Theory of wavelet-based coarse-graining hierarchies for molecular dynamics

Berend Christopher Rinderspacher,[1] Jaydeep P. Bardhan,[2] and Ahmed E. Ismail[3,4]

[1]*Weapons and Materials Directorate, United States Army Research Laboratory, Adelphi, Maryland 20783-1138, USA*

[2]*Department of Mechanical and Industrial Engineering, Northeastern University, Boston, Massachusetts 02115, USA*

[3]*Faculty of Mechanical Engineering, RWTH Aachen University, Aachen, Germany*

[4]*Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, West Virginia 26505, USA*

We present a multiresolution approach to compressing the degrees of freedom and potentials associated with molecular dynamics, such as the bond potentials. The approach suggests a systematic way to accelerate large-scale molecular simulations with more than two levels of coarse graining, particularly applications of polymeric materials. In particular, we derive explicit models for (arbitrarily large) linear (homo)polymers and iterative methods to compute large-scale wavelet decompositions from fragment solutions. This approach does not require explicit preparation of atomistic-to-coarse-grained mappings, but instead uses the theory of diffusion wavelets for graph Laplacians to develop system-specific mappings. Our methodology leads to a hierarchy of system-specific coarse-grained degrees of freedom that provides a conceptually clear and mathematically rigorous framework for modeling chemical systems at relevant model scales. The approach is capable of automatically generating as many coarse-grained model scales as necessary, that is, to go beyond the two scales in conventional coarse-grained strategies; furthermore, the wavelet-based coarse-grained models explicitly link time and length scales. Furthermore, a straightforward method for the reintroduction of omitted degrees of freedom is presented, which plays a major role in maintaining model fidelity in long-time simulations and in capturing emergent behaviors.

I. INTRODUCTION

Balancing the competing needs of computational complexity and fidelity to the underlying chemical and physical phenomena being studied is a significant challenge in molecular simulations. The challenge is made more acute when the properties under investigation evolve over time, as many orders of magnitude exist between the fundamental time scale of molecular motions and the time scale of collective material processes such as heat transfer, diffusion, or elastic deformations. Coarse-graining methodologies have been frequently used to reduce the complexity of molecular-dynamics (MD) simulations, by reducing the number of "particles" being studied [1–5]. Although the elimination of degrees of freedom often leads to minor gains in the size of the time steps that can be used in dynamic simulations [6–10], such improvements have not led to significant breakthroughs in simulation capabilities. More advanced coarse-graining techniques will be required to enable simulations of systems exhibiting structure on the micron or submillimeter scales, such as semicrystalline materials [11].

In this paper we present an approach towards that end, focusing on polymeric materials composed of homopolymers or block copolymers. Our approach is based on the concept of diffusion wavelets [12,13], which both automatically identifies chemical structures that can be reduced into coarse-grained units and also allows for repeated application, thus permitting much greater customization of the level of simulation than the traditional single-step coarse-graining approach. Unlike traditional wavelet decompositions, which are based on predetermined scaling and wavelet functions [14,15], the functional form of diffusion wavelets is based on the topological structure to which they are applied, making them ideally suitable for use in molecular systems. Such a technique enables both the potential reduction of the number of degrees of freedom by orders of magnitude and a vastly increased time step. Taken together, these improvements can allow for vastly extended time scales to be simulated using currently available computational resources. Our approach is similar to the wavelet transformed Gaussian network model of [16], but our method directly accesses the underlying molecular Hamiltonian rather than starting from a coarse-grained approximation. This notion is fundamentally distinct from the use of wavelets to efficiently compute coarse-grained (CG) potentials, for example, in the multiscale coarse-graining approach [17–20].

Among the shortcomings of typical coarse-graining approaches has been that they normally involve only two levels of description, the "original" atomistic level, and the coarse representation, which has typically been based on the developer's judgment or intuition. These techniques rely on partitionings of the atoms as the foundation to deriving coarse degrees of freedom (DOFs), which typically replace the atomistic groupings with single beadlike entities [21–27]. While approaches such as the force-matching method of Lu *et al.* [28] and the reversible coarse-graining method for phenolic polymers by Kremer and co-workers [21,22,29] have greatly impacted accessible simulation scales for these materials, they do not fundamentally change the underlying computational process. Furthermore, these CG approaches introduce the difficult problem of generating consistent atomistic reconstructions [30,31], because particlelike coarse-grained beads lose information about the particles they subsume. In order to address this loss of information, approaches have been developed to reintroduce specific information selectively [32,33] or to use mixed resolutions [34]. On the other hand, mixed-resolution approaches can pose theoretical challenges regarding the fundamental underlying Hamiltonian [35].

The present work focuses on polymeric materials, which are the target application for many classes of coarse-graining methodologies. In particular, polymers' long chain lengths and low defect concentrations compound the problems of two-scale representations and fine-scale reconstruction, by requiring a large number of particles and their associated DOFs. Moreover, the equilibration time for polymeric liquids is itself computationally challenging: A melt whose chains contain N beads each will require $O(N^3)$ time to equilibrate, making simulation essentially impossible without advanced simulation approaches [36]. (For a review of this problem, see [37].) However, coarse graining need not be restricted only to polymers; even relatively simple fluids, such as water, have been the subject of coarse-grained models [38].

Many of the above problems can be alleviated by selecting an appropriate basis set for describing the internal structure of individual molecules. Consider the analogy of a time-varying quantity $f(t)$. One naturally represents $f(t)$ in terms of an infinitely local basis (Dirac δ functions). A more sophisticated approach, suitable for analyzing certain average properties of the signal, might employ the Fourier basis consisting of sines of varying frequency. However, the latter basis is global in nature, presenting a challenge to processing and storage of signals that are infrequent or have sharply varying features over only small durations in time. In signal processing, wavelets are often used as a basis to differentiate between local and increasingly global features of the signal, because wavelet bases can be flexibly defined to efficiently capture features of varying localization [14,15]. Few other models provide the on-the-fly adaptivity required for important problems, which in the structural modeling sense might be problems including crack initiation, crack propagation, and interfacial phenomena [39–41]. By analogy to the time-varying signal, purely atomistic models and standard two-scale CG models represent infinitely local Dirac distributions, which are expensive; coarse-graining methods based on globally periodic functions (e.g., plane waves) are inefficient for modeling localized properties. In the context of time-varying signals, the mathematical inefficiency can be quantified precisely: Except for limited special cases, either basis requires infinitely many coefficients to approximate a signal to within tolerance of a given error metric, e.g., the l_2 or l_∞ metrics. In contrast, the number of coefficients required for a wavelet-based representation is usually $O(\ln_2 N)$, where N is the size of the signal (e.g., number of time samples).

Wavelet ideas have already been used extensively to analyze time series data [15], including some relatively early applications in molecular dynamics [42]. Our application of wavelet ideas to a *structural* representation extends the work of Ismail *et al.* [43,44]. In particular, the earlier work employed the Monte Carlo method and did not capture dynamical information. The work here develops an approach suitable for molecular dynamics simulation by interpreting the wavelet transform of Ismail *et al.* in terms of the equations of motion. The method provides a consistent and systematic framework to derive multiple levels of model resolution while also reducing simulation complexity.

This approach has numerous advantages, whose theoretical basis we address in this paper. The general theme of the advantages is that for the dynamical and nonequilibrium metrics of interest, this approach especially captures molecular information relevant to both thermodynamics and kinetics. In particular, application of diffusion wavelets to the chemical topology of the molecule leads to the identification of what we call collective action modes (CAMs) that represent coordinated motions within the molecule at various length and time scales. Our approach is rigorously tied to the underlying physics and has the potential to increase simulation size and duration by several orders of magnitude. Moreover, the approach is agnostic to the kind of material being studied and can be applied to structures of arbitrary chemical complexity, including both relatively simple molecules such as water or benzene and more complicated chain molecules such as polymers and biopolymers (for a multiscale model based on Lie groups, see [45]). Finally, it is also capable of capturing minor effects such as mass effects from chemical substitution, such as partial deuteration or fluorination. This may be of special importance for materials design, where the task is to link macroscopic behavior (e.g., Young's modulus) to the atomic structure of the monomer unit.

We proceed by introducing the type of classical Hamiltonian model we seek to accelerate, as well as the wavelet methodology and its basic properties. We then discuss how wavelets can be applied to uncover the CAMs for a given molecule and how to use CAMs to systematically reconstruct finer resolutions, as well as deriving mixed resolutions. We present several examples showing the application of the method for small molecules such as water and HCN before summarizing our conclusions.

II. METHODOLOGY

We begin by analyzing the equations of motion in MD for a physics-based wavelet construction. The foundations of MD lie in the application of Newtonian mechanics to the energy functional

$$E = \tfrac{1}{2}\,\mathrm{tr}(\dot{x}^T M \dot{x}) + V(x), \tag{1}$$

where $x \in \mathbb{R}^{N \times 3}$ are particle positions, $\dot{x} \in \mathbb{R}^{N \times 3}$ are particle velocities, N is the number of particles, M is the diagonal matrix of particle masses, and V is the potential. For the macromolecular systems we are interested in, V is usually partitioned as

$$V = V_{\mathrm{harmonic}} + V_{\mathrm{nonharmonic}}, \tag{2}$$

where $V_{\mathrm{harmonic}}(x) = \sum_{i \neq j} K_{ij}(\|x_i - x_j\| - r_0^{(ij)})^2/2$, K_{ij} is the force constant of the harmonic oscillator, x_k is the position of particle k, and $r_0^{(ij)}$ is the equilibrium distance of particles i and j. The atoms and the bonds between them define a graph in which the atoms are the vertices and an edge between atoms i and j has weight K_{ij}. The maximum number of edges for a vertex in organic materials is 4 and even in organometallic complexes the number of edges is unlikely to exceed 6 (the typical maximum coordination number). Consequently, the matrix representation of this graph, defined by K, should be highly sparse.

Since generally the bond vibrations associated with the harmonic potential limit time-step size and define molecular units outside of reactive potentials [46,47], our multiresolution approach begins with the graph Laplacian of the weighted
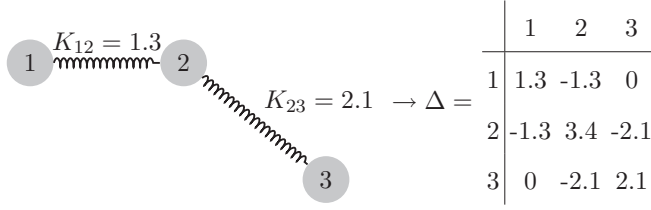
FIG. 1. Example of a weighted graph Laplacian derived from a simple weighted graph.
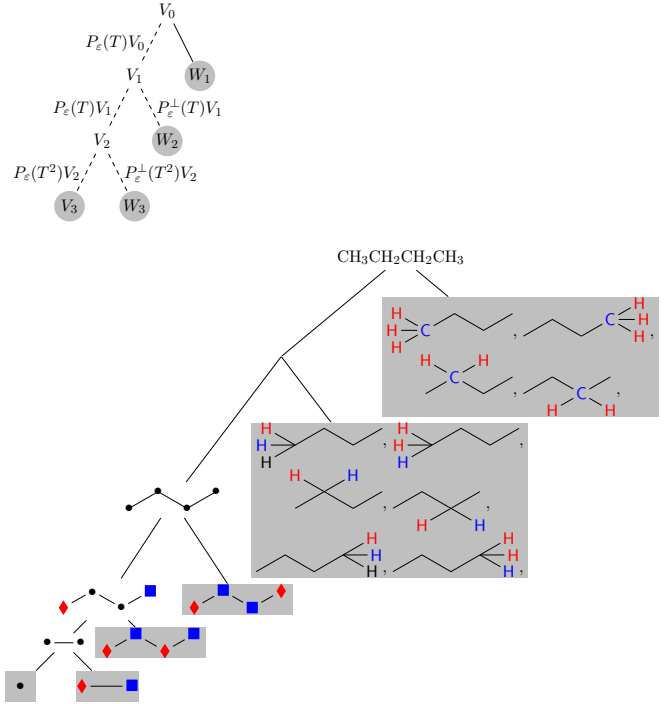

FIG. 2. Dyadic trees generated by multiresolution formalism using a filter T. Shaded spaces are final subspaces of the multiresolution. All other spaces are intermediates of the construction. Shown on top is the generic scheme and on bottom is an example using butane. The final spaces are in gray; diamonds (red) and squares (blue) denote opposite signs of weights in the construction from the finer scale.

graph defined by using particles as vertices and the bonds as edges weighted by the harmonic force constant. This graph Laplacian, denoted by Δ, is here a matrix that can be defined as

$$\Delta := \mathrm{diag}(K\mathbf{1}) - K, \qquad (3)$$

$$\Delta_{ij} = \begin{cases} \sum_j K_{ij}, & i = j \\ -K_{ij}, & i \neq j, \end{cases} \qquad (4)$$

where $\mathbf{1}$ denotes a vector of all ones and $\mathrm{diag}(v)$ denotes a diagonal matrix defined in terms of a vector v, so that if $A = \mathrm{diag}(v)$ then $A_{ii} = v_i$. An example of the weighted graph Laplacian Δ for a linear triatomic molecule can be found in Fig. 1.

Using Eq. (1) and our definition of Δ, we derive the equations of motion

$$M\ddot{x} = -\boldsymbol{\nabla} V(x) = -\Delta x - V'(x), \qquad (5)$$

where $V'(x) = \boldsymbol{\nabla} V(x) - \Delta x$ is the force for x not due to Δ; Δ is the graph Laplacian of the weighted graph with particles as vertices and the bonds as edges weighted by the harmonic force constant (see Fig. 1). By letting $r = M^{1/2}x$, $p = \dot{r}$, and $q = \tilde{\Delta}^{1/2}M^{1/2}x$, where $\tilde{\Delta} = M^{-1/2}\Delta M^{-1/2}$, the harmonic energy term can be expressed as $(\|p\|^2 + \|q\|^2)/2$ and the equations of motion become

$$\dot{p} = -\tilde{\Delta}^{1/2}q - V'(x) = -\tilde{\Delta}^{1/2}q - \tilde{V}(r), \qquad (6)$$

where $\tilde{V}(r) = V'(M^{-1/2}r)$ is the effective potential in terms of r. If $\tilde{V} = 0$, the system can be solved analytically,

$$y(t) = e^{\Upsilon t}y_0, \qquad (7)$$

where

$$\Upsilon = \begin{pmatrix} 0 & \tilde{\Delta}^{1/2} \\ -\tilde{\Delta}^{1/2} & 0 \end{pmatrix}, \quad y(t) = \begin{pmatrix} q(t) \\ p(t) \end{pmatrix}, \quad y_0 = \begin{pmatrix} q_0 \\ p_0 \end{pmatrix}. \qquad (8)$$

It can be shown that $\tilde{\Delta}$ is positive semidefinite and that the eigenvectors of the exponential operator in Eq. (7) are $(U_j, \pm i U_j)$, where U_j is an eigenvector of $\tilde{\Delta}$. The solutions to Eq. (7) oscillate with frequencies $\pm\omega_j$, respectively, where ω_j^2 is the eigenvalue of $\tilde{\Delta}$ corresponding to the eigenvector U_j. For this analytical case, any eigensolution of Υ can be propagated through time independently of any other solution. Unfortunately, this simplicity is in general broken by $\tilde{V}$, which nonlinearly couples all of the eigensolutions to one another. As a consequence, all solutions have to be simulated concurrently according to the highest frequency associated with an eigenvector. Hence, the transformation from particle space coordinates $(x, \dot{x})^{\dagger}$ to the harmonic solution coordinates

offers no advantage. To circumvent this issue, a basis that isolates the coupling effects from high- and low-frequency components is needed.

The key motivation for our work is the recognition that the mass-weighted graph Laplacian $\tilde{\Delta}$ relates spatial coordinates to temporal frequencies, which suggests that its eigenvector matrix is a promising basis for compression. The weighted graph Laplacian $\tilde{\Delta}$ and its matrix of eigenvectors are then analogous to the Laplace operator and the Fourier transform, respectively, in conventional wavelet theory.

A. Basic multiresolution

Here we introduce the wavelet transform used in this work and its derivation. We use the multiresolution analysis for diffusion wavelets as introduced by Coifman and Maggioni [12]. In essence, the multiresolution decomposition partitions the eigenvalues and eigenvectors of $\tilde{\Delta}$, effectively strongly coupling high frequencies in the time domain to high-frequency eigenvectors of $\tilde{\Delta}$ in the particle domain. This is an important point for the applicability of our approach: Not only can DOFs be reduced, but the time step may also be increased considerably, approximately by a factor of 2 at each subsequent resolution.

The multiresolution scheme (see Fig. 2) relies on a positive-definite low-pass filter T with $\|T\|_\infty = 1$, i.e., an operator that suppresses high-frequency vectors, and an accuracy operator P_ε that projects eigenvectors of a matrix $X \in \mathrm{span}\{T^{2^n} | n \in \mathbb{N}\}$ with associated eigenvalue less than a given accuracy $\varepsilon > 0$

to zero. The effect and purpose of the filter will be to project out DOFs associated with high frequencies, thereby producing a hierarchy of CAMs in respective vector spaces. After each application, the associated frequencies roughly halve and concomitantly the minimum time step size roughly doubles.

The recursively defined vector spaces $V_n = V_{n+1} \oplus W_{n+1}$, where

$$V_{n+1} = P_\varepsilon(T^{2^{n+1}})V_n \qquad (9)$$

and

$$W_{n+1} = \ker P_\varepsilon(T^{2^{n+1}}|_{V_n}) \qquad (10)$$

are iteratively associated with orthonormal bases via QR decompositions,

$$T^{2^n} = Q_n R_n, \qquad (11)$$

where Q_n is unitary, R_n is upper triangular, and both are dependent on the basis used to express T^{2^n}. The algorithm terminates due to the finite spectrum of T when only a single eigenvector remains.

One advantage of this approach is the inherent invariance of the wavelet spaces with respect to permutations of the graph vertices and its ability to deal with arbitrary matrices T with $\|T\|_\infty \leqslant 1$. The wavelet space at scale n (W_n) is approximately spanned by eigenvectors of T whose eigenvalues obey

$$2^{-n+1} \ln \varepsilon < \ln \lambda \leqslant 2^{-n} \ln \varepsilon. \qquad (12)$$

Only a limited degree of mixing with eigenvectors associated with eigenvalues outside of these bounds is possible (see Appendix B for details). Since T is positive semidefinite, T can be rewritten as an exponential $e^{\ln T}$. The application of the previous procedures are equivalent to doubling the spectrum of $\ln T$ followed by projecting out the high-frequency components of the spectrum of $\ln T$.

1. Properties of $\tilde{\Delta}$ as a filter

With an efficient means of computing the wavelet transform in hand, we consider exclusively in the following the low-pass filter $T = I - \tilde{\Delta}/C$, where C is a constant sufficiently large to render T positive semidefinite. In order to minimize the number of matrix-matrix multiplications, in particular with $\ker P_\varepsilon(T^{2^n})|_{V_n} = \emptyset$, the normalization constant $C = \|\Delta\|_\infty$ would generally be optimal as $\Psi_1 \neq \emptyset$. Eigenvalues for graph Laplacians, such as Δ, are known to lie in $[0, 2 \max_i \Delta_{ii}]$ [48]. Hence, C is chosen to be between $\max_i \tilde{\Delta}_{ii}$ and $2 \max_i \tilde{\Delta}_{ii}$ for all numerical examples.

As implied by Eq. (12), the frequency range $\Delta \ln \lambda = 2^{-n} \ln \varepsilon$ for each wavelet space W_n is drawn ever tighter with each iteration while the eigenvalues λ approach 1. Thus, generally, more CAMs correspond to scales of high frequency than scales of low frequency. As a result, each successive scale corresponds to an increase in the minimum time step size of that scale in MD simulations of approximately ε^{-2^n} as well as a significant reduction of CAMs. Furthermore, unless the effective potential $\tilde{V}$ couples involved scales strongly,
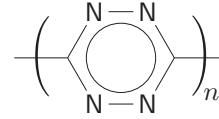

FIG. 3. Chemical structure of polytetrazine.

sufficiently coarse scales ($n \gg m$) are quasistatic compared to a given scale m, while sufficiently fine scales ($n \ll m$) only influence the target scale via their static mean. Therefore, only the relevant scales need to be propagated through time, reducing the DOFs and allowing one to raise the time step size to match the scales.

Assuming that eigenvalues of T are distributed approximately exponentially, $\varepsilon = e^{-1/2}$ would lead to $\ln_2 N$ scales, where N is the dimensionality of $\tilde{\Delta}$. This leads to many DOFs per scale due to issues discussed below. Instead a higher resolution of $\varepsilon_{\text{machine}}^{1/2}$ is used. Although this wastes some computation on the first few iterations because $\ker P_\varepsilon T^{2^n}|_{V_n} = \emptyset$, it is equivalent to choosing a tolerance on the scale of $\delta_{\max}$, $\varepsilon_{\text{effective}} = \varepsilon_{\text{machine}}^{2^{-m}} \approx 1 - \delta_{\max}$.

2. How molecular information influences $\tilde{\Delta}$

Since the filter T and the weighted graph Laplacian $\tilde{\Delta}$ derived from the MD potentials share the same eigenvectors albeit with reversed order of eigenvalues, we discuss its properties in greater detail. These properties have a major impact on the performance of the wavelet transform. In the following, we discuss shortly the conditions under which separate groups of atoms are strictly independent of each other leading to highly localized CAMs.

Due to the small number of bonds an atom generally has, vertices are also generally of low degree. As a result, there are highly localized modes due to invariant subspaces of chemical motifs. For example, any CH_2 group has an associated medium-frequency, highly localized eigenvector of $\tilde{\Delta}$. This follows from the fact that the hydrogens are leafs on the graph, i.e., the weighted subgraph Laplacian

$$\tilde{\Delta}_{CH_2} = \begin{pmatrix} \frac{2K_{CH}}{12} + o & -\frac{K_{CH}}{\sqrt{12}} & -\frac{K_{CH}}{\sqrt{12}} \\ -\frac{K_{CH}}{\sqrt{12}} & K_{CH} & 0 \\ -\frac{K_{CH}}{\sqrt{12}} & 0 & K_{CH} \end{pmatrix}, \qquad (13)$$

where o collects the contributions from further bonds with the carbon atom, shows only contributions from CH_2 for the hydrogens. If the hydrogen atoms are identified with indices i and j on the full graph Laplacian, then $(e_i - e_j)/\sqrt{2}$ is an eigenvector with an associated frequency $\sqrt{K_{CH}/m_H}$ of a CH vibration.

A slightly more involved example that also shows that this phenomenon is not restricted to leaves on the graph is the repeat unit of the energetic polymer poly-1,2,4,5-tetrazine (see Fig. 3), which has an invariant subspace spanned by two independent eigenvectors. The repeat unit block of the weighted

graph Laplacian $\tilde{\Delta}$ is

$$
A = \begin{pmatrix}
\frac{2X+Z_1}{12} & -\frac{X}{\sqrt{12\times14}} & 0 & -\frac{X}{\sqrt{12\times14}} & & & \\
-\frac{X}{\sqrt{12\times14}} & \frac{X+Y}{14} & -\frac{Y}{\sqrt{12\times14}} & & & & \\
0 & -\frac{Y}{\sqrt{12\times14}} & \frac{X+Y}{14} & & & & -\frac{X}{\sqrt{12\times14}} \\
-\frac{X}{\sqrt{12\times14}} & & & \frac{X+Y}{14} & -\frac{Y}{\sqrt{12\times14}} & 0 \\
& & & -\frac{Y}{\sqrt{12\times14}} & \frac{X+Y}{14} & -\frac{X}{\sqrt{12\times14}} \\
& & -\frac{X}{\sqrt{12\times14}} & 0 & -\frac{X}{\sqrt{12\times14}} & \frac{2X+Z_6}{12}
\end{pmatrix},
\tag{14}
$$

where X is the bond constant for the C–N bond, Y for the N–N bond, and $Z_{1/6}$ for the contributions of vertices outside the repeat unit. The two vectors $(0,1,1,-1,-1,0)$ and $(0,1,-1,1,-1,0)$ span a subspace invariant under application of A but that is mapped to zero for operators B with entries $B_{ij} \neq 0 \iff i/j \notin \{2,3,4,5\}$. Hence, these vectors are highly localized (see Appendix C for details of the general case).

In large linear homopolymers, discussed in later sections, the invariant subspaces represent highly degenerate eigenvalues due to the polymer's repetitive structure. Degeneracies in $\tilde{\Delta}$ reduce the effectiveness of the wavelets in the sense that no choice of frequency accuracy alone can be used to separate them into individual subscales. On the other hand, when these degeneracies exist outside of the focus of the targeted time scale, the degeneracy allows for a large reduction in tracked CAMs. Especially for polymers, the high degeneracies associated with CAMs internal to the repeat unit are generally high-frequency CAMs and are thus dropped at longer time scales. In such cases, when large numbers of degenerate frequencies coincide with the targeted time scale, it may be possible to incorporate more information from the potential $\tilde{V}$, but this is outside of the current scope.

Example: Butane $[H(CH_2)_4H]$. Butane has 14 atoms and therefore the graph Laplacian associated with $H(CH_2)_4H$ is the 14-dimensional square matrix

$$
\Delta \propto \begin{pmatrix}
1 & -1 & & & & & & & & & & & & \\
-1 & 3+\kappa & -1 & -1 & & -\kappa & & & & & & & & \\
& -1 & 1 & & & & & & & & & & & \\
& -1 & & 1 & & & & & & & & & & \\
& -\kappa & & & 2+2\kappa & -1 & -1 & & -\kappa & & & & & \\
& & & & -1 & 1 & & & & & & & & \\
& & & & -1 & & 1 & & & & & & & \\
& & & & -\kappa & & & 2+2\kappa & -1 & -1 & & -\kappa & & \\
& & & & & & & -1 & 1 & & & & & \\
& & & & & & & -1 & & 1 & & & & \\
& & & & & & & -\kappa & & & 3+\kappa & -1 & -1 & -1 \\
& & & & & & & & & & -1 & 1 & & \\
& & & & & & & & & & -1 & & 1 & \\
& & & & & & & & & & -1 & & & 1
\end{pmatrix},
\tag{15}
$$

where $\kappa = 44/31$ is the ratio of force constants for a C–C bond to a C–H bond in the polymer-consistent force field (PCFF) [49] and the units of kcal/mol/Å/Å have been subsumed in the proportionality constant. The individual CH_2 repeat units have been boxed for emphasis. Since 660 kcal/mol/Å^2 is an upper bound for Δ, using the filter $T = I - \Delta /(660 \text{ kcal/mol/Å}^2)$ and $\varepsilon_{\text{machine}}$ as the cutoff criterion yields six scales. The first four application of the scheme ($T, T^2, T^4, T^8,$ and T^{16}) did not lead to any unit vectors below the threshold. At $n = 5$, the four highest frequency modes of Δ ($\lambda = 517.1, 524.0, 556.3,$ and 556.3 kcal/mol/Å^2,

respectively),

$$
\left(
\begin{array}{cccccccccccccc}
\text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{H} \\
\hline
-0.50 & \mathbf{0.46} & -0.50 & -0.50 & & & & & & & & & & \\
 & & & & \mathbf{0.33} & -0.53 & -0.53 & \mathbf{0.21} & -0.36 & -0.36 & & & & \\
 & & & & -0.24 & \mathbf{0.35} & \mathbf{0.35} & \mathbf{0.35} & -0.52 & -0.52 & & & & \\
 & & & & & & & & & & -0.46 & \mathbf{0.50} & \mathbf{0.50} & \mathbf{0.50}
\end{array}
\right),
\tag{16}
$$

approximately map to zero for T^{32}, i.e., the logarithm of the expectation values with T do not exceed $2^{-5}\ln\varepsilon \approx -1.13$ [$\ln(1-\lambda/660) < -1.53$]. In Eq. (16) the coefficients are coded by their signs to highlight the inherent symmetries; negative values are in italics and positive values are in bold. These CAMs correspond to the symmetric stretches of the methyl groups. Hence the coefficients on the nonzero weighted carbon indices are opposite in sign to their bound hydrogens. This is less obvious in the middle block of Eq. (16), as these are superpositions of the more localized vectors.

The six second highest frequencies [$\lambda = 440\,\text{kcal/mol/Å}^2$ for all $\ln(1-\lambda/660) \approx -1.10$]

$$
\left(
\begin{array}{cccccccccccccc}
\text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{H} \\
\hline
\mathbf{0.81} & 0 & -0.39 & -0.43 & & & & & & & & & & \\
0 & \mathbf{0.72} & -0.69 & & & & & & & & & & & \\
 & & 0.10 & 0 & -0.70 & \mathbf{0.70} & 0 & & & & & & & \\
 & & & & 0 & -0.71 & \mathbf{0.71} & & & & & & & \\
 & & & & & & & 0 & \mathbf{0.82} & -0.41 & -0.41 & & & \\
 & & & & & & & & & \mathbf{0.71} & -0.71 & & &
\end{array}
\right),
\tag{17}
$$

are computed for $n = 6$ and their expectations with T are between $2^{-5}\ln\varepsilon \approx -1.13$ and $2^{-6}\ln\varepsilon \approx -0.56$. These CAMs correspond to individual HH stretches. The CAMs on the first and last two rows cover the degeneracy between the three hydrogens within the respective methyl groups, whereas the third and fourth rows show isolated HH vibration modes.

The next two powers of T (T^{128} and T^{256}) do not filter out any new spaces. The third wavelet subspace is spanned by a single vector [$\lambda = 72.7\,\text{kcal/mol/Å}^2$ for $\ln(1-\lambda/660) \approx -0.12$]

$$
\left(
\begin{array}{cccccccccccccc}
\text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{H} \\
\hline
0 & -0.22 & 0 & 0 & \mathbf{0.59} & \mathbf{0.20} & \mathbf{0.20} & -0.59 & -0.20 & -0.20 & \mathbf{0.22} & 0 & 0 & 0
\end{array}
\right),
\tag{18}
$$

with exponent $n = 9$ for T^{2^n}. The logarithm of its expectation value with T is between $2^{-6}\ln\varepsilon \approx -0.56$ and $2^{-9}\ln\varepsilon \approx -0.070$. This CAM isolates the asymmetric stretch between the terminal carbons and the centers of mass of the bridging methylene groups.

The fourth wavelet subspace is spanned by another vector [$\lambda = 42.1\,\text{kcal/mol/Å}^2$ for $\ln(1-\lambda/660) \approx -0.066$],

$$
\left(
\begin{array}{cccccccccccccc}
\text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{H} \\
\hline
\mathbf{0.13} & \mathbf{0.43} & \mathbf{0.13} & \mathbf{0.13} & -0.46 & -0.15 & -0.15 & -0.46 & -0.15 & -0.15 & \mathbf{0.43} & \mathbf{0.14} & \mathbf{0.14} & \mathbf{0.14}
\end{array}
\right),
\tag{19}
$$

with exponent $n = 10$ for $T^{2^n} = T^{1024}$. The logarithm of its expectation value with T is between $2^{-9}\ln\varepsilon \approx -0.070$ and $2^{-10}\ln\varepsilon \approx -0.035$. This CAM captures the symmetric stretch between the center of mass of the bridging methylenes and the centers of mass of the terminal methyl groups.

The last nontrivial wavelet subspace also is spanned by a single vector [$\lambda = 12.2$ kcal/mol/$\mathring{A}^2$ for $\ln(1 - \lambda/660) \approx -0.019$]

$$\left(\begin{array}{c|ccccccccccccc} \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{H} \\ \hline -0.17 & -0.59 & -0.17 & -0.17 & -0.24 & 0 & 0 & \mathbf{0.24} & 0 & 0 & \mathbf{0.59} & \mathbf{0.17} & \mathbf{0.17} & \mathbf{0.17} \end{array}\right), \tag{20}$$

with exponent $n = 11$ for $T^{2^n} = T^{2048}$. The logarithm of its expectation value with T is between $2^{-10}\ln\varepsilon \approx -0.035$ and $2^{-11}\ln\varepsilon \approx -0.018$. This CAM captures the symmetric stretch of the bridging carbons and the centers of mass of the terminal methyl groups.

The coarsest level (0 kcal/mol/$\mathring{A}^2$)

$$\left(\begin{array}{c|ccccccccccccc} \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{C} & \text{H} & \text{H} & \text{H} \\ \hline 0.13 & 0.45 & 0.13 & 0.13 & 0.45 & 0.13 & 0.13 & 0.45 & 0.13 & 0.13 & 0.45 & 0.13 & 0.13 & 0.13 \end{array}\right) \tag{21}$$

This last CAM is just the center of mass.

3. Multiresolution construction from fragments

Since the previously mentioned invariant subspaces are inherent to molecular fragments and some molecular fragments are particularly common, the question of how much can be gained by precomputing the internal scales of these fragments arises. For example, proteins are long heterogeneous polymers, but they are mostly comprised of only 20 amino acids. Hence it is instructive to characterize the effects of modifying bonds or substituting different atoms. Indeed, connecting fragments (see Appendix D for details) affect precomputed CAMs such that only a few CAMs need to be adjusted. Precomputed fragments can therefore speed up the computation of scales considerably when the perturbations due to connecting them are relatively small.

4. Example: Linear homopolymers

Linear homopolymers are an important class of materials, whose graph Laplacians exhibit convenient structures that we will exploit in the following to derive their CAMs. Linear homopolymers are a successive addition of edges between identical building blocks. We can derive computationally inexpensive algorithms to compute the eigenvalues of their weighted graph Laplacians and thereby the successive construction of the respective wavelet spaces. The eigensystems are computed by exploiting the recursive structure of $\tilde{\Delta}$ to solve n much smaller eigensystems, where n is the number of repeat units in a single chain of the polymer. The graph Laplacian of linear homopolymers can be ordered to have a block Toeplitz structure, where each nonterminal block is a constant $m \times m$ matrix for the off-diagonal B and diagonal A, respectively,

$$\tilde{\Delta} = \begin{pmatrix} A & B^* & 0 & B \\ B & \ddots & \ddots & 0 \\ 0 & \ddots & & B^* \\ B^* & 0 & B & A \end{pmatrix}, \tag{22}$$

where m is the number of particle DOFs in the repeat unit of the polymer. Furthermore, the off-diagonal block B generally consists of a single nonzero entry representing a single bond connecting repeat units, such as $B_{1,m}$ [see off-diagonal blocks in Eq. (15) for an example]. This highly repetitive structure can be exploited to compute eigenvectors and eigenvalues very efficiently and hence the CAMs (see Appendix E).

The following simulation exemplifies the effectiveness of the approach. The molecular-dynamics simulation was performed in LAMMPS [50]. A crystalline model of polyethylene (PE) consisting of 100 800 atoms in a single infinite chain, wrapped due to periodic boundary conditions, was simulated in the NVT ensemble with 2-fs time steps, PCFF [49] at 500 K for 1000 time steps. The monoclinic unit cell has edge vectors $\mathbf{a} = (103.432\ \mathring{A}, 0, 0)$, $\mathbf{b} = (0, 147.87\ \mathring{A}, 0)$, and $\mathbf{c} = (73.88\ \mathring{A}, 0, 50.78\ \mathring{A})$. The initial structure is depicted in Fig. 4.
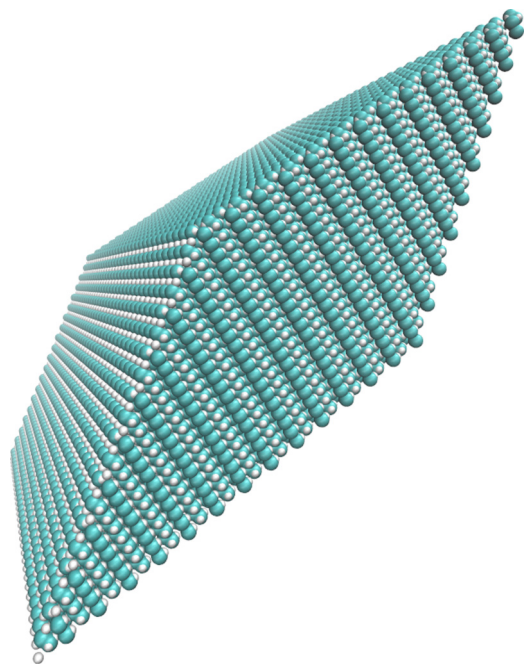

FIG. 4. Simulated 3D structure of a 100 800-atom polyethylene single crystal.
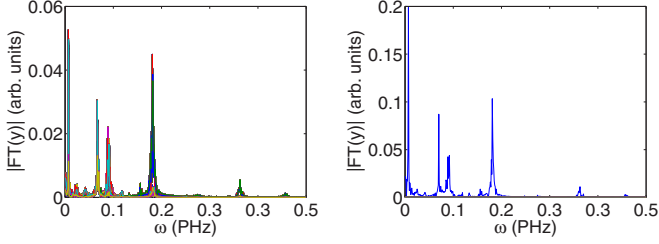
FIG. 5. Fourier transform of the y component [FT(y)] of 1000 atoms in crystalline PE (100 800 atoms) with frequency ω, excluding zero frequency to allow detail at other frequencies. The MD simulations were at 500 K and 1 atm. Shown on the left are individual power spectra per atom and on the right is the power spectrum of magnitude of optimal representation.

Figure 5 shows the superimposed Fourier-transformed time series of 1000 atoms from the 500 K trajectory, which due to the high temperature shows the highest mobility of atoms. Although, the zero frequency is by far the most intense signal (and is omitted from the figure for clarity), a wide range of other frequencies is active, most importantly around 0.45 PHz, which limits the time step of atomistic simulations of PE to 2.4 fs or less.

On the other hand, Fig. 6 shows the effects of scales on time series analysis. The finest scale out of 25 still retains the high-frequency components [Fig. 6(a)] as may be expected, but they are much less intense than the remaining modes. Traversing the scales, it is noteworthy that a decreasing number of DOFs at the coarser scales show significant peaks at all. At the medium scale [Fig. 6(b)], no high-frequency components are found anymore. Therefore, medium-scale and coarser DOFs are quasistatic compared to the finer scales. Furthermore, the signals are clearly separated and very sharp despite the high temperature, which speaks for strongly decoupled modes and justifies dropping the finer scales, which in turn facilitates
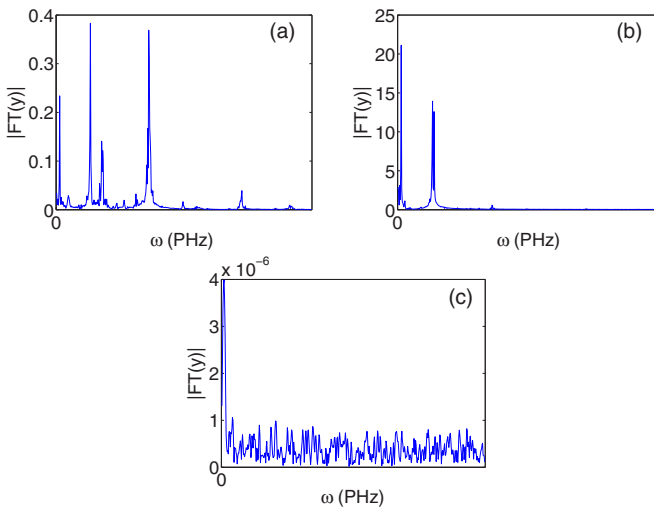

FIG. 6. Fourier transform of the y component [FT(y)] of a 100 800-atom crystalline PE sampled at 1 fs with frequency ω. Three scales are shown: (a) the first (finest) scale, (b) the fifth scale, and (c) the 12th scale (out of 25).

speed-up by not only reducing DOFs but also increasing the propagation time step.

B. Adaptive multiresolution

In the previous sections, our goal was to identify scales of the physical system to remove unimportant CAMs, thereby increasing time step size and reducing DOFs. However, for some phenomena, such as phase transitions such as melting of a polymer, fine details that are unimportant at one point in time can play a major role at another. So we now turn to the problem of reconstruction, that is, reintroducing dropped CAMs. We find that reconstruction is systematically possible for numerical (e.g., derived from iterative Boltzmann inversion) and analytical (where such exists, e.g., quadratic potentials) coarse-graining hierarchies.

1. Reconstruction theory

We begin by putting coarse graining into a wider context. In general, a coarsening $\gamma : \alpha \to \beta$ is a continuous surjection between two topological state spaces α, the fine space, and β, the coarse space, which can be parametrized by $n > m$ state variables, respectively. As an example, one can consider mapping the positions of groups of particles to their centers of mass. In statistical mechanics, the fine-grained state space α is associated with a probability measure $P_\alpha : \{X \subset \alpha\} \to [0,1]$ and a corresponding probability space. For systems in equilibrium, α would further follow a Boltzmann distribution. The coarsening γ thereby induces a probability space on the coarse states in β as well, with the probability measure $P_\beta(k) = P_\alpha(\gamma^{-1}(k))$, where $k \subset \beta$ and $\gamma^{-1}(k) \subset \alpha$ is the preimage of k. Here P_β constitutes the mean probability distribution of the coarsened DOFs. It is thus possible to select (reconstruct) a precursor for a state $b \in \beta$ by sampling $\gamma^{-1}(\{b\})$ with P_α via the conditional probability $P(a|b) = P_\alpha(\{a\} \cap \gamma^{-1}(b))/P_\alpha(\gamma^{-1}(b))$, e.g., using Monte Carlo sampling. Since $n > m$, there is also a complementary coarsening $\gamma^\perp : \alpha \to \ker \gamma$ with its complementary probability measure.

In MD, the state spaces consist of the positions and their associated momenta and thermodynamic state variables, such as temperature or pressure. The probability distributions are Boltzmann distributions that depend on the studied thermodynamical ensemble. In a sequence of coarsenings $(\gamma_n : \beta_n \to \beta_{n+1})_n$ spanning several scales, such as that derived from the preceding multiresolution scheme, it is generally not cost effective to sample fully in the largest space $\alpha =: \beta_0$ and analytical derivations for P_{β_n} are rarely available. Hence, approximations need to be made. Common solutions in the MD community are probability measures from iterative Boltzmann inversion or (successive) force matching to generate effective potentials that are themselves Boltzmann distributed. Hierarchical iteration thereby produces not only probability distributions on the coarser space, but also conditional probability distributions for a fine state mapping to a coarse state. Furthermore, the probability distributions can be used to indicate when a previously undersampled coarse state subspace is encountered at some state x, e.g., using an expected improvement measure [51] or a sensitivity analysis of the potential $-\ln P_\beta(x)$ with respect to the sampled points, for which the trust boundaries can be precomputed.
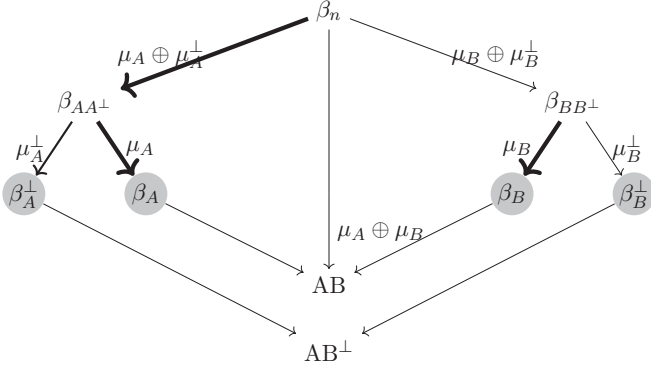
FIG. 7. Dyadic tree generated by separable coarsenings. Shaded spaces are final subspaces of the multiresolution. All other spaces are intermediates of the construction with $\gamma_1 = \mu_A \oplus \mu_A^\perp$ and $\gamma_2 = \mu_A \oplus \mu_B$.

2. Coarse-graining hierarchies

In the context of MD, mixed resolutions are necessary for the concurrent treatment of, e.g., gross rigid protein orientations and flexible active site residue interactions with a binding molecule. State spaces with mixed levels of fine and coarse CAMs arise naturally from the multiresolution scheme laid out in Eqs. (9) and (10) (see Appendix F and Fig. 7 for details).

The sequence of coarsenings $(\gamma_n : V_n \to V_{n+1})_n$ generates a full dyadic tree, due to the complementary coarsenings $\gamma_n^\perp$, i.e., each state space can be split into $\ker \gamma_n^\perp$ and $\ker \gamma_n$. A sequence of separable coarsenings $(\gamma_n, \beta_n, \mu_{A_n}, \mu_{B_n})$ thereby induces a hierarchy of coarsenings and associated probability measures, induced as described above.

3. Wavelet hierarchy

A minimal set of separable coarsenings that generates a full given hierarchy is fundamental and characterizes said hierarchy. One such sequence for coarsenings based on CAMs for $V_m = (\bigoplus_{n=1}^{\lfloor 2^{-m}N \rfloor} W_{2^m n-1})_{m=0}^{\ln_2 N}$ is

$$\mu_{A_m} : \bigoplus_{n=1}^{\lfloor 2^{-m-1}N \rfloor} W_{2^m n-1} \to \bigoplus_{n=1}^{\lfloor 2^{-m-2}N \rfloor} W_{2^{m+1} n-1}, \quad (23)$$

$$\mu_{B_m} = \bigoplus_{n=\lfloor 2^{-m-1}N \rfloor+1}^{\lfloor 2^{-m}N \rfloor} W_{2^m n-1} \to \bigoplus_{n=\lfloor 2^{-m-2}N \rfloor+1}^{\lfloor 2^{-m-1}N \rfloor} W_{2^{m+1} n-1}, \quad (24)$$

where N is the number of particles in the finest resolution.

We note that these intermediate probability distributions are available both analytically and numerically, since an accurate fundamental coarsening has to include proper statistics for the intermediate state space to be consistently sampled. Recursive application of conditional probabilities enables concurrent mixed resolutions. Since the construction of CAMs from the multiresolution analysis produces a hierarchy of frequencies, it induces a hierarchy of coarsenings by dropping successively higher-frequency CAMs, i.e., by applying the low-pass filter $P_\varepsilon(T^{2^n})$.

4. A priori approximations for reconstruction

Implementation of reconstruction algorithms as discussed above requires a starting point. In the following, methods are proposed for finding good starting points based on $\tilde{\Delta}$ and other molecular information that is available prior to simulation, i.e., without the need for MD or Monte Carlo sampling. To second order, a quadratic potential around the equilibrium positions of the transformed coordinates approximates the full potential. We assume dominance of harmonic terms, in both the original and the transformed basis:

$$V_{\text{bond}}(M^{-1/2}U^\mathsf{T}\tilde{r}) \approx \tfrac{1}{2}(\tilde{r}-\tilde{r}_0)^\mathsf{T} U \tilde{\Delta} U^\mathsf{T}(\tilde{r}-\tilde{r}_0), \quad (25)$$

where $\tilde{r} = U M^{1/2} x$ is the position vector in the wavelet basis and $U M^{1/2}$ is the wavelet transformation matrix. From statistical mechanics the root mean square deviation from equilibrium of a harmonic oscillator is $\sqrt{kT/\lambda}$, where λ is the force constant. In other words, the higher-frequency components are found increasingly close to their energy minima. This implies that finer scales only have small deviations from their equilibrium positions, while coarser scales may access a much larger space.

We start by approximating $\|x_i - x_j\|$ by a Taylor expansion around $r_{ij}^{(0)}$. This transforms the bond potential V_{bond} into

$$V_{\text{bond}} \approx \frac{1}{8} \sum_{ij} \frac{K_{ij}}{r_{ij}^{(0)^2}} \big(\|x_i - x_j\|^2 - r_{ij}^{(0)^2}\big)^2. \quad (26)$$

Hence, to find equilibrium distances for $\tilde{r} = U M^{1/2} x$, we solve the minimization problem

$$\min_{\tilde{r}_0} \sum C_{ij} \big[\|(M^{-1/2}U^T \tilde{r}_0)_i - (M^{-1/2}U^T \tilde{r}_0)_j\|^2 - r_{ij}^{(0)^2} \big]^2,$$
$$(27)$$

where $C_{ij} = K_{ij}/r_{ij}^{(0)^2}$.

For example, the two nonzero eigenvalues of H_2O correspond to a unique solution for reconstructing H_2O. The harmonic Laplacian for H_2O,

$$\tilde{\Delta}_{H_2O} = \begin{pmatrix} \frac{2K_{OH}}{m_O} & -\frac{K_{OH}}{m_O^{1/2}} & -\frac{K_{OH}}{m_O^{1/2}} \\ -\frac{K_{OH}}{m_O^{1/2}} & K_{OH} & 0 \\ -\frac{K_{OH}}{m_O^{1/2}} & 0 & K_{OH} \end{pmatrix}, \quad (28)$$

shares the same structure as CH_2. The eigenvalues $\lambda_{0,1,2}$ of this simple matrix are 0, K_{OH}, and $(1 + 2/m_O)K_{OH}$, respectively, with corresponding eigenvectors $(m_O^{1/2}, 1, 1)$, $(0, 1, -1)$, and $(2m_O^{-1/2}, -1, -1)$. The transformation matrix from $\tilde{r}$ to x is

$$W = \begin{pmatrix} m_O^{-1/2} & & \\ & 1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} m_O^{1/2} & 0 & 2m_O^{-1/2} \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix}$$
$$\times \begin{pmatrix} (m_O+2)^{-1/2} & & \\ & 2^{-1/2} & \\ & & (2+4/m_O)^{-1/2} \end{pmatrix}. \quad (29)$$

The corresponding harmonic potential for water is

$$V_{H_2O}(x_0, x_1, x_2) = \tfrac{1}{2} K_{OH}(\|x_0 - x_1\| - r_{OH})^2$$
$$+ K_{OH}(\|x_0 - x_2\| - r_{OH})^2. \quad (30)$$

This attains its minimum when

$$\|x_0 - x_1\|^2 = \|\tilde{r}_2\|^2 \left(\tfrac{1}{2} + m_{\mathrm{O}}^{-1}\right) - \sqrt{1 + 2m_{\mathrm{O}}^{-1}}\,\tilde{r}_1^T \tilde{r}_2 + \tfrac{1}{2}\|\tilde{r}_1\|^2$$
$$= r_{\mathrm{OH}}^2, \tag{31}$$

$$\|x_0 - x_2\|^2 = \|\tilde{r}_2\|^2 \left(\tfrac{1}{2} + m_{\mathrm{O}}^{-1}\right) + \sqrt{1 + 2m_{\mathrm{O}}^{-1}}\,\tilde{r}_1^T \tilde{r}_2 + \tfrac{1}{2}\|\tilde{r}_1\|^2$$
$$= r_{\mathrm{OH}}^2, \tag{32}$$

where we have used that $x_0 - x_1 = \tilde{r}_2\sqrt{1/2 + 1/m_{\mathrm{O}}} - \tilde{r}_1/\sqrt{2}$ and $x_0 - x_2 = \tilde{r}_2\sqrt{1/2 + 1/m_{\mathrm{O}}} + \tilde{r}_1/\sqrt{2}$. Hence, $\tilde{r}_1$ and $\tilde{r}_2$ must be perpendicular to each other [subtracting Eqs. (31) and (32)] and $\|\tilde{r}_2\|^2(1/2 + 1/m_{\mathrm{O}}) + \tfrac{1}{2}\|\tilde{r}_1\|^2 = r_{\mathrm{OH}}^2$. As these solutions are exact, the hydrogens are always at a distance of r_{OH} from the oxygen. Since there is no angular potential, the solution is indeterminate with two extreme solutions, the first being $\|\tilde{r}_2\| = 0$. In this case, the molecule is linear, while for $\|\tilde{r}_1\| = 0$ the two hydrogens overlap. Adding the constraint that $\|\tilde{r}_1\|^2 = 2r_{\mathrm{OH}} \sin 104.5°$ at equilibrium renders $\|\tilde{r}_2\| = r_{\mathrm{OH}} \cos 104.5°/\sqrt{1/2 + 1/m_{\mathrm{O}}}$.

Another instructive example is HCN. Its harmonic Laplacian

$$\tilde{\Delta}_{\mathrm{HCN}} = \begin{pmatrix} \frac{K_{\mathrm{CH}} + K_{\mathrm{CN}}}{m_{\mathrm{C}}} & -\frac{K_{\mathrm{CH}}}{\sqrt{m_{\mathrm{C}} m_{\mathrm{H}}}} & -\frac{K_{\mathrm{CN}}}{\sqrt{m_{\mathrm{C}} m_{\mathrm{N}}}} \\ -\frac{K_{\mathrm{CH}}}{\sqrt{m_{\mathrm{C}} m_{\mathrm{H}}}} & \frac{K_{\mathrm{CH}}}{m_{\mathrm{H}}} & 0 \\ -\frac{K_{\mathrm{CN}}}{\sqrt{m_{\mathrm{C}} m_{\mathrm{N}}}} & 0 & \frac{K_{\mathrm{CN}}}{m_{\mathrm{N}}} \end{pmatrix} \tag{33}$$

is no longer as simple as for H_2O, nor are the eigenvalues except 0 simple functions of the variables in Eq. (33); for the general AMBER force field (GAFF) [52], they are 14.7 and 42.8. Equation (27) was numerically solved. The numerical GAFF CAM distances in one dimension are $(\tilde{r}_1^{(0)}, \tilde{r}_2^{(0)}) = (3.26, 0.66)$ with $\tilde{r}_1^{(0)} = 2.61$ and $\tilde{r}_2^{(0)} = -1.32$. Knowing that the molecule is linear at equilibrium selects the first solution to reconstruct the equilibrium structure.

In both examples it was necessary to include angle information to make the best choice. The numerical solution to Eq. (27) can be computed efficiently using a variety of nonlinear least-squares algorithms, but more direct methods are still under investigation. Similar derivations are possible for angle potentials.

III. CONCLUSION

We have characterized a coarse-graining procedure for accelerating molecular simulations through a systematic hierarchical algorithm based on multiresolution diffusion wavelets. The proposed wavelet-CG approach goes beyond conventional approaches based on expert knowledge because our proposed method can accelerate calculations of different classes of molecules without requiring extensive expert insight and model parametrization. This advantage is especially important for inverse problems in materials design, wherein the materials engineer aims to optimize material performance in an essentially infinite design space (the chemical space of polymer repeat units). Our demonstration of the perturbation theory for chemical variations in the repeat unit illustrates this key advantage for the wavelet CG approach.

Importantly, these advantages are obtained in a framework that automatically recapitulates the physical insights underlying existing coarse-graining methods such as united-atom models. On the other hand, diffusion-wavelet CG models are simultaneously more general (they do not require *a priori* expert modeling and parametrization) and more specific. In fact, the diffusion-wavelet CG approach leads to system-specific CG models derived automatically from the system's underlying bonding topology and atomistic force field, without further input other than an error tolerance. The systematic and purely algorithmic basis offers the opportunity for adaptive error control, whose obvious importance has motivated significant analysis already [53–55]. It is possible that this framework provides a mechanism to automatically adjust model resolution dynamically in regions that require atomistic accuracy, e.g., at grain boundaries. Further work will establish the relationship between time steps and simulation accuracy [56] and demonstrate algorithms for accelerating MD simulations by compressing force and energy calculations using the wavelet CG basis.

APPENDIX A: DERIVATION OF WAVELET SPACES

The wavelets will be derived iteratively from a filter T using the QR decomposition. At each iteration, T is cast in the Q basis of the previous iteration,

$$\left(\prod_{i=0}^{n} Q_i\right)^{\mathsf{T}} T^{2^{n+1}} \left(\prod_{i=0}^{n} Q_i\right) = Q_{n+1} R_{n+1}. \tag{A1}$$

Therefore, repeated application (with infinite precision) is equivalent to the QR algorithm for finding eigenvectors. Since T is positive definite and $\|T\|_\infty = 1$, the squaring introduces a *de facto* projection operator P_ε via the machine precision. It is this projection that distinguishes a conventional QR algorithm for finding eigenvalues and eigenvectors from the wavelet decomposition into CAMs.

We separate Q_n into a low-frequency submatrix Φ_n and a high-frequency submatrix Ψ_n, where the latter are the columns $Q_{n,i}$ of Q_n for which $Q_{n,i}^T T^{2^n} Q_{n,i} < \varepsilon$. The matrix Φ_n collects the remaining columns of Q_n. Thus,

$$\left(\prod_{i=0}^{n-1} Q_i\right)^{\mathsf{T}} T^{2^{n+1}} \left(\prod_{i=0}^{n-1} Q_i\right) \approx \Phi_n^{\mathsf{T}} Q_n R_n R_n^{\mathsf{T}} Q_n^{\mathsf{T}} \Phi_n^{\mathsf{T}}$$
$$\approx \left(\prod_{i=0}^{n} \Phi_i^{\mathsf{T}}\right) T^{2^{n+1}} \left(\prod_{i=0}^{n} \Phi_i\right).$$

APPENDIX B: ERROR BOUNDS ON SCALES

The contamination of the wavelet spaces W_n by eigenvectors $U_>$ of larger eigenvalues $\lambda^{2^n} > \varepsilon$ of the filter operator T is limited by

$$S_T^{(n)} := \frac{\varepsilon - \omega_{n,\max}}{\omega_{n,\min} - \varepsilon},$$

where $\omega_{n,\min} = \min\{\lambda^{2^n} > \varepsilon | \lambda \in \sigma(T)\}$, $\omega_{n,\max} = \max\{\lambda^{2^n} \leqslant \varepsilon | \lambda \in \sigma(T)\}$, and $\sigma(T)$ denotes the spectrum of T. In classical wavelet theory, wavelets are localized in both real and reciprocal space, e.g., the Fourier transform of the Mexican

hat wavelet transform $Wf(x) = -\sigma^{-2}\Delta \int e^{-(y-x)^2/2\sigma^2} f(y)dy$ is $\tilde{W}\tilde{f}(\omega) = \omega^2 \sigma^{-2} e^{-\omega^2/4\sigma^2} \tilde{f}(\omega)$. The sensitivity $S_T^{(n)}$ of T at scale n measures the localization of W_n in the frequency domain. In practical terms, small $S_T^{(n)}$ implies that the bases Φ_n and Ψ_n are generally sparse matrices if T^{2^n} is sparse, and the wavelet transform can thus be computed efficiently.

APPENDIX C: INVARIANT SUBSPACES

The weighted graph Laplacian $\tilde{\Delta}$ is positive semidefinite and $\tilde{\Delta}_{JJ}M^{1/2}\mathbf{1}_J = 0$ for each set of indices J, where $\tilde{\Delta}_{JJ}$ is the square submatrix of $\tilde{\Delta}$ with indices in J and $\mathbf{1}_J$ is the vector of ones on indices in J and 0 otherwise. It is possible to block tridiagonalize $\tilde{\Delta}$ using only transpositions with diagonal blocks $\tilde{\Delta}_{J,J}$ and off-diagonal blocks $\tilde{\Delta}_{J,K}$, where J and K are disjoint index sets and $J, K, J \cup K$ are contiguous index sets. Without loss of generality, we let $j < k \forall j \in J$ and $k \in K$. The rank of $\tilde{\Delta}_{JK}$ is generally low due to the low maximum degree of a vertex in $\tilde{\Delta}$. If $\Lambda = \text{span } \tilde{\Delta}_{KJ}$ and $\tilde{\Delta}_{JJ}$ has a nontrivial invariant vector space Γ perpendicular to Λ, then Γ is localized to indices preceding K. Examples include linear homopolymers discussed in detail in Sec. II A 4 and Appendix E, but also disconnected graphs from individual molecules. Invariant subspaces $\{\Gamma\}$ are computationally convenient because they allow a separation of the problem into independent smaller subproblems.

APPENDIX D: PERTURBATION THEORY FOR MOLECULAR FRAGMENTS

Changing the mass of atom i by δ leads to $M'^{-1/2} = M^{-1/2} + \delta e_i e_i^T$ and similarly $\tilde{\Delta}$ becomes

$$\tilde{\Delta}' = \left(M^{-1/2} + \delta e_i e_i^T\right)\Delta\left(M^{-1/2} + \delta e_i e_i^T\right) \tag{D1}$$

$$= \tilde{\Delta} + \delta m_i^{1/2}\left(e_i e_i^T \tilde{\Delta} + \tilde{\Delta} e_i e_i^T\right) + \delta^2 m_i \tilde{\Delta}_{ii} e_i e_i^T. \tag{D2}$$

For an eigenvector $\vec{v}$ of $\tilde{\Delta}$ contained in some wavelet space W_n with nondegenerate eigenvalue $\eta_{\vec{v}}$, the first-order correction

to the eigenvalue is $m_i^{1/2}\delta|\vec{v}_i|^2(2\eta_{\vec{v}} + m_i^{1/2}\tilde{\Delta}_{ii}\delta)$ and the first-order correction to the eigenvector is $\sum_{\vec{\mu}\neq\vec{v}}\vec{\mu} \cdot \delta\vec{\mu}_i\vec{v}_i m_i^{1/2}(\eta_{\vec{v}} + \eta_{\vec{\mu}} + \delta m_i\tilde{\Delta}_{ii})/(2\eta_{\vec{\mu}} - 2\eta_{\vec{v}})$, where $\{\vec{\mu}\}$ is an orthonormalized set of eigenvectors of $\tilde{\Delta}$. Hence the scale of $\vec{v}$ is unaffected if $|\vec{v}_i|^2$ is 0 or sufficiently negligible. Corrections to the eigenvector are only of significance if some eigenvector $\vec{\mu}$ outside of W_n additionally has a significant component $\vec{\mu}_i$.

By a similar argument, first-order corrections can be computed for a perturbation of the edge weights, e.g., changing bond spring constants,

$$\tilde{\Delta}' = \tilde{\Delta} + \delta M^{-1/2}\left(e_i e_i^T + e_j e_j^T - e_i e_j^T - e_j e_i^T\right)M^{-1/2}, \tag{D3}$$

$$\eta_{\vec{v}}^{(1)} = \delta\left|m_i^{-1/2}\vec{v}_i - m_j^{-1/2}\vec{v}_j\right|^2, \tag{D4}$$

$$\vec{v}^{(1)} = \sum_{\vec{\mu}\neq\vec{v}} \delta\frac{\left(m_i^{-1/2}\vec{\mu}_i - m_j^{-1/2}\vec{\mu}_j\right)\left(m_i^{-1/2}\vec{v}_i - m_j^{-1/2}\vec{v}_j\right)}{\eta_{\vec{\mu}} - \eta_{\vec{v}}}\vec{\mu}. \tag{D5}$$

Hence, a wavelet subspace W_n only changes if $\exists\vec{v} \in W_n$: $\|T\vec{v}\|_2 + \delta\|\vec{v}\|_\infty \geqslant \varepsilon^{1/2^n}$ or $\|T\vec{v}\|_2 - \delta\|\vec{v}\|_\infty \leqslant \varepsilon^{1/2^{n-1}}$.

APPENDIX E: DERIVATION OF HOMOPOLYMER WAVELETS

The computation of wavelet spaces for linear homopolymers with n repeat units can be subdivided into n subproblems. The recursive structure of linear homopolymers implies that $\tilde{\Delta}$ for a linear homopolymer of n repeat units can be reordered by a permutation κ of indices such that $\kappa^T\tilde{\Delta}\kappa = A \otimes I_n + B \otimes (\Sigma_n^T + e_n e_1^T) + B^T \otimes (\Sigma_n + e_1 e_n^T)$, where I_n is the $n \times n$ identity matrix and Σ_n is the $n \times n$ matrix with all ones on the first subdiagonal only and zeros elsewhere. Then $\kappa^T\tilde{\Delta}\kappa$ takes the simple form

$$\begin{pmatrix} A_{1,1}I_n & A_{1,2}I_n & \cdots & A_{1,m}I_n + B_{1,m}\left(\Sigma_n + e_1 e_n^T\right) \\ A_{2,1}I_n & A_{2,2}I_n & \cdots & A_{1,m}I_n \\ \vdots & & \ddots & \vdots \\ A_{m,1}I_n + B_{1,m}\left(\Sigma_n^T + e_n e_1^T\right) & \cdots & \cdots & A_{m,m}I_n \end{pmatrix}. \tag{E1}$$

Let $U = (e^{i2\pi kl/m}/\sqrt{m})_{kl}$ be a unitary matrix that diagonalizes $A_{1,m}I_n + B_{1,m}(\Sigma_n^T + e_n e_1^T)$; then $\mathbb{U} := I_m \otimes U$ transforms $\kappa^T\tilde{\Delta}\kappa$ into

$$\begin{pmatrix} A_{1,1}I_n & A_{1,2}I_n & \cdots & D^* \\ A_{2,1}I_n & A_{2,2}I_n & \cdots & A_{1,m}I_n \\ \vdots & \vdots & \ddots & \vdots \\ D & A_{m,2}I_n & \cdots & A_{m,m}I_n \end{pmatrix} = \mathbb{U}^*\kappa^T\tilde{\Delta}\kappa\,\mathbb{U}, \tag{E2}$$

where D is a complex diagonal matrix with eigenvalues $D_{k,k} = A_{1,m} + B_{1,m}e^{ik\pi/m}$. Applying the similarity transform by κ^T produces a block diagonal matrix $(\kappa\mathbb{U}^*\kappa^T\tilde{\Delta}\kappa\mathbb{U}\kappa^T)$ with square blocks $\tilde{A}_i = A - (A_{1,m} + D_{i,i})e_1 e_m^T - (A_{1,m} + D_{i,i}^*)e_m e_1^T$. Hence, the computation of the wavelet spaces

has been reduced to n problems of size m instead of one problem of size nm. Furthermore, any eigenvector v of A with $e_m^T v = e_1^T v = 0$ has an n-fold degenerate eigenvalue.

APPENDIX F: DERIVATION OF MIXED RESOLUTION FROM SEPARABLE COARSENINGS

Let $\beta_n^\tau : \mathbb{R}^{m_n} \to \beta_n$ denote a parametrization of the m_n-dimensional state space β_n, e.g., the positions and momenta of particles. A coarsening γ_n can be separated into components if there exist parametrizations $\beta_n^\tau : A_n \oplus B_n \to \beta_n$ and $\beta_{n+1}^\tau : A_{n+1} \oplus B_{n+1} \to \beta_{n+1}$ such that $A_n \oplus B_n = \mathbb{R}^{m_n}$, $A_{n+1} \oplus B_{n+1} = \mathbb{R}^{m_{n+1}}$, and there exist continuous surjective

mappings $\mu_{A_n} : A_n \to A_{n+1}$ and $\mu_{B_n} : B_n \to B_{n+1}$ with

$$\beta_n^\tau\left(\mu_{A_n}^{-1}(t_A), \mu_{B_n}^{-1}(t_B)\right) = \gamma_n^{-1}\left(\beta_{n+1}^\tau(t_A, t_B)\right) \qquad \text{(F1)}$$

for $t_A \in A_{n+1}$ and $t_B \in B_{n+1}$. Here μ_A and μ_B map a fine parametrization to a coarse parametrization. An example parametrization is the representation of V_n as $V_{n+1} \oplus W_{n+1}$ with fine-to-coarse mappings as per Eqs. (9) and (10). See Fig. 7.

A separable coarsening for which neither μ_{A_n} nor μ_{B_n} is a bijection induces intermediate coarsenings. The state space $A_n \times B_{n+1}$ is an intermediate state space with the coarsenings $\gamma : \beta_n^\tau(t_{A_n}, t_{B_n}) \mapsto (t_{A_n}, \mu_{B_n}(t_{B_n})) \in A_n \times B_{n+1}$ and $\gamma' : (t_{A_n}, t_{B_{n+1}}) \mapsto \beta_{n+1}^\tau(\mu_{A_n}(t_{A_n}), t_{B_{n+1}}) \in \beta_{n+1}$. Finally, mixed resolution spaces $A_n \times B_{n+2}$ are induced via function composition, $\gamma : \beta^\tau(t_{A_n}, t_{B_n}) \mapsto (t_{A_n}, \mu_{B_{n+1}} \circ \mu_{B_n}(t_{B_n}))$.

The reconstruction from a fundamental coarsening can be achieved by reconstructing its separable components separately and independently via the conditional probabilities $P(t_{A_n} | t_{A_{n+1}}, t_{B_{n+1}}) = P_{\beta_n}((\{t_{A_n}\} \cap \mu_{A_n}^{-1}(t_{A_{n+1}})) \times \mu_{B_n}^{-1}(t_{B_{n+1}}))/P_{\beta_{n+1}}(t_{A_{n+1}}, t_{B_{n+1}})$. The same can be achieved for the complements analogously.

[1] R. L. C. Akkermans and W. J. Briels, J. Chem. Phys. **115**, 6210 (2001).

[2] J. Baschnagel, K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Mattice, F. Müller-Plathe, M. Murat, W. Paul, S. Santos, U. W. Suter, and V. Tries, *Viscoelasticity, Atomistic Models, Statistical Chemistry*, Advances in Polymer Science Vol. 152 (Springer, Berlin, Heidelberg, 2000), pp. 41–156.

[3] A. P. Lyubartsev, M. Karttunen, I. Vattulainen, and A. Laaksonen, Soft Mater. **1**, 121 (2003).

[4] R. K. Z. Tan, A. S. Petrov, and S. C. Harvey, J. Chem. Theory Comput. **2**, 529 (2006).

[5] C. Clementi, Curr. Opin. Struct. Biol. **18**, 10 (2008).

[6] P. K. Depa and J. K. Maranas, J. Chem. Phys. **123**, 094901 (2005).

[7] N. Goga, M. N. Melo, A. J. Rzepiela, A. H. de Vries, A. Hadar, S. J. Marrink, and H. J. C. Berendsen, J. Chem. Theory Comput. **11**, 1389 (2015).

[8] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidid, O. Runborg, and C. Theodoropoulos, Commun. Math. Sci. **1**, 715 (2003).

[9] P. Ilg, H. C. Öttinger, and M. Kröger, Phys. Rev. E **79**, 011802 (2009).

[10] P. Ilg, J. Non-Newtonian Fluid Mech. **165**, 973 (2010).

[11] J. E. Shepherd, Multiscale modeling of the deformation of semi-crystalline polymers, Ph.D. thesis, Georgia Institute of Technology, 2006.

[12] R. R. Coifman and M. Maggioni, Appl. Comput. Harmon. Anal. **21**, 53 (2006).

[13] J. Bremer, R. Coifman, M. Maggioni, and A. Szlam, Appl. Comput. Harmon. Anal. **21**, 95 (2006).

[14] I. Daubechies, *Ten Lectures on Wavelets, Proceedings of the CBMS-NSF Regional Conference Series in Applied Mathematics* (SIAM, Philadelphia, 1992), Vol. 61.

[15] G. Strang and T. Nguyen, *Wavelets and Filter Banks* (Wellesley Cambridge, Wellesley, 1996).

[16] M. Zhan, S. Li, and F. Li, J. Theor. Comput. Chem. **13**, 1450053 (2014).

[17] M. Maiolo, A. Vancheri, R. Krause, and A. Danani, J. Comput. Phys. **300**, 592 (2015).

[18] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, J. Chem. Theory Comput. **9**, 2466 (2013).

[19] A. Das and H. C. Andersen, J. Chem. Phys. **136**, 194113 (2012).

[20] A. Das and H. C. Andersen, J. Chem. Phys. **136**, 194114 (2012).

[21] C. F. Abrams, L. Delle Site, and K. Kremer, Phys. Rev. E **67**, 021807 (2003).

[22] C. Abrams and K. Kremer, Macromolecules **36**, 260 (2003).

[23] S. J. Marrink, A. H. de Vries, and A. E. Mark, J. Phys. Chem. B **108**, 750 (2004).

[24] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).

[25] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, J. Chem. Theory Comput. **4**, 819 (2008).

[26] A. Gautieri, A. Russo, S. Vesentini, A. Redaelli, and M. J. Buehler, J. Chem. Theory Comput. **6**, 1210 (2010).

[27] B. Fan and J. Maranas, Cellulose **22**, 31 (2015).

[28] L. Lu, S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth, J. Chem. Theory Comput. **6**, 954 (2010).

[29] V. A. Harmandaris, D. Reith, N. F. A. Van der Vegt, and K. Kremer, Macromol. Chem. Phys. **208**, 2109 (2007).

[30] A. J. Rzepiela, L. V. Schäfer, N. Goga, H. J. Risseleda, A. H. de Vries, and S. J. Marrink, J. Comput. Chem. **31**, 1333 (2010).

[31] J. T. Padding and W. J. Briels, J. Phys.: Condens. Matter **23**, 233101 (2011).

[32] P. Espanol, Europhys. Lett. **40**, 631 (1997).

[33] M. Ripoll and P. Espanol, Int. J. Mod. Phys. C **9**, 1329 (1998).

[34] P. Espanol, R. Delgado-Buscalioni, R. Everaers, R. Potestio, D. Donadio, and K. Kremer, J. Chem. Phys. **142**, 064115 (2015).

[35] J. Zhu, R. Klein, and L. Delle Site, Phys. Rev. E **94**, 043321 (2016).

[36] *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, Oxford, 1995).

[37] N. Karayiannis and M. Kröger, Int. J. Mol. Sci. **10**, 5054 (2009).

[38] M. Johnson, T. Head-Gordon, and A. Louis, J. Am. Chem. Soc. **126**, 144509 (2007).

[39] A. Heyden and D. G. Truhlar, J. Chem. Theory Comput. **4**, 217 (2008).

[40] M. Praprotnik, L. Delle Site, and K. Kremer, J. Chem. Phys. **123**, 224106 (2005).

[41] C. Peter, L. Delle Site, and K. Kremer, Soft Matter **4**, 859 (2008).

[42] A. Askar, A. E. Cetin, and H. Rabitz, J. Phys. Chem. **100**, 19165 (1996).

[43] A. E. Ismail, G. C. Rutledge, and G. Stephanopoulos, J. Chem. Phys. **122**, 234901 (2005).

[44] A. E. Ismail, G. Stephanopoulos, and G. C. Rutledge, J. Chem. Phys. **122**, 234902 (2005).

[45] K. C. Wolfe, W. A. Hastings, S. Dutta, A. Long, B. A. Shapiro, T. B. Woolf, M. Guthold, and G. S. Chirikjian, J. Phys. Chem. B **116**, 8556 (2013).

[46] A. Srikanth, R. S. Hoy, B. C. Rinderspacher, and J. W. Andzelm, Phys. Rev. E **88**, 042607 (2013).

[47] T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, and A. C. T. van Duin, NPJ Comput. Mater. **2**, 15011 (2016).

[48] F. R. K. Chung, *Spectral Graph Theory* (American Mathematical Society, Providence, 1997).

[49] H. Sun, S. J. Mumby, J. R. Maple, and A. T. Hagler, J. Am. Chem. Soc. **116**, 2978 (1994).

[50] S. Plimpton, J. Am. Chem. Soc. **117**, 1 (1995).

[51] E. Vazquez and J. Bect, J. Stat. Plan. Infer. **140**, 3088 (2010).

[52] http://ambermd.org/antechamber/gaff.html.

[53] M. A. Katsoulakis, P. Plecháč, L. Rey-Bellet, and D. K. Tsagkarogiannis, J. Non-Newtonian Fluid Mech. **152**, 101 (2008).

[54] M. S. Shell, J. Chem. Phys. **129**, 144108 (2008).

[55] M. A. Katsoulakis and P. Plecháč, J. Chem. Phys. **139**, 074115 (2013).

[56] A. A. Mansour and P. J. Ortoleva, J. Chem. Theory. Comput. **12**, 1965 (2016).