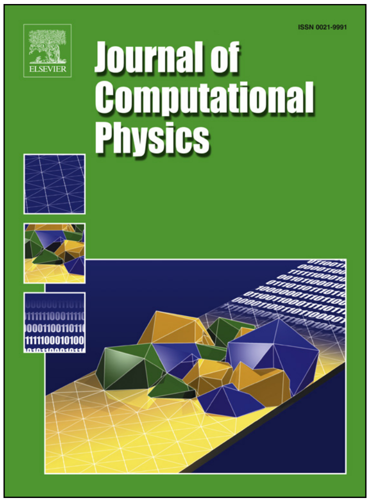

B-spline tight frame based force matching method

Jianbin Yang, Guanhua Zhu, Dudu Tong, Lanyuan Lu, Zuowei Shen

PII:	S0021-9991(18)30106-2
DOI:	https://doi.org/10.1016/j.jcp.2018.02.024
Reference:	YJCPH 7862
To appear in:	*Journal of Computational Physics*
Received date:	20 July 2017
Revised date:	23 December 2017
Accepted date:	12 February 2018

Please cite this article in press as: J. Yang et al., B-spline tight frame based force matching method, *J. Comput. Phys.* (2018), https://doi.org/10.1016/j.jcp.2018.02.024

Highlights

- We proposed an L1-regularized least squares model to form the force functions in the force matching method, which makes additional use of the B-spline wavelet tight frame. Numerical results for molecular systems involving pairwise non-bonded, three and four-body bonded interactions are obtained to demonstrate the effectiveness of our approach.
- In our approach, the B-spline tight frames system was first used for representing our force functions which has a simple explicit expression. Moreover, the redundancy of the system offers more resilience to the effects of noise and is useful in the case of lossy data.

B-spline tight frame based force matching method

Jianbin Yang[a,b], Guanhua Zhu[c], Dudu Tong[c], Lanyuan Lu[c,*], Zuowei Shen[b]

[a]*Department of Mathematics, Hohai University, Nanjing, 211100, China*
[b]*Department of Mathematics, National University of Singapore, 117543, Singapore*
[c]*School of Biological Sciences, Nanyang Technological University, 637551, Singapore*

Abstract

In molecular dynamics simulations, compared with popular all-atom force field approaches, coarse-grained (CG) methods are frequently used for the rapid investigations of long time- and length-scale processes in many important biological and soft matter studies. The typical task in coarse-graining is to derive interaction force functions between different CG site types in terms of their distance, bond angle or dihedral angle. In this paper, an ℓ_1-regularized least squares model is applied to form the force functions, which makes additional use of the B-spline wavelet frame transform in order to preserve the important features of force functions. The B-spline tight frames system has a simple explicit expression which is useful for representing our force functions. Moreover, the redundancy of the system offers more resilience to the effects of noise and is useful in the case of lossy data. Numerical results for molecular systems involving pairwise non-bonded, three and four-body bonded interactions are obtained to demonstrate the effectiveness of our approach.

Keywords: force matching, coarse-grained methods, wavelet tight frames, ℓ_1-regularized least squares

[*]Corresponding author

Email addresses: jbyang@hhu.edu.cn (Jianbin Yang), gzhu001@e.ntu.edu.sg (Guanhua Zhu), tongdudu@uchicago.edu (Dudu Tong), lylu@ntu.edu.sg (Lanyuan Lu), matzuows@nus.edu.sg (Zuowei Shen)

1. Introduction

Molecular dynamics (MD) simulation is a widely applied technique to study biomacromolecules in computational biology [1]. In its most common form, target systems are modeled at the atomistic level. The interactions between atoms are defined by some empirical force fields, which usually contain bond, angle, dihedral, van de Waals and Coulombic interactions. Then the Newton's equation of motion is solved to model the conformational changes of biomacromolecules. Despite the intensive computational resources available nowadays, the time-scale and length-scale of all-atom (AA) MD simulations are still limited, thus limiting our understanding to important biological processes. In this situation, coarse-grained (CG) models are often proposed to replace AA models [2]. In CG models, nearby atoms are grouped into a virtual CG bead. Thus fewer particles are needed to represent a target system compared with AA models. The dynamics of the system is also accelerated due to reduced degrees of freedom. As a result, both time-scale and length-scale of CG MD simulations are greatly extended.

In the so called multiscale coarse-graining methods, the interactions between CG beads are usually parameterized to fit the behavior of AA models. There are several methods in this category available in literature, including iterative inverse Boltzmann [3], inverse Monte Carlo [4], and force matching [5, 6]. The force matching method, sometimes called the multiscale coarse-graining method (MS-CG), aims at reproducing many-body potential of mean force (PMF) of atomistic configurations by fitting the total forces on the CG beads during the atomistic simulations. As proposed by Noid et al. [6], the CG forces as the fitting target correspond to the derivatives of PMFs. Force matching fits the derivative of many-body PMF through a number of over-determined equations, which can usually be solved in a least squares sense.

Reconstructing the interactions between different CG sites using atomistic simulation data can be formulated as a functional reconstruction problem [7, 8]. Assume that we are given a set of scattered data sites, i.e., the Cartesian coordinates for N CG sites in a single configuration: $\Xi = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N\} \subset \mathbb{R}^3$ and associated function values $\sum_{j\neq i} f(|\mathbf{R}_j - \mathbf{R}_i|)e^{\zeta}_{i,j} = \mathbf{f}^{\zeta}_i$, where f is the force function we want to solve, $e^{\zeta}_{i,j}$ is the component of unit vector $(\mathbf{R}_j - \mathbf{R}_i)/|\mathbf{R}_j - \mathbf{R}_i|$, and $\mathbf{f}^{\zeta}_i$ is the component of force which possibly contains noise. Our goal is then to reconstruct the force function f under the assumption that f is a piecewise smooth function. It is emphasized that the point of inflection of f need to be well preserved in the reconstruction, because they encode important information. Also, note that the input data sites Ξ are scattered, i.e. they are non-uniformly sampled, with large gaps and even sparsity. Moreover, the obtained function values $\mathbf{f}_i^\zeta$ could be very noisy. All these challenges make the reconstruction a difficult problem.

Among all available functional reconstruction methods, a regularized least squares model is one of the most widely used methods. For our force matching model, the function f is determined by solving the variational problem

$$\min_{f\in V}\sum_i(\sum_{j\neq i} f(|\mathbf{R}_j-\mathbf{R}_i|)e_{i,j}^\zeta-\mathbf{f}_i^\zeta)^2+\Gamma(f)\,, \tag{1}$$

where $\mathbf{R}_i$ and $\mathbf{R}_j$ refer to Cartesian coordinates of CG sites, and V is a function space where f is derived from. Here, the first term measures the fitting error while the regularization term $\Gamma(f)$ gives preferences to properties of the approximant f. It can for instance be chosen such that the roughness of f is penalized or such that f comes close to a piecewise continuous function.

There are several choices for the function space V in (1), often considered spaces are the Sobolev space, C^2, polynomial space or as we will use in this paper, a principal shift invariant space

$$\begin{aligned} S^h(B) := closure\{&\sum_{\alpha\in\mathbb{Z}}\mathbf{u}(\alpha)B(\frac{r}{h}-\alpha):\mathbf{u}(\alpha)\in\mathbb{R}\\ &\text{and only a finite number of }\mathbf{u}(\alpha)\neq 0,\quad \alpha\in\mathbb{Z}\},\end{aligned}$$

which is spanned by those h-dilates and h-shifts of compactly supported function $B(r)$, and $h>0$ is a scaling parameter that controls the refinement of the space. Then any function in $S^h(B)$ can be written as a finite expansion

$$f(r)=\sum_{\alpha=0}^{n-1}\mathbf{u}(\alpha)B(\frac{r}{h}-\alpha),\quad r\in\mathbb{R},$$

and our aim is to find those coefficients $\mathbf{u}(\alpha)$. In this paper, we choose B-spline function as $B(r)$. Several desirable properties that the space $S^h(B)$ enjoys motivated us to choose it as an approximation space for fitting force functions. First, it has a simple structure and provides a good approximation to smooth functions [9], which naturally leads to simple and accurate algorithms. The compact support of B-spline results in sparse system matrices which is of computational interest. Furthermore, it can be associated to a wavelet frame system and hence one can solve the data fitting problem by taking the advantages that a frame system can offer.

Inspired by some recent wavelet frame based image restoration methodologies [10, 11], we determine the approximating function $f \in S^h(B)$ by minimizing the functional

$$\min_{\mathbf{u}} \sum_i \Big(\sum_{j=1, j\neq i}^{N} f(|\mathbf{R}_j - \mathbf{R}_i|) e^{\zeta}_{i,j} - \mathbf{f}^{\zeta}_i \Big)^2 + \|\mathrm{diag}(\lambda)\mathcal{W}\mathbf{u}\|_{\ell_1}, \tag{2}$$

where $\mathbf{R}_i$ and $\mathbf{R}_j$ refer to Cartesian coordinates of CG sites, $\mathbf{u}$ are the coefficients of f, $\mathcal{W}$ is the wavelet frame transform and $\mathrm{diag}(\lambda)$ is a diagonal parameter matrix which scales the different wavelet channels. The first term in this minimization characterizes the fitting error. The second term $\|\mathrm{diag}(\lambda)\mathcal{W}\mathbf{u}\|_{\ell_1}$ suppresses noise and penalizes the roughness of the solution on one hand, and preserves the features of the resulting curves on the other hand. We will give a more detailed discussion on this in the next section.

In [12], a similar model to (2) was applied to approximate the solution of smoothing spline for fitting a curve or surface to scattered data. The model was also used in [13] to approximate range data and an asymptotic approximation analysis of the model and its minimizer was presented in [14].

Recently, Larini and Shea [15] investigated how far a system can be coarse-grained using functional forms that are commonly employed in standard atomistic simulations and assessed the impact of poor initial sampling on the quality of the resulting CG model. Das and Andersen [16] constructed hierarchical basis functions associated with the elastic net method to derive force functions. M. Maiolo et al. [17] applied Daubechies' orthogonal wavelets to represent the coarse-graining potential. More recently, Schöberl et al. [18] prescribed a probabilistic coarse-to-fine map and presented a data-driven coarse-graining scheme of atomistic ensembles in equilibrium.

The rest of this paper is organized as follows. In section 2.1 we introduce the multiscale coarse grained force matching model. To fit the force functions in B-spline function space with the wavelet smoothing method, in section 2.2 we review some properties of B-spline tight frames. Then, in section 2.3 we establish the wavelet frame based ℓ_1-regularized least squares model to derive forces between different CG sites. In section 3.1 we explain how to treat the model numerically. In the rest of section 3, we present some numerical experiments and compare our results with some other known models. Finally, conclusive remarks are given in section 4.

2. Theory Details and Mathematical Model

2.1. The Multiscale Coarse-Graining Approach

The multiscale coarse-graining method (MS-CG) aims to optimize a CG potential to reproduce many-body potential of mean force (PMF) calculated from atomistic configurations. It is developed by Voth and co-workers [7]. For an atomistic system with n atoms with coordinates ($\mathbf{r}^n = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n\}$) and force functions ($\mathbf{f}^n(\mathbf{r}^n) = \{\mathbf{f}_1(\mathbf{r}^n), \mathbf{f}_2(\mathbf{r}^n), \dots, \mathbf{f}_n(\mathbf{r}^n)\}$), a CG system is constructed by defining $N \leq n$ linear mapping operators ($\mathbf{M}_1(\mathbf{r}^n), \mathbf{M}_2(\mathbf{r}^n), \dots, \mathbf{M}_N(\mathbf{r}^n)$) that map the positions of the atomistic particles ($\mathbf{r}^n$) to CG sites ($\mathbf{R}^N = \{\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N\}$). Let $\mathbf{F}^N = \{\mathbf{F}_1(\mathbf{R}^N), \mathbf{F}_2(\mathbf{R}^N), \dots, \mathbf{F}_N(\mathbf{R}^N)\}$ be the set of CG force functions, then the MS-CG (i.e., force-matching) method aims at minimizing the difference between atomistic and CG forces through a least squares fitting [19]:

$$\min \Big\langle \sum_{I=1}^{N} (\hat{\mathbf{f}}_I(\mathbf{r}^n) - \mathbf{F}_I(\mathbf{M}_I(\mathbf{r}^n)))^2 \Big\rangle. \tag{3}$$

Here $\hat{\mathbf{f}}_I(\mathbf{r}^n) = \sum_{i \in \mathcal{I}_I} \mathbf{f}_i(\mathbf{r}^n)$, and $\mathcal{I}_I$ is the set of indices of the atomistic particles that are involved in the definition of the I-th CG site and angular brackets denote an ensemble average. For simplicity, we abbreviate (3) as

$$\min \|\mathbf{F}\mathbf{u} - \mathbf{f}\|_{\ell_2}^2, \tag{4}$$

where $\mathbf{F}$ is a matrix which is related to the input atomistic configurations, $\mathbf{f}$ is a vector composed of atomistic force data, $\mathbf{u}$ is an unknown vector containing all CG force field parameters [8]. In the MS-CG theory the optimal CG potential represented by $\mathbf{u}$ corresponds to an approximation of the many-body potential of mean force derived from an atomistic trajectory.

Solving Eq. (4) can be converted to the following normal equation

$$\mathbf{F}^T\mathbf{F}\mathbf{u} = \mathbf{F}^T\mathbf{f}, \quad or \quad \mathbf{G}\mathbf{u} = \mathbf{b}, \tag{5}$$

in which $\mathbf{G}$ denotes the square matrix $\mathbf{F}^T\mathbf{F}$ and $\mathbf{b}$ represents the vector $\mathbf{F}^T\mathbf{f}$. If the CG potential is pairwise, the MS-CG normal equation (5) is related to the well-known YBG equation in the liquid state theory [20, 6, 21], in which the two- and three- body distribution functions are connected. The dimension of the matrix $\mathbf{G}$ is determined by the total number of parameters in the molecular forces. For complex protein systems, this dimension can be in the order of 10^4 to 10^5, while for simple homogeneous liquids the size of the matrix is usually several hundreds.

In MS-CG, a pairwise CG force $f(r)$ is usually represented by a number of delta or spline functions, and the entire range of the pair distance r is divided into a number of bins accordingly. The spline functions with certain orders are preferred for higher computational efficiency. For example, the MS-CG software package MSCGFM supports both linear splines and B-splines of high order [8]. Mathematically speaking, the force functions are derived from the following principal shift invariant space

$$S^h(B) := closure\{\sum_{\alpha\in\mathbb{Z}} \mathbf{u}(\alpha)B(\frac{r}{h}-\alpha) : \mathbf{u}(\alpha)\in\mathbb{R} \\ \text{and only a finite number of } \mathbf{u}(\alpha)\neq 0,\ \ \alpha\in\mathbb{Z}\},$$

where B is a B-spline function, $h > 0$ is a dilation. The concept of shift invariant spaces arises in approximation theory, wavelet analysis, finite elements, etc. (see e.g. [22]). Besides its structural simplicity, shift invariant spaces have the beneficial property that they provide good approximation orders to smooth functions (see [9]). The compact support of B-spline results in sparse system matrices which is of computational interest. Moreover, the space gives rise to associated wavelet tight frame systems, as we shall discuss in the next subsection.

Specifically, the pairwise force at an arbitrary distance r is calculated by the formula

$$f(r) = \sum_{\alpha=0}^{n-1} \mathbf{u}(\alpha)B_m(\frac{r}{h}-\alpha), \tag{6}$$

where B_m is an m-th order B-spline basis function with the polynomial order $m-1$ and $\mathbf{u}(\alpha)$ are the corresponding coefficients we want to solve. The dilation h is determined by the atomistic data; the number n of basis functions for a pairwise force in (6) is determined by the number of break points in the chosen distance range [23].

While Eq. (6) is for pairwise distance dependent interactions, similar expressions can be obtained for other intermolecular and intramolecular coordinates such as angles and dihedrals. For instance, for a CG angular interaction, the left hand side of Eq. (6) turns to $f(\theta)$, where θ is the CG angle; and for a CG dihedral interaction, the left hand side of Eq. (6) turns to $f(\gamma)$, where γ is the CG dihedral angle.

The corresponding terms of the system potential energy function can then be obtained by integrating the expressions of the forces (6), and the CG potential energy function consists of the sum of all non-bonded and bonded components. Then, each of the CG interactions can be calculated according to (6) and the system potential energy function is the sum of these three types of interaction potentials.

If the experimental data are sufficient and noiseless, solution to (4) will fit the CG force functions well with a reasonably flexible CG potential. However, insufficient sampling and noise are unavoidable in MD simulations. In this case, solving (4) in the least squares sense cannot obtain a reasonable solution. For instance, if the sampling for the distance range is poor, i.e. there are few molecular configurations corresponding to the distance range of the support, the optimized spline coefficient will cause very large fluctuations, this is because the support of the B-spline function is compact and there exist very few or even none samples in the support of $B(\frac{r}{h} - \alpha)$ for some α.

Therefore, we want to establish a regularized method to tackle this problem. In the following, a B-spline wavelet frame based approach will be applied to fit the force functions. Before introducing this method, we review some properties of wavelet frames first.

2.2. B-spline wavelet tight frames

We present here some basics of B-spline wavelet tight frames. Interested readers should consult [24, 9] and the references therein to get a complete picture of it. A countable subset $X \subset L_2(\mathbb{R})$ is called a tight frame of $L_2(\mathbb{R})$ if

$$f = \sum_{g \in X} \langle f, g \rangle g.$$

This is equivalent to

$$\|f\|^2 = \sum_{g \in X} |\langle f, g \rangle|^2, \quad \forall f \in L_2(\mathbb{R}),$$

where

$$\langle f, g \rangle := \int_{\mathbb{R}} f(x)\overline{g(x)}dx$$

and

$$\|f\| := (\int_{\mathbb{R}} |f(x)|^2 dx)^{1/2}.$$

A wavelet system $X(\Psi)$ is defined to be a collection of dilations and shifts of a finite set of functions $\Psi = \{\psi_1, \dots, \psi_r\} \subset L_2(\mathbb{R})$, where

$$X(\Psi) := \{\psi_{\ell,j,k} := 2^{j/2}\psi_\ell(2^j x - k), \psi_\ell \in \Psi; j, k \in \mathbb{Z}\}.$$

When the countable set of functions $X(\Psi)$ forms a tight frame of $L_2(\mathbb{R})$, it is called a wavelet tight frame and each $\psi_\ell \in \Psi$ is called a framelet.

To construct wavelet tight frames, one usually starts with a refinable function ϕ satisfying

$$\phi(x) = 2\sum_{\alpha\in\mathbb{Z}} h_0(\alpha)\phi(2x - \alpha),$$

where h_0 is a finitely supported sequence called refinement mask. It is well known that B-splines are refinable. For example, the B-spline of order 1,

$$B_1(x) = \begin{cases} 1, & if \quad 0 \le x \le 1, \\ 0, & otherwise, \end{cases}$$

can be used as ϕ with $h_0 = [\frac{1}{2}, \frac{1}{2}]$. The piecewise linear B-spline

$$B_2(x) = \begin{cases} x+1, & if \quad -1 \le x \le 0, \\ -x+1, & if \quad 0 \le x \le 1, \\ 0, & otherwise, \end{cases}$$

is refinable with $h_0 = [\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$.

For a given compactly supported refinable function ϕ, the construction of a wavelet tight frame is to find an appropriate set of framelets $\Psi = \{\psi_1, \dots, \psi_r\}$ defined by

$$\psi_\ell(x) = 2\sum_{\alpha\in\mathbb{Z}} h_\ell(\alpha)\phi(2x - \alpha), \qquad \ell = 1, \dots, r, \tag{7}$$

where the framelet masks h_ℓ are finitely supported sequences.

By using B-spline as the refinable function ϕ, a family of wavelet tight frame system is derived by the Unitary Extension Principle (UEP) [24]. For example, the simplest system in this family is piecewise linear B-spline tight frame which uses B_2 as ϕ and two framelets ψ_1 and ψ_2 as defined in (7) with

$$h_1 = \frac{\sqrt{2}}{4}[-1, 0, 1], \quad h_2 = \frac{1}{4}[-1, 2, -1].$$

The plot of ϕ, ψ_1, ψ_2 is given in Fig 1 (a).

A smoother wavelet tight frame system is the cubic B-spline tight frame which uses B_4 as ϕ with $h_0 = [\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}]$. Define h_1, h_2, h_3, h_4 as follows:

$$h_1 = [\frac{1}{16}, -\frac{1}{4}, \frac{3}{8}, -\frac{1}{4}, \frac{1}{16}], \quad h_2 = [-\frac{1}{8}, \frac{1}{4}, 0, -\frac{1}{4}, \frac{1}{8}],$$
$$h_3 = [\frac{\sqrt{6}}{16}, 0, -\frac{\sqrt{6}}{8}, 0, \frac{\sqrt{6}}{16}], \quad h_4 = [-\frac{1}{8}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{8}]. \tag{8}$$

Then the system $X(\Psi)$ where $\Psi = \{\psi_1, \psi_2, \psi_3, \psi_4\}$ defined in (7) by h_1, h_2, h_3, h_4 above is a tight frame of $L_2(\mathbb{R})$ (see Fig. 1 (b)). Other constructions of wavelet tight frames from any B_m can be found in [24].

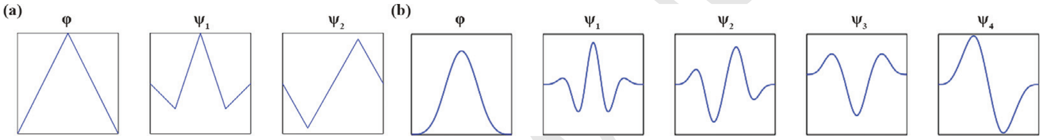

Figure 1: Refinable functions and framelets. (a) Piecewise linear spline and corresponding framelets (b) Piecewise cubic spline and corresponding framelets

Let $V_0 = \overline{span\{\phi(\cdot - k), k \in \mathbb{Z}\}}$. The dilations $\phi_{1,k} := 2^{1/2}\phi(2\cdot -k), k \in \mathbb{Z}$, form a Riesz basis for a space $V_1 \supset V_0$. In fact, we have $\cdots \supset V_1 \supset V_0 \supset V_{-1} \supset \cdots$ and $\overline{\cup_j V_j} = L_2(\mathbb{R})$, where each V_j is spanned by $\phi_{j,k} := 2^{j/2}\phi(2^j \cdot -k), k \in \mathbb{Z}$. By [25], for any give $L \in \mathbb{Z}$, for $f \in L_2(\mathbb{R})$, we have

$$f = \sum_{k\in\mathbb{Z}} \langle f, \phi_{L,k}\rangle \phi_{L,k} + \sum_{\ell=1}^{r} \sum_{j\geq L} \sum_{k\in\mathbb{Z}} \langle f, \psi_{\ell,j,k}\rangle \psi_{\ell,j,k},$$

where $\psi_{\ell,j,k} := 2^{j/2}\psi_\ell(2^j \cdot -k)$.

Here, we represent a function by a component in V_L plus the component in W_j, written as $L_2(\mathbb{R}) = V_L + \cup_{j\geq L} W_j$, where W_j is spanned by $\psi_{\ell,j,k}$, $1 \leq \ell \leq r$, $k \in \mathbb{Z}$. The component in V_L represents rough components of f and the component in W_j represents the detail, and we might wish to set some elements of the latter component to zero by shrinking and selecting the coefficients toward a sparse representation [26].

The difference between B-spline tight frames and Daubechies' s orthonormal wavelet basis [27] is the follows: first, the scaling function B_m of B-spline tight frames has very simple expression which is easy to generate the data matrix, while the scaling function of orthonormal wavelet basis has no explicit formula except Haar. Second, the B-spline tight frames are a redundant system, thus it offers more resilience to the effects of noise, and is useful especially in the case of lossy data.

B-spline tight frames are very popular in signal and image processing, since they are able to represent both smooth and/or locally bumpy functions in an efficient way and provide time and frequency localization [22]. The effectiveness of the B-spline tight frame has been proved in many applications in signal and image processing [28, 13, 10, 11]. We will show that such a simple system can also be used to effectively reconstruct the interaction functions from MD simulation data.

Numerical computation of the wavelet frame transform is done by using the wavelet frame decomposition algorithm [22, 25, 11]. Let h_i, $0 \leq i \leq r$ be the framelet masks. For the $(\ell+1)$-th level of wavelet frame transform, the filters are defined by $h_{\ell,i} := \tilde{h}_{\ell,i} * \tilde{h}_{\ell-1,i} * \ldots * \tilde{h}_{0,i}$, where

$$\tilde{h}_{\ell,i}[k] = \begin{cases} h_i[2^{-\ell}k], & k \in 2^{\ell}\mathbb{Z}^d, \\ 0, & k \notin 2^{\ell}\mathbb{Z}^d. \end{cases}$$

Then, the discrete framelet transform without down-sampling are defined by $\mathcal{W}_{\ell,i}\mathbf{u} := h_{\ell,i}[-\cdot] * \mathbf{u}$, where $\mathbf{u} \in \ell_1(\mathbb{Z})$ and $*$ is the discrete convolution operator. We denote the discrete framelet transform with L levels of decomposition as

$$\mathcal{W}\mathbf{u} = \{\mathcal{W}_{\ell,i}\mathbf{u} : 0 \leq \ell \leq L-1, 0 \leq i \leq r\}.$$

The transform can be represented by a matrix whose construction depends on the boundary conditions. We omit the detailed discussions here and the interested reader should consult [25] for more details.

2.3. Wavelet Smoothing Model

In this section, we propose a B-spline tight wavelet frame based ℓ_1-regularized model to fit CG forces. Let $f : \mathbb{R} \to \mathbb{R}$ be a force function we want to solve. In MS-CG, for pairwise distance dependent interactions, the position of each CG site $(\mathbf{R}_i = (x_i, y_i, z_i))_{i=1}^{N}$ is obtained by the centers of geometry of the corresponding atoms, and $(|\mathbf{R}_j - \mathbf{R}_i|)_{i,j=1}^{N}$ are considered as our input variables.

We approximate f between different CG sites in terms of Euclidean distance r from $S^h(B_m)$. That is, let

$$f(r) = \sum_{\alpha} \mathbf{u}(\alpha) B_m(\frac{r}{h} - \alpha), \tag{9}$$

where the range of α is determined by the number of break points in the chosen distance range, and the dilation h is determined by the density of input variables. The force function f will be given once the coefficients $\mathbf{u}(\alpha)$ are determined. By the Boltzmann distribution in statistical mechanics [1], most of the CG sites are located in a low energy state, whereas only a few CG sites are located in the high energy state. Moreover, due to inadequate sampling of phase space and random fluctuations in the measurements, the data $\mathbf{R}_i$ and resultant force $\mathbf{f}_i$ are usually noisy. These factors make it challenge to derive force functions.

Let $e^x_{i,j} := \frac{x_j - x_i}{|\mathbf{R}_j - \mathbf{R}_i|}$, $e^y_{i,j} := \frac{y_j - y_i}{|\mathbf{R}_j - \mathbf{R}_i|}$, $e^z_{i,j} := \frac{z_j - z_i}{|\mathbf{R}_j - \mathbf{R}_i|}$ and $\{|\mathbf{R}_j - \mathbf{R}_i|\}^N_{i,j=1}$ be considered as our input variables, where (x_i, y_i, z_i) is the Cartesian coordinate of $\mathbf{R}_i$. The resultant force $\mathbf{f}_i = (\mathbf{f}^x_i, \mathbf{f}^y_i, \mathbf{f}^z_i)$ which is acting on each coarse site W_i is calculated as the sum of corresponding atomistic forces, $W_j (j \neq i)$ acting upon W_i. We determine the coefficients $\mathbf{u}(\alpha)$ in (9) by solving

$$\min_{\mathbf{u}} \sum_i \sum_{\zeta \in \{x,y,z\}} \Big(\sum_{j=1, j\neq i}^{N} f(|\mathbf{R}_j - \mathbf{R}_i|) e^{\zeta}_{i,j} - \mathbf{f}^{\zeta}_i \Big)^2 + \|\mathrm{diag}(\lambda) \mathcal{W} \mathbf{u}\|_{\ell_1}, \tag{10}$$

where $\mathcal{W}$ is the wavelet frame transform and $\mathrm{diag}(\lambda)$ is a diagonal parameter matrix which scales the different wavelet channels. Here, for a sequence $\mathbf{a}$, $\|\mathbf{a}\|_{\ell_1} := \sum_\alpha |\mathbf{a}(\alpha)|$. If $\mathbf{u}^*$ is the minimizer of (10), then we have

$$f(r) = \sum_{\alpha \in I} \mathbf{u}^*(\alpha) B_m(\frac{r}{h} - \alpha).$$

In the numerical experiments in section 3.2, we choose the cubic spline ($m = 4$), i.e.

$$B_4(x) = \begin{cases} x^3/6 & \text{if } 0 \leq x < 1 \\ (-3x^3 + 12x^2 - 12x + 4)/6 & \text{if } 1 \leq x < 2 \\ (3x^3 - 24x^2 + 60x - 44)/6 & \text{if } 2 \leq x < 3 \\ (4 - x)^3/6 & \text{if } 3 \leq x < 4 \\ 0 & \text{else} \end{cases}$$

and its associated wavelet frame transform $\mathcal{W}$ with the corresponding masks h_1, h_2, h_3, h_4 in (8).

When all parameters λ are chosen zero, the model (10) is the usual least squares fitting, corresponding to (4). The idea behind the regularization term in (10) is to make use of the interaction between the framelet transform and the ℓ_1-norm. Here, the force functions are derived from a B-spline function space V_J $(h = 2^{-J}$) with certain high resolution which can be decomposed into a coarse resolution space and wavelet frame subspaces [24, 22, 9]. It is well known that the wavelet frame coefficient of a signal, which is sampled from a piecewise smooth function, is sparse (i.e. large number of wavelet frame coefficients are equal or close to zero and negligible, see [22]). Thus, for a large class of functions, we can get a good approximation by neglecting small coefficients [26]. Furthermore, since the ℓ_1-norm minimization annihilates small coefficients in the wavelet frame domain, the regularization term $\|\text{diag}(\lambda)\mathcal{W}\mathbf{u}\|_{\ell_1}$ gives preference to a solution whose wavelet coefficient sequence is sparse, and to keep important features of functions.

While model (10) is for distance dependent interactions, similar models can be obtained for other intermolecular and intramolecular coordinates such as angles and dihedrals. For a general system, for a CG angular interaction, $f(r)$ turns to $f(\theta)$, where θ is the CG angle; and for a CG dihedral interaction, $f(r)$ turns to $f(\gamma)$, where γ is the CG dihedral angle. Let f_k, $k = 1, \ldots, K$ be those force functions in a system we want to solve. Then the model (10) becomes

$$\min_{\mathbf{u}} \sum_i \sum_{\zeta \in \{x,y,z\}} \Big(\sum_{k=1}^{K} \sum_{j=1}^{N_k} f_k(X_i^j) e_{i,j}^{\zeta} - \mathbf{f}_i^{\zeta} \Big)^2 + \|\text{diag}(\lambda)\mathcal{W}\mathbf{u}\|_{\ell_1}.$$

Here, $\{X_i^j\}$ are the scalar variables (i.e., distance, bond length or dihedral angle), $f_k = \sum_\alpha \mathbf{u}_k(\alpha) B_m(\frac{\cdot}{h_k} - \alpha)$ and $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K)^T$ are the vector of sequences we want to solve.

In the next section we explain how to treat the minimization problem (10) numerically, and we expect that our method will be effective for the realistic cases with poor sampling, which has been challenging in previous studies.

3. Numerical Results and Discussions

3.1. Numerical algorithms

In this section we explain how to solve the optimization problem (10). This problem can be written in the matrix vector form as

$$\min_{\mathbf{u}} \|\mathbf{Fu} - \mathbf{f}\|_{\ell_2}^2 + \|\mathrm{diag}(\lambda)\mathcal{W}\mathbf{u}\|_{\ell_1} \,, \tag{11}$$

where $\mathbf{f} = [\cdots, \mathbf{f}_i^{\zeta}, \cdots]^T$, $\mathbf{F}_{ij} = \sum_{\ell=1,\ell\neq i}^{N} B_m(\frac{|\mathbf{R}_\ell - \mathbf{R}_i|}{h} - k_j)$ and $I = \{k_1, \ldots, k_n\}$. Thus, (11) is an ordinary least squares problem with an ℓ_1-regularization term. The alternating direction method of multipliers (ADMM) method [29, 28] has been proved to be very efficient in solving (11) with various successful applications, see e.g. [29, 28, 13].

Note that (11) is equivalent to

$$\min_{\mathbf{u},\mathbf{d}} \|\mathbf{Fu} - \mathbf{f}\|_{\ell_2}^2 + \|\mathrm{diag}(\lambda)\mathbf{d}\|_{\ell_1} \text{ subject to } \mathbf{d} = \mathcal{W}\mathbf{u}. \tag{12}$$

Then the ADMM algorithm that solves (12) is as follows:

$$\begin{cases} \mathbf{u}^{i+1} = \arg\min_{\mathbf{u}} \|\mathbf{Fu} - \mathbf{f}\|_{\ell_2}^2 + \dfrac{\mu}{2}\|\mathcal{W}\mathbf{u} - \mathbf{d}^i + \mathbf{b}^i\|_{\ell_2}^2 & (13) \\ \mathbf{d}^{i+1} = \arg\min_{\mathbf{d}} \|\mathrm{diag}(\lambda)\mathbf{d}\|_{\ell_1} + \dfrac{\mu}{2}\|\mathbf{d} - \mathcal{W}\mathbf{u}^i - \mathbf{b}^i\|_{\ell_2}^2 & (14) \\ \mathbf{b}^{i+1} = \mathbf{b}^i + \mathcal{W}\mathbf{u}^{i+1} - \mathbf{d}^{i+1} & (15) \end{cases}$$

with initial $\mathbf{u}^0 = 0$, $\mathbf{d}^0 = 0$ and $\mathbf{b}^0 = 0$. In the following computations, we choose $\mu = 0.1$.

The solution to (13) is determined by solving the system of equations

$$(2\mathbf{F}^T\mathbf{F} + \mu\mathcal{W}^T\mathcal{W})\mathbf{u} = 2\mathbf{F}^T\mathbf{f} + \mu\mathcal{W}^T(\mathbf{d}^i - \mathbf{b}^i)$$

which, because of $\mathcal{W}^T\mathcal{W} = I$, can be simplified to

$$(2\mathbf{F}^T\mathbf{F} + \mu\mathbf{I})\mathbf{u} = 2\mathbf{F}^T\mathbf{f} + \mu\mathcal{W}^T(\mathbf{d}^i - \mathbf{b}^i). \tag{16}$$

Since $(2\mathbf{F}^T\mathbf{F} + \mu\mathbf{I})$ is symmetric positive definite, the system of equations (16) can be efficiently solved by applying the conjugate gradient method. The solution to (14) is given by

$$\mathbf{d}^{i+1} = T_{\lambda/\mu}(\mathcal{W}\mathbf{u}^{i+1} + \mathbf{b}^i).$$

For $\mu > 0$, $T_{\lambda/\mu}$ is the soft-threshold operator

$$T_{\lambda/\mu}(\mathbf{x}) := [t_{\lambda/\mu}(x_1), t_{\lambda/\mu}(x_2), \ldots, t_{\lambda/\mu}(x_M)] \,,$$

with $t_{\lambda/\mu}(x_i) := \operatorname{sgn}(x_i) \max\{0, |x_i| - \frac{\lambda}{\mu}\}$.

Further note, that $\mathcal{W}^T(\mathbf{d}^i - \mathbf{b}^i)$ in (16) is determined by performing the inverse framelet transform rather than by using its matrix representation, similar in the iterations (14) and (15). The stoping criteria of the iteration is

$$\|\mathbf{d}^i - \mathcal{W}\mathbf{u}^i\|_{\ell_2} \leq \epsilon$$

for some positive constant ϵ.

In order to highlight the effectiveness of our wavelet smoothing method, in the next part, we will show some numerical results and compare our method with the Tikhonov regularization method, which is defined as

$$\min_{\mathbf{u}} \|\mathbf{F}\mathbf{u} - \mathbf{f}\|^2_{\ell_2} + \nu \|\mathbf{u}\|^2_{\ell_2}, \tag{17}$$

and the Laplacian regularization method, i.e.,

$$\min_{\mathbf{u}} \|\mathbf{F}\mathbf{u} - \mathbf{f}\|^2_{\ell_2} + \nu \mathbf{u}^* \mathcal{L} \mathbf{u} \tag{18}$$

with $\mathcal{L}$ the discrete Laplacian operator.

3.2. Numerical Results

3.2.1. One Site Coarse-Grained Water

The atomistic simulation of 999 water molecules contained in a cubic box of size 3.111 nm was carried out using the OPLS-AA [30, 31] all atom force field at 300K. In the atomistic simulations, hydrogens are constrained using Lincs [32]. The thermostat used is the velocity rescaling method in Gromacs [33] with 0.1 ps coupling constant. The SPC/E [34] water model was used. Totally 10 ns simulation was conducted under the constant NVT condition. The integration step was set to 2 fs. The force matching method was applied to the atomistic configurations to generate the CG potentials. In the CG model three atoms in one water molecule were combined to one singe CG-site “W”.

In Fig. 2 the curve of fitting with sufficient data can be seen as a benchmark, which was based on 10^5 frames of the trajectory data. For 30 frames of the trajectory data, we applied the wavelet smoothing model (10) with

h = 0.005 nm to derive the force function. It can be seen that compared with the Tikhonov (17) and Laplacian regularization (18) methods, our approach preserves the minima better, which is important for CG modeling. To assess the statistic errors in our modeling, we calculated 5 independent numerical results and each experiment randomly sampled 30 frames throughout the trajectory. The unbiased standard deviation of these 5 results around the force minima point is 0.8522.

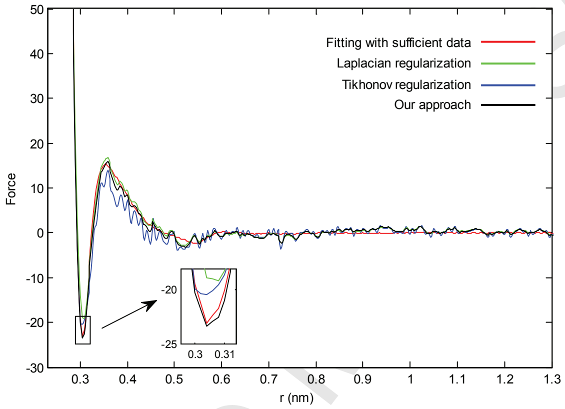

Figure 2: Approximation of interaction force (kJ/(mol nm)) of Water-Water molecules by 30 frames

3.2.2. Angular potential from a single propane molecule

A single propane molecule was simulated in vacuum condition using the Amber-99SB [35] force field at 300 K. The time step was set to 2 fs. Totally 4 us Langevin dynamic simulation was performed and in Langevin dynamics, the coupling constant was 0.02 ps. Here, CG sites were placed at C atoms, as shown in Fig. 3. The beads were positioned at C atoms instead of centers of mass, because this coarse-grained treatment permit a direct comparison of its bond and angle distributions with AA simulations whose bonds and angles are physical connections.

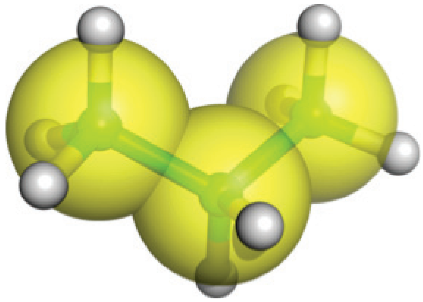

Figure 3: CG representation of a single propane molecule

Totally, we had 2×10^6 frames. In the CG model, the bond was fitted to a harmonic potential by using all the frames. The angle term was subjected to a tabulated potential by using different number of frames. By using all frames to fit the angle, the CG simulation produced an ensemble of conformations that was able to fully reproduce the angle distribution of all-atom simulations (see Fig. 4). With less frames, the force and energy functions became less smooth with more noise. If the number of frames was reduced to 850, the force and energy would be too noisy to reproduce the angle distribution with all frames. In this case, though the peak position in the angle distribution was correctly located, the peak intensity deviated dramatically. By using our method, the force can be recovered as a smooth pattern and the angle distribution was recovered significantly, closer to the all-frame pattern (see Fig. 5). If fewer frames were used, the force would be noisier even with very large values (see Fig. 6) and the angle distribution would deviate further away from that of the all-frame case. This case with noise and very large forces would produce a random pattern of angle distribution. The exclusion of outliers (forces larger than 1000) could drive the angle distribution from the random pattern, but still failed to display the correct distribution. Our wavelet smoothing model can remove the side-effect of the noise caused by insufficient frames, successfully reproducing the correct angle distribution. Besides, we also found that the bond distribution in CG simulations was also strongly affected by the force noise of angle. Our approach also remarkably improved the bond distribution compared with the classical least squares method (see Fig. 7).

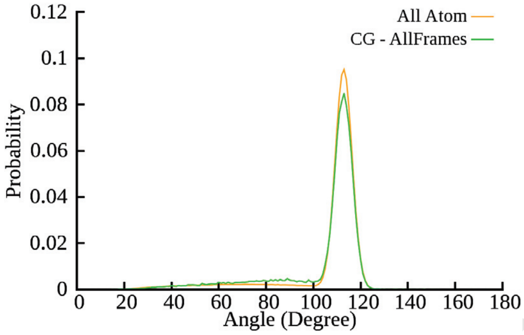

Figure 4: A comparison of angle distributions between all-atom simulations and CG simulations of all frames (2×10^6 frames).

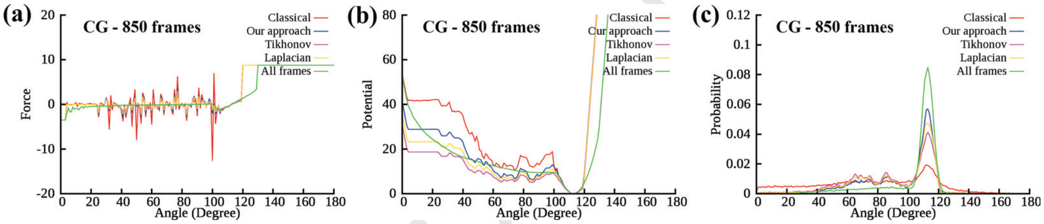

Figure 5: 850 fames of propane molecules. (a) force (kJ/(mol degree)). (b) potential (kJ/mol). (c) angle distribution. Results by the classical least squares method, our approach, Tikhonov and Laplacian regularizations with 850 frames, and the classical least squares method with all the 2×10^6 frames.

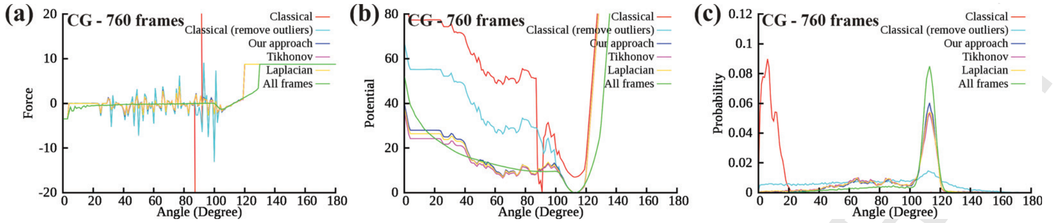

Figure 6: 760 frames of propane molecules. (a) force (kJ/(mol degree)). (b) potential (kJ/mol). (c) angle distribution. Results by the classical least squares method, the classical least squares method removing outliers (refer to the main text), our approach, Tikhonov and Laplacian regularizations with 760 frames, and the classical least squares method with all the 2×10^6 frames.

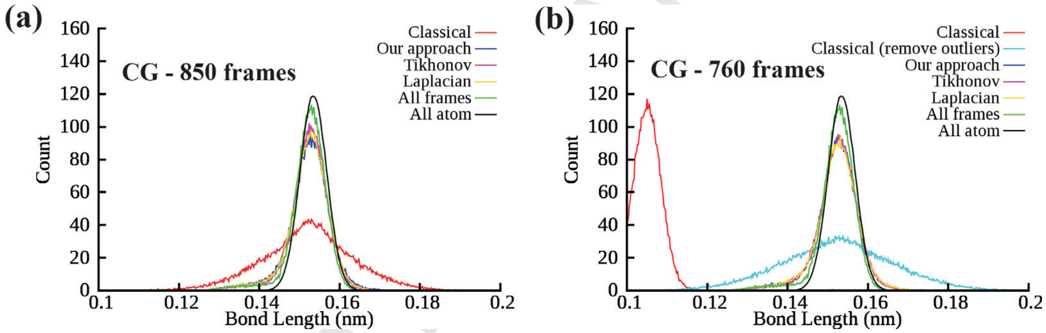

Figure 7: A comparison of bond distributions between different methods, with 850 frames (a) and 760 frames (b) used.

3.2.3. Hexane

The atomistic simulation of 600 hexane molecules contained in a cubic box of size 5.08 nm was carried out using the OPLS-AA all atom force field at 300 K. Totally 10 ns simulation was conducted under the constant NVT condition. The integration step was set to 2 fs. In the atomistic simulations, hydrogens are constrained using Lincs. The time constant was 0.1 ps using the V-rescale method in Gromacs that is a modified Berendsen thermostat [36]. Two types of CG sites of hexane were implemented, namely CA and CB (see Fig. 8). The total number of CG sites is 2400.

A total of 5000 frames were saved from MD simulations and each frame was a box of 600 hexane molecules. CG MD simulations were performed to investigate whether our method is able to improve the accuracy of CG force field of the dihedral angle when the sampling is insufficient. In the CG model, the bond and angle potentials were fitted to harmonic potentials and non-bonded interactions were fitted as tabulated potentials. For the dihedral angle, if less frames were used, the angle distribution would deviate away from that of the all-frame case (see Fig. 9 (c)). It is noted that, a symmetric treatment was subjected to the CG dihedral force that was derived from all-atom simulations because of the symmetric nature of dihedral angles. As shown in Fig. 9 (a), using 6 frames undermines the accuracy of force, compared with the force from all 5000 frames, especially an opposite trend around $-50 \sim +50°$ that was a high-energy region with less sampled data. As a result, this bad-sampling case failed to capture the high-energy peak of the dihedral potential at $-35 \sim +35°$ (see Fig. 9 (b)). The CG MD simulation of this insufficient sampling case produced a dihedral angle distribution that deviate from the result from all-atom simulations (see Fig. 9 (c)). The inaccurate area of dihedral distribution was around 0° caused by the inaccurate CG force of bad sampling. In comparison, our method can improve the force and potential accuracy at worse-sampled dihedral range around $-50 \sim +50°$, consequently recovering a correct pattern of dihedral angle distribution as the all-atom one.

Fig. 9 shows that our approach can preserve the force function well, while the Tikhonov regularization (17) reduces the energy of force function. The reason behind this is that for the Tikhonov method, the model minimizes $\|\mathbf{u}\|_{\ell_2}^2$ and causes the force curves to deviate to x-axis. In our approach, a redundant wavelet frame system was used to represent functions, but then shrink and select the coefficients toward a sparse representation. This wavelet smoothing method fit represents locally bumpy functions and nicely localizes the spikes.

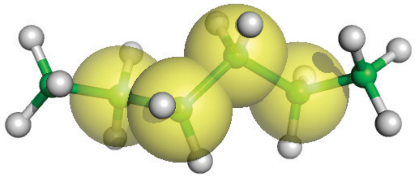

Figure 8: CG representation of hexane molecule. Four CG beads are CA-CB-CB-CA.

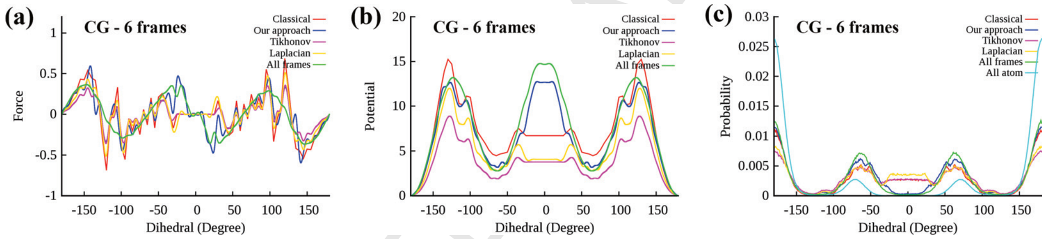

Figure 9: The dihedral angles of hexane in the case of 6 frames. (a) and (b) are force (kJ/(mol degree)) and potential (kJ/mol) respectively for the classical least squares method (red), our approach (blue), Tikhonov (magenta) and Laplacian (gold), in comparison with the counterpart of all the 5000 frames (green). The dihedral distribution in (c) also shows the dihedral distribution from an all-atom MD simulation (cyan).

3.3. Discussions

The normal equation (5) was proposed by W. G. Noid et al. [6, 7] to minimize the MS-CG residual. When the data are sufficient and noiseless, the solution can approximate the CG force functions well. When the data are poorly sampling or noisy, as it often happens, the regularized models are preferred to derive the force functions.

The Tikhonov regularization model (17) which minimize the ℓ_2-norm of $\mathbf{u}$ will reduce the energy of underlying solution. This causes the force curves to deviate to x-axis (see Fig. 9). The idea in the Laplacian regularization method (18) is to restrain the fluctuations of neighbouring $\mathbf{u}$ which gives preference to a smooth solution. However, this method will erase some features and details of the curve which is important in force functions (see e.g. [14] and Fig. 2). Therefore, the ℓ_1-regularized term is preferred in our model. In addition, when the sampling is very poor, for example, no samples lie in several adjacent support of $B_m(\frac{r}{h}-\alpha)$, high-pass filters with longer support than Laplacian operator are desirable. In this case, the wavelet frame transform with long supported masks and multiscale transform can be chosen.

In our approach the B-spline functions and associated wavelet tight frames are used to derive the force functions between different CG sites. The Daubechies orthogonal wavelet functions ϕ and ψ are not considered [27, 17], since they are non-symmetric and have no explicit expression. Furthermore, wavelet frame, as a generalization to orthonormal basis, relaxes the requirement of $X(\Psi)$ being a basis and brings in redundancy, while the redundancy offers more resilience to the effects of noise, provides more numerical stability, and is useful in the case of lossy data.

The support of the B-spline functions $B_m(\frac{r}{h}-\alpha)$ or scaling parameter h is determined by the density and noise level of experimental data. A large support B-spline functions can fit the lossy data, but it fails to represent the details of force functions which is important in MD simulations. Moreover, due to the large support, the noisy data at r_1 can change the behavior of $f(r_2)$, even those r_2 far away from r_1. This property is undesirable for most applications. On the contrary, a small support B-spline functions can detect the details of force functions, but the curve fluctuates in the poor sampling case. Therefore, multiresolution hierarchical basis functions are preferred to represent the force functions.

Recently, P. Liu et al. [37] presented a Bayesian statistics approach to improve the MS-CG force field obtained from the CG model. Our wavelet smoothing model can also be cast in a Bayesian framework. The fitting term of (10) corresponds to the Gaussian distribution of sample noise, and the regularization term corresponds to the Laplace distribution of wavelet coefficients. This optimization problem amounts to finding the posterior mode and usually can be solved by fast algorithm and be constructed for functions in any dimension. In the approach of Das and Andersen [16], they constructed hierarchical basis functions associated with the elastic net method to derive force functions. M. Maiolo et al. [17] applied Daubechies scaling functions and orthogonal wavelets to approximate the MSCG. The idea of both methods is to approximate functions from a relatively coarser resolution subspace plus the fluctuated detail subspaces. On the contrary, our model is to approximate f from the high resolution space directly, which leads to generating the matrix $\mathbf{F}$ easily. In addition, we assume that the wavelet frame coefficients of f is sparse, i.e. most of the coefficients of wavelet frame transform of f are negligible. Thus, the threshold procedure should be given after the wavelet frame transform. In [16], certain hierarchical basis functions should be carefully constructed, and in [17] subdivision scheme [27] should be applied to interpolate the scaling function values, since Daubechies scaling functions and wavelets functions have no explicit formula. Compared to the construction in [16, 17], our approach has the double advantages of being multi-resolution hierarchical basis functions and being easier in the implementation. The structure of B-spline principle shift invariant space $S^h(B_m)$ is very simple and the tight wavelet frame transform and inverse transform can be implemented very fast. Furthermore, the B-spline tight frame system has a simple explicit formula and can reconstruct any f in L_2 space in theory.

4. Conclusion and Future work

In this paper, we proposed an ℓ_1-regularized least squares force matching method based on the wavelet frame transform in order to preserve the important features of the force functions and suppress noise. The force functions were derived from a B-spline function space with certain high resolution which can be decomposed into a coarser resolution B-spline function space and wavelet frame subspaces. Here, the wavelet frame system has simple explicit expression which is useful for representing our force functions, and we expect that the wavelet coefficients of the underlying functions are sparse. Furthermore, the redundancy of the frame system offers more resilience to the effects of noise, and is useful especially in case of lossy data.

In the future work, it is of interest to implement our approach on protein structure data from experimental database, as proposed by Mullinax and Noid [21, 38], since the experimental data are in general more noisy than those from MD simulations.

Acknowledgment

The work of Jianbin Yang was partially supported by the research grant 11771120 from NSFC and the Fundamental Research Funds for the Central

Universities 2015B19514, China; the work of Lanyuan Lu and Zuowei Shen was partially supported by Singapore MOE Research Grant MOE2014-T2-1-065 and MOE2012-T3-1-008. The computations were supported in part by National Supercomputing Centre (NSCC) Singapore.

References

[1] D. Frenkel, B. Smit, Understanding molecular simulation: from algorithms to applications, Academic press, 2001.

[2] G.A. Voth, Coarse-graining of condensed phase and biomolecular systems, CRC Press: Boca Raton, 2008.

[3] D. Reith, M. Pütz, F. Müller-Plathe, Deriving effective mesoscale potentials from atomistic simulations, J. Comput. Chem. 24 (13) (2003) 1624-1636.

[4] A.P. Lyubartsev, A. Laaksonen, Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach, Phys. Rev. Lett. 52 (4) (1995) 3730-3737.

[5] S. Izvekov, G.A. Voth, Multiscale coarse-graining method for biomolecular systems, J. Phys. Chem. B 109 (7) (2005) 2469-2473.

[6] W.G. Noid, J.W. Chu, G.S. Ayton, V. Krishna, S. Izvekov, G.A. Voth, A. Das, H.C. Andersen, The multiscale coarse-graining method, I. A rigorous bridge between atomistic and coarse-grained models, J. Chem. Phys. 128 (24) (2008) 244114.

[7] W.G. Noid, P. Liu, Y. Wang, J.W. Chu, G.S. Ayton, S. Izvekov, H.C. Andersen, G.A. Voth, The multiscale coarse-graining method, II. Numerical implementation for coarse-grained molecular models, J. Chem. Phys. 128 (24) (2008) 244115.

[8] L. Lu, S. Izvekov, A. Das, H.C. Andersen, G.A. Voth, Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining, J. Chem. Theory Comput. 6 (3) (2010) 954-965.

[9] I. Daubechies, B. Han, A. Ron, Z. Shen, Framelets: MRA-based constructions of wavelet frames, Appl. Comput. Harmon. Anal. 14 (1) (2003) 1-46.

[10] Z. Shen, Wavelet frames and image restorations. Proceedings of the International Congress of Mathematicians, Hyderabad, India, 2010.

[11] J.F. Cai, B. Dong, S. Osher, Z. Shen, Image restoration: total variation, wavelet frames, and beyond, J. Amer. Math. Soc. 25 (4) (2012) 1033-1089.

[12] M.J. Johnson, Z. Shen, Y.H. Xu, Scattered data reconstruction by regularization in B-spline and associated wavelet spaces, J. Approx. Theory. 159 (2) (2009) 197-223.

[13] H. Ji, Z. Shen, Y.H. Xu, Wavelet frame based scene reconstruction from range data, J. Comput. Phys. 229 (6) (2010) 2093-2108.

[14] J. Yang, D. Stahl, Z. Shen, An analysis of wavelet frame based scattered data reconstruction, Appl. Comput. Harmon. Anal. 42 (3) (2017) 480–507.

[15] L. Larini, J.E. Shea, Coarse-grained modeling of simple molecules at different resolutions in the absence of good sampling, J. Phys. Chem. B 116 (29) (2012) 8337-8349.

[16] A. Das, H.C. Andersen, The multiscale coarse-graining method, VIII. Multiresolution hierarchical basis functions and basis function selection in the construction of coarse-grained force fields, J. Chem. Phys. 136 (2012) 194113.

[17] M. Maiolo, A. Vancheri, R. Krause, A. Danani, Wavelets as basis functions to represent the coarse-graining potential in multiscale coarse graining approach, J. Comput. Phys. 300 (2015) 592-604.

[18] M. Schöberl, N. Zabaras, P.S. Koutsourelakis, Predictive coarse-graining, J. Comput. Phys. 333 (2017) 49-77.

[19] A. Das, L. Lu, H.C. Andersen, G.A. Voth, The multiscale coarse-graining method. X. Improved algorithms for constructing coarse-grained potentials for molecular systems, J. Chem. Phys. 136 (19) (2012) 194115.

[20] J.-P. Hansen, I.R. McDonald, Theory of Simple Liquids, 3rd ed. Academic Press, Amsterdam, 2006.

[21] J.W. Mullinax, W.G. Noid, Generalized Yvon-Born-Green Theory for Molecular Systems, Phys. Rev. Lett. 103 (19) (2009) 198104.

[22] S. Mallat, A wavelet tour of signal processing: the sparse way, Academic press, 2008.

[23] L. Lu, J.F. Dama, G.A. Voth, Fitting coarse-grained distribution functions through an iterative force-matching method. J. Chem. Phys. 139 (12) (2013) 121906.

[24] A. Ron, Z. Shen, Affine systems in $L_2(\mathbb{R}^d)$: The analysis of the analysis operator, J. Funct. Anal. 148 (2) (1997) 408-447.

[25] B. Dong, Z. Shen, MRA Based Wavelet Frames and Applications, IAS Lecture Notes Series, Summer Program on The Mathematics of Image Processing, Park City Mathematics Institute, 2010.

[26] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, biometrika, 81 (3) (1994) 425-455.

[27] I. Daubechies, Ten lectures on wavelets, Philadelphia: Society for industrial and applied mathematics, 1992.

[28] J.F. Cai, S. Osher, Z. Shen, Split Bregman methods and frame based image restoration, Multiscale Model. Simul. 8 (2) (2009) 337-369.

[29] T. Goldstein, S. Osher, The split Bregman method for L1-regularized problems, SIAM J. Im. Sc. 2 (2) (2009) 323-343.

[30] W.L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, J. Am. Chem. Soc. 110(6) (1988) 1657-1666.

[31] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, J. Am. Chem. Soc. 118(45) (1996) 11225-11236.

[32] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, LINCS: A linear constraint solver for molecular simulations, J. Comput. Chem. 18 (12) (1997) 1463-1472.

[33] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation, J. Chem. Theory Comput. 4 (3) (2008) 435-447.

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

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

[36] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. 81 (8) (1984) 3684.

[37] P. Liu, Q. Shi, H. Daumé III, G.A. Voth, A Bayesian statistics approach to multiscale coarse graining, J. Chem. Phys. 129 (2008), 214114.

[38] J.W. Mullinax, W.G. Noid, A Generalized-Yvon-Born-Green Theory for Determining Coarse-Grained Interaction Potentials, J. Phys. Chem. C 114 (12) (2010) 5661-5674.