13}	1.799×10^{-2}	1.776×10^{-1}
c_{14}	1.450×10^{-2}	-1.566×10^{-1}
c_{15}	3.008×10^{-3}	-1.769×10^{-1}
c_{16}	-3.199×10^{-3}	3.083×10^{-2}
c_{17}	-4.755×10^{-3}	1.915×10^{-2}
c_{18}	-3.450×10^{-3}	-1.795×10^{-2}
c_{19}	-7.776×10^{-4}	-1.196×10^{-2}
c_{20}	1.839×10^{-3}	6.953×10^{-3}

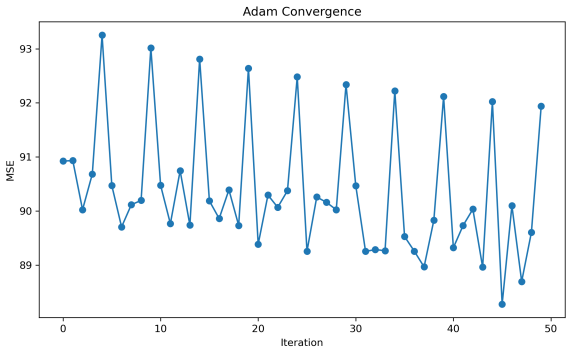


Figure 20: Training loss curve for the first experiment (LJ initialization, $\alpha = 1 \times 10^{-3}$, 50 iterations).

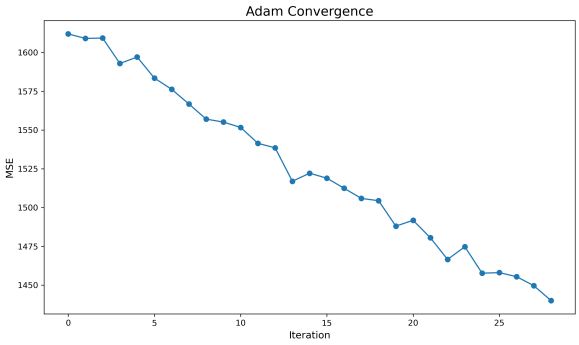


Figure 21: Training loss curve for the second experiment (SPC/E initialization, $\alpha = 1 \times 10^{-2}$, 30 iterations).

9.4. Structural Analysis of the CG water

Following the Force Matching optimization and subsequent CG MD simulations, a detailed structural analysis was performed to evaluate the properties of the resulting CG water model. This analysis primarily focused on the RDF and Voronoi tessellation characteristics, providing insights into the local packing and environment of the CG beads. The CG simulations utilized the potential derived from Experiment 1 (Lennard-Jones initialization, $\alpha = 1 \times 10^{-3}$, 50 iterations). These simulations consisted of a 2.5×10^6 steps (5 ns) production run, followed by an additional 1×10^5 step (0.2 ns) run for dedicated Voronoi data collection.

9.4.1. RDF and Coordination Analysis

The RDF, $g(r)$, and coordination number for the CG system are presented in Figure 22 (longer simulation) and Figure 23 (shorter simulation). Both RDFs exhibit characteristic liquid-like ordering with distinct first peaks indicating nearest neighbor shells and subsequent, less pronounced peaks for further coordination shells. The coordination number rises accordingly with increasing radius. Comparison with the reference atomistic RDF reveals that the CG model qualitatively reproduces the liquid structure, though discrepancies in peak positions and heights are apparent. The CG RDF shows an overestimation of the first peak's sharpness and height compared to the atomistic reference, consistent across both simulation lengths.

9.4.2. Voronoi Tessallation Analysis

The distribution of Voronoi cell volumes (rightmost panels in Figures 22-23) is unimodal and somewhat skewed, peaking around 50-70 Å^3, reflecting variations in local packing density within the liquid. The distribution of Voronoi neighbor counts (lower left panels) is centered around 15-17 neighbors, providing a direct measure of local connectivity.

The 2D histograms (upper right panels in Figures 22-23) correlating Voronoi cell volume with neighbor count reveal a positive correlation, indicating that particles with more neighbors tend to occupy larger Voronoi cells on average, though with considerable spread. The time evolution plots (center panels) show stable fluctuations around mean values for both Voronoi cell volume and neighbor count during the sampling periods, confirming structural equilibration.

9.4.3. Simulation Length Effects

Comparison between the longer ($2.5 \times 10^6 + 5 \times 10^5$ steps) and shorter ($1 \times 10^6 + 1 \times 10^5$ steps) simulations reveals remarkably similar structural characteristics. The RDF profiles, Voronoi property distributions, and time evolution behaviors are nearly identical, indicating that the primary structural features are well-converged even with the shorter

production run. This robustness across different simulation lengths demonstrates the stability of the CG system's structural ensemble and suggests adequate sampling for capturing liquid-like behavior.

The Voronoi analysis complements the RDF information by providing detailed characterization of local environment and packing, with consistent metrics across simulation conditions confirming the reliability of the CG model's structural representation.

10. Conclusion

In this project, we developed and applied the FM method for the parametrization of a CG model for homogeneous liquid water. The theoretical formulation established a clear pathway from detailed atomistic descriptions to simplified CG representations, with a particular focus on approximating the many-body Potential of Mean Force (PMF) using effective pairwise potentials. These pairwise interactions were flexibly represented using a cubic B-spline basis set, allowing for a data-driven determination of their functional form without a priori assumptions, as advocated in contemporary coarse-graining strategies (e.g., [25]).

A comprehensive computational pipeline was implemented, starting from all-atom (AA) simulations of 1000 TIP3P water molecules (e.g., [24]). This pipeline included a center-of-mass mapping of atomistic configurations and forces to the CG level, followed by a robust FM algorithm designed to optimize the spline parameters of the CG interaction potential. The FM implementation leveraged modern computational techniques, including stochastic optimization via the Adam algorithm for large datasets and numerical optimizations such as JIT compilation for performance-critical routines.

The investigation into FM optimization strategies, specifically varying the initial guess for the spline parameters and the optimizer's learning rate, revealed significant sensitivity of the outcomes to these choices. An experiment employing a naive Lennard-Jones-like initialization with a lower learning rate ($\alpha = 10^{-3}$) resulted in a lower final Mean Squared Error (MSE of 89.41) between CG and reference atomistic forces. However, the qualitative agreement of the force distributions was suboptimal, with CG forces being systematically underestimated. Conversely, an SPC/E-inspired initialization combined with a higher learning rate ($\alpha = 10^{-2}$) yielded a higher MSE (1439.47) but produced CG force distributions that, despite a tendency to overestimate forces, captured a broader dynamic range more comparable to the atomistic reference. These results underscore that while MSE is a valuable quantitative metric, a holistic assessment including qualitative aspects of force reproduction is crucial in FM (e.g., [11]; [20]). The substantial computational investment for these optimizations (approximately 2 days per experiment) highlights the cost associated with large-scale FM.

Subsequent CG MD simulations, utilizing the potential derived from the first experiment (LJ initialization), demonstrated the stability and utility of the parametrized model. In-depth structural analysis, primarily through RDFs and Voronoi tessellation, indicated that the CG model qualitatively reproduced the liquid structure of water. The CG RDF exhibited characteristic liquid ordering, though quantitative deviations in peak positions and heights were noted when compared to the atomistic reference. The Voronoi analysis, offering a detailed view of local packing and coordination environments, showed stable structural properties over extended simulation times, with these properties being robust even with moderate production run lengths. The consistency of Voronoi metrics across different simulation lengths (2 ns vs. 5 ns production runs) suggests good structural convergence for the CG model under the derived potential.

Despite these achievements, the study acknowledges certain limitations. The CG model developed herein relies exclusively on pairwise non-bonded interactions, thereby neglecting potential many-body effects or explicit orientation-dependent terms which can be critical for accurately capturing the complex behavior of associated liquids like water [47]. The exploration of FM hyperparameters, while insightful, was not exhaustive [60]. Furthermore, the validation of the CG model was predominantly structural; an assessment of dynamic properties (e.g., diffusion coefficients, viscosity) was beyond the current scope but is essential for a complete characterization of the model's fidelity [51].

Future work could beneficially extend this framework by incorporating many-body terms into the CG potential (e.g., through explicit three-body splines or machine-learned density-dependent functionals [4],[23]) or by investigating more sophisticated mapping schemes that retain some degree of orientational information, which might improve the accuracy for systems where directionality is key [41]. A more systematic and potentially automated hyperparameter optimization study for the FM process, possibly leveraging machine learning approaches (e.g., Bayesian optimization, neural-network-accelerated fitting [22],[57]), could lead to more robust and transferable CG potentials. Crucially, the validation of dynamic properties will be necessary to assess the broader applicability of the derived CG models for simulating time-dependent

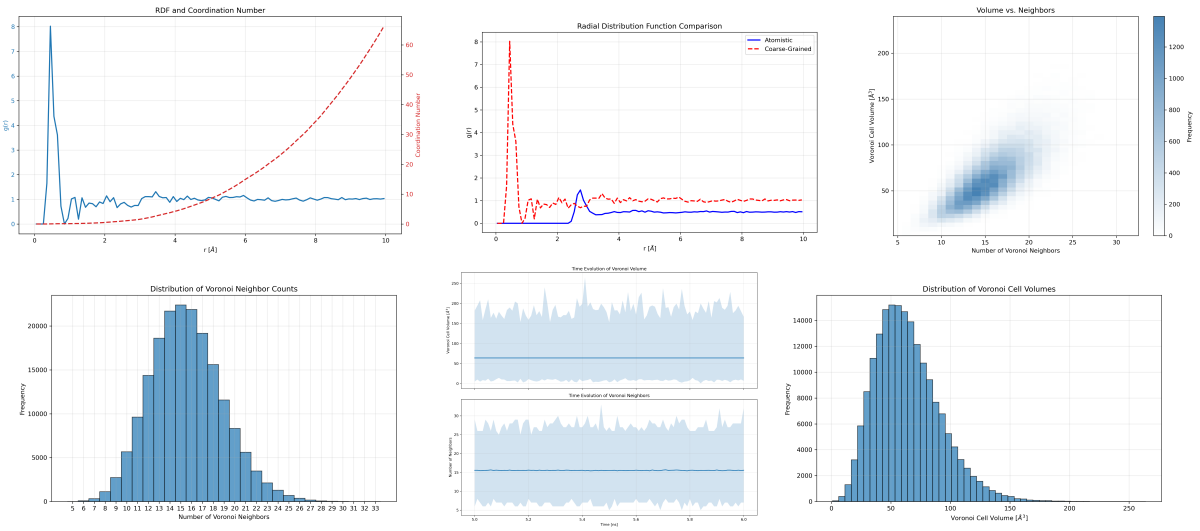


Figure 22: Structural analysis of CG water model from $2.5 \times 10^6 + 5 \times 10^5$ step simulation using LJ initialization with $\alpha = 10^{-3}$. First row: (left) RDF $g(r)$ (blue solid line, left y-axis) and coordination number (red dashed line, right y-axis); (center) comparison of atomistic reference RDF (blue solid line) with CG RDF (red dashed line); (right) 2D histogram correlating Voronoi cell volume with number of Voronoi neighbors. Second row: (left) distribution of Voronoi neighbor counts centered around 15-17 neighbors; (center) time evolution of mean Voronoi cell volume (top panel) and mean number of Voronoi neighbors (bottom panel) with shaded areas indicating min-max ranges; (right) distribution of Voronoi cell volumes peaking around 50-70 Å^3.

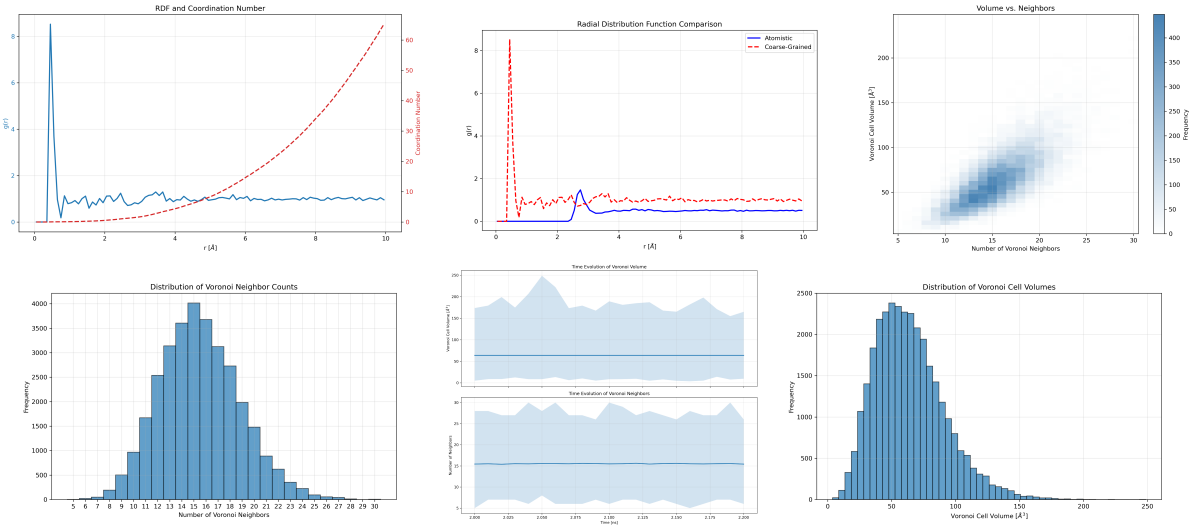


Figure 23: Structural analysis of CG water model from shorter $1 \times 10^6 + 1 \times 10^5$ step simulation. Panel arrangement identical to Figure 22, demonstrating similar structural characteristics despite reduced simulation length. The consistency between results validates the robustness of the CG model's structural properties and confirms adequate sampling even with shorter production runs.

phenomena [15].

In conclusion, this project provides a functional and well-documented FM pipeline for developing spline-based CG potentials. It offers valuable practical insights into the parametrization process, the impact of optimization choices, and the structural validation of CG models for molecular simulations, laying a foundation for further refinements and applications to more complex systems.

References

[1] B. J. Alder and T. E. Wainwright, «Phase transition for a hard sphere system», The Journal of Chemical Physics **27**, 1208–1209, ISSN: 0021-9606 (1957) 10.1063/1.1743957.

[2] H. C. Andersen, «Molecular dynamics simulations at constant pressure and/or temperature», The Journal of Chemical Physics **72**, 2384–2393, ISSN: 0021-9606 (1980) 10.1063/1.439486.

[3] M. Andrade, R. Car, and A. Selloni, «Probing the self-ionization of liquid water with ab initio deep potential molecular dynamics», Proceedings of the National Academy of Sciences of the United States of America **120**, e2302468120 (2023) 10.1073/pnas.2302468120.

[4] J. Behler and M. Parrinello, «Generalized neural-network representation of high-dimensional potential-energy surfaces», Phys. Rev. Lett. **98**, 146401 (2007) 10.1103/PhysRevLett.98.146401.

[5] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, «Molecular dynamics with coupling to an external bath», The Journal of Chemical Physics **81**, 3684–3690, ISSN: 0021-9606 (1984) 10.1063/1.448118.

[6] F. E. Boas and P. B. Harbury, «Potential energy functions for protein design», Current Opinion in Structural Biology **17**, Theory and simulation / Macromolecular assemblages, 199–204, ISSN: 0959-440X (2007) https://doi.org/10.1016/j.sbi.2007.03.006.

[7] M. Campo, «Structural and dynamic properties of spc/e water», Papers in Physics **2**, 10.4279/pip.020001 (2010) 10.4279/pip.020001.

[8] C.-E. Chang, Y.-M. Huang, L. Mueller, and W. You, «Investigation of structural dynamics of enzymes and protonation states of substrates using computational tools», Catalysts **6**, 82 (2016) 10.3390/catal6060082.

[9] R. L. Davidchack, R. Handel, and M. V. Tretyakov, «Langevin thermostat for rigid body dynamics», The Journal of Chemical Physics **130**, 10.1063/1.3149788, ISSN: 1089-7690 (2009) 10.1063/1.3149788.

[10] S. Dey, *Minimal modification to nosé-hoover barostat enables correct npt sampling*, 2020.

[11] F. Ercolessi and J. B. Adams, «Interatomic potentials from first-principles calculations: the force-matching method», Europhysics Letters (EPL) **26**, 583–588, ISSN: 1286-4854 (1994) 10.1209/0295-5075/26/8/005.

[12] D. J. Evans and B. L. Holian, «The nose–hoover thermostat», The Journal of Chemical Physics **83**, 4069–4074, ISSN: 0021-9606 (1985) 10.1063/1.449071.

[13] R. Everaers, H. A. Karimi-Varzaneh, F. Fleck, N. Hojdis, and C. Svaneborg, «Kremer–grest models for commodity polymer melts: linking theory, experiment, and simulation at the kuhn scale», Macromolecules **53**, 1901–1916 (2020) 10.1021/acs.macromol.9b02428.

[14] J. H. Friedman, J. L. Bentley, and R. A. Finkel, «An algorithm for finding best matches in logarithmic expected time», ACM Trans. Math. Softw. **3**, 209–226, ISSN: 0098-3500 (1977) 10.1145/355744.355745.

[15] M. G. Guenza, M. Dinpajooh, J. McCarty, and I. Y. Lyubimov, «Accuracy, transferability, and efficiency of coarse-grained models of molecular liquids», The Journal of Physical Chemistry B **122**, PMID: 30153027, 10257–10278 (2018) 10.1021/acs.jpcb.8b06687.

[16] M. Hazewinkel, *Encyclopaedia of mathematics: supplement*, Encyclopaedia of Mathematics τ. 1 (Springer Netherlands, 1997), ISBN: 9780792347095.

[17] R. Henderson, «A uniqueness theorem for fluid pair correlation functions», Physics Letters A **49**, 197–198, ISSN: 0375-9601 (1974) https://doi.org/10.1016/0375-9601(74)90847-0.

[18] T. S. Hofer, B. M. Rode, A. B. Pribil, and B. R. Randolf, «Simulations of liquids and solutions based on quantum mechanical forces», in *Theoretical and computational inorganic chemistry*, Vol. 62, edited by R. van Eldik and J. Harvey, Advances in Inorganic Chemistry (Academic Press, 2010), pp. 143–175, https://doi.org/10.1016/S0898-8838(10)62004-1.

[19] W. Humphrey, A. Dalke, and K. Schulten, «Vmd: visual molecular dynamics», Journal of molecular graphics **14**, 33–38 (1996).

[20] S. Izvekov and G. A. Voth, «A multiscale coarse-graining method for biomolecular systems», The Journal of Physical Chemistry B **109**, PMID: 16851243, 2469–2473 (2005) 10.1021/jp044629q.

[21]J. Jin, A. J. Pak, A. E. P. Durumeric, T. D. Loose, and G. A. Voth, «Bottom-up coarse-graining: principles and perspectives», Journal of Chemical Theory and Computation **18**, PMID: 36070494, 5759–5791 (2022) 10.1021/acs.jctc.2c00643.

[22]R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse, and R. Asahi, «On-the-fly active learning of interatomic potentials for large-scale atomistic simulations», The Journal of Physical Chemistry Letters **11**, 10.1021/acs.jpclett.0c01061 (2020) 10.1021/acs.jpclett.0c01061.

[23]S. T. John, *Many-body coarse-grained interactions using gaussian approximation potentials*, 2017.

[24]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, «Comparison of simple potential functions for simulating liquid water», The Journal of Chemical Physics **79**, 926–935, ISSN: 0021-9606 (1983) 10.1063/1.445869.

[25]E. Kalligiannaki, A. Chazirakis, A. Tsourtis, M. A. Katsoulakis, P. Plecháč, and V. Harmandaris, «Parametrizing coarse grained models for molecular systems at equilibrium», The European Physical Journal Special Topics **225**, 1347–1372, ISSN: 1951-6401 (2016) 10.1140/epjst/e2016-60145-x.

[26]E. Kalligiannaki, V. Harmandaris, M. Katsoulakis, and P. Plechac, «The geometry of generalized force matching in coarse-graining and related information metrics», The Journal of Chemical Physics, 10.1063/1.4928857 (2015) 10.1063/1.4928857.

[27]H. A. Karimi-Varzaneh, H.-J. Qian, X. Chen, P. Carbone, and F. Müller-Plathe, «Ibisco: a molecular dynamics simulation package for coarse-grained simulation», Journal of Computational Chemistry **32**, 1475–1487 (2011) https://doi.org/10.1002/jcc.21717.

[28]M. Karplus, «Molecular dynamics simulations of biomolecules», Accounts of Chemical Research **35**, PMID: 12069615, 321–323 (2002) 10.1021/ar020082r.

[29]D. P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, 2017.

[30]J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell Jr, and R. W. Pastor, «Update of the charmm all-atom additive force field for lipids: validation on six lipid types», The journal of physical chemistry B **114**, 7830–7843 (2010).

[31]K. Kremer and G. S. Grest, «Dynamics of entangled linear polymer melts: a molecular-dynamics simulation», The Journal of Chemical Physics **92**, 5057–5086, ISSN: 0021-9606 (1990) 10.1063/1.458541.

[32]S. Kullback and R. A. Leibler, «On Information and Sufficiency», The Annals of Mathematical Statistics **22**, 79–86 (1951) 10.1214/aoms/1177729694.

[33]G. Kumar, P. Kalra, and S. Dhande, «Parameter optimization for b-spline curve fitting using genetic algorithms», in The 2003 congress on evolutionary computation, 2003. cec '03. Vol. 3 (2003), 1871–1878 Vol.3, 10.1109/CEC.2003.1299902.

[34]W. Li, C. Burkhart, P. Polińska, V. Harmandaris, and M. Doxastakis, «Backmapping coarse-grained macromolecules: an efficient and versatile machine learning approach», The Journal of Chemical Physics **153**, 041101, ISSN: 0021-9606 (2020) 10.1063/5.0012320.

[35]Y. Liao, *Practical electron microscopy and database: www.globalsino.com/em/* (Feb. 2013).

[36]L. Lu and G. A. Voth, «The multiscale coarse-graining method. vii. free energy decomposition of coarse-grained effective potentials», The Journal of Chemical Physics **134**, 224107, ISSN: 0021-9606 (2011) 10.1063/1.3599049.

[37]A. P. Lyubartsev and A. Laaksonen, «Calculation of effective interaction potentials from radial distribution functions: a reverse monte carlo approach», Phys. Rev. E **52**, 3730–3737 (1995) 10.1103/PhysRevE.52.3730.

[38]A. P. Lyubartsev, A. Naômé, D. P. Vercauteren, and A. Laaksonen, «Systematic hierarchical coarse-graining with the inverse monte carlo method», The Journal of Chemical Physics **143**, 243120, ISSN: 0021-9606 (2015) 10.1063/1.4934095.

[39]D. Macuglia, «Shake and the exact constraint satisfaction of the dynamics of semi-rigid molecules in cartesian coordinates, 1973–1977», Archive for History of Exact Sciences **77**, 345–371 (2023).

[40]P. Mark and L. Nilsson, «Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k», The Journal of Physical Chemistry A **105**, 9954–9960 (2001) 10.1021/jp003020w.

[41]S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, «The martini force field: coarse grained model for biomolecular simulations», The journal of physical chemistry B **111**, 7812–7824 (2007).

[42]L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, «Packmol: a package for building initial configurations for molecular dynamics simulations», Journal of Computational Chemistry **30**, 2157–2164 (2009) https://doi.org/10.1002/jcc.21224.

[43]S. Mashayak, M. Jochum, K. Koschke, N. Aluru, V. Rühle, and C. Junghans, «Relative entropy and optimization-driven coarse-graining methods in votca», PloS one **10**, e0131754 (2015) 10.1371/journal.pone.0131754.

[44]A. May, R. Pool, E. van Dijk, J. Bijlard, S. Abeln, J. Heringa, and K. A. Feenstra, «Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins», Bioinformatics **30**, 326–334, ISSN: 1367-4803 (2013) 10.1093/bioinformatics/btt675.

[45]I. Nadkarni, J. Jeong, B. Yalcin, and N. R. Aluru, «Modulating coarse-grained dynamics by perturbing free energy landscapes», The Journal of Physical Chemistry A **128**, PMID: 39540849, 10029–10040 (2024) 10.1021/acs.jpca.4c04530.

[46]C. Navarro, M. Majewski, and G. de Fabritiis, *Top-down machine learning of coarse-grained protein force-fields*, 2023.

[47]H. T. L. Nguyen and D. M. Huang, «Systematic bottom-up molecular coarse-graining via force and torque matching using anisotropic particles», The Journal of Chemical Physics **156**, 184118, ISSN: 0021-9606 (2022) 10.1063/5.0085006.

[48]H. Ohtaki and H. Yamatera, «Chapter 3 - molecular dynamics simulations of liquids and solutions», in *Structure and dynamics of solutions*, Vol. 79, edited by H. Ohtaki and H. Yamatera, Studies in Physical and Theoretical Chemistry (Elsevier, 1992), pp. 57–132, https://doi.org/10.1016/B978-0-444-89651-3.50008-5.

[49]M. Parrinello and A. Rahman, «Polymorphic transitions in single crystals: a new molecular dynamics method», Journal of Applied Physics **52**, 7182–7190, ISSN: 0021-8979 (1981) 10.1063/1.328693.

[50]R. Potestio, «Is henderson's theorem practically useful», JUnQ **3**, 13–15 (2013).

[51]D. Reith, M. Pütz, and F. Müller-Plathe, «Deriving effective mesoscale potentials from atomistic simulations», Journal of Computational Chemistry **24**, 1624–1636 (2003) https://doi.org/10.1002/jcc.10307.

[52]M. Rosellen, «Using advanced computational methods to model the binding of antibody complexes: a case study from the coagulation cascade», in (June 2021).

[53]T. Sanyal and M. S. Shell, «Coarse-grained models using local-density potentials optimized with the relative entropy: application to implicit solvation», The Journal of Chemical Physics **145**, 034109, ISSN: 0021-9606 (2016) 10.1063/1.4958629.

[54]A. Savitzky and M. J. E. Golay, «Smoothing and differentiation of data by simplified least squares procedures.», Analytical Chemistry **36**, 1627–1639 (1964) 10.1021/ac60214a047.

[55]R. Shi, H.-J. Qian, and Z.-Y. Lu, «Coarse-grained molecular dynamics simulation of polymers: structures and dynamics», WIREs Computational Molecular Science **13**, e1683 (2023) https://doi.org/10.1002/wcms.1683.

[56]M. Skrodzki, *The k-d tree data structure and a proof for neighborhood computation in expected logarithmic time*, 2019.

[57]J. Smith, O. Isayev, and A. Roitberg, «Ani-1: an extensible neural network potential with dft accuracy at force field computational cost», Chem. Sci. **8**, 10.1039/C6SC05720A (2017) 10.1039/C6SC05720A.

[58]T. Sun, V. Minhas, N. Korolev, A. Mirzoev, A. P. Lyubartsev, and L. Nordenskiöld, «Bottom-up coarse-grained modeling of dna», Frontiers in Molecular Biosciences **8**, eCollection 2021, 645527, ISSN: 2296-889X (2021) 10.3389/fmolb.2021.645527.

[59]T. Tadros, «Flory-huggins interaction parameter», in *Encyclopedia of colloid and interface science*, edited by T. Tadros (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), pp. 523–524, ISBN: 978-3-642-20665-8, 10.1007/978-3-642-20665-8_89.

[60]H. Wang, C. Junghans, and K. Kremer, «Comparative atomistic and coarse-grained study of water: what do we lose by coarse-graining?», The European physical journal. E, Soft matter **28**, 221–9 (2009) 10.1140/epje/i2008-10413-5.

[61]W. Wang and R. Gómez-Bombarelli, «Coarse-graining auto-encoders for molecular dynamics», npj Computational Materials **5**, 10.1038/s41524-019-0261-5, ISSN: 2057-3960 (2019) 10.1038/s41524-019-0261-5.

[62]X. Zhang and R. Bridson, «A pppm fast summation method for fluids and beyond», ACM Trans. Graph. **33**, 10.1145/2661229.2661261, ISSN: 0730-0301 (2014) 10.1145/2661229.2661261.

[63]C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, «Algorithm 778: l-bfgs-b: fortran subroutines for large-scale bound-constrained optimization», **23**, 550–560, ISSN: 0098-3500 (1997) 10.1145/279232.279236.