ARL-TR-7989 • Apr 2017

US Army Research Laboratory

Theory of Wavelet-Based Coarse-Graining Hierarchies for Molecular Dynamics

by Berend Christopher Rinderspacher, Jaydeep P Bardhan, and Ahmed E Ismail

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7989 ● APR 2017

US Army Research Laboratory

Theory of Wavelet-Based Coarse-Graining Hierarchies for Molecular Dynamics

by Berend Christopher Rinderspacher
Weapons and Materials Directorate, ARL

Jaydeep P Bardhan
Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA

Ahmed E Ismail
Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
Apr 2017	Technical Report	October 2013-September 2016

4. TITLE AND SUBTITLE
Theory of Wavelet-Based Coarse-Graining Hierarchies for Molecular Dynamics

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Berend Christopher Rinderspacher, Jaydeep P Bardhan, and Ahmed E Ismail

5d. PROJECT NUMBER
FY14-WMR-038

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-WMM-G
Aberdeen Proving Ground, MD 21005-5066

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7989

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
primary author's email: berend.c.rinderspacher.civ@mail.mil

14. ABSTRACT
We present a multiresolution approach to compressing the degrees of freedom (DoFs) and potentials associated with molecular dynamics (MD). We suggest a systematic way to accelerate large-scale MD with more than 2 levels of coarse-graining, particularly for simulation of polymeric materials. We derive explicit models for linear 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 (CG) mappings, but instead uses diffusion wavelets for graph Laplacians to develop system-specific mappings. Our methodology leads to a hierarchy of system-specific CG DoFs that provide a conceptually clear and rigorous framework for modeling chemical systems at relevant model scales. The approach is capable of automatically generating as many CG model scales as necessary, that is, to go beyond the 2 scales in conventional CG strategies. Furthermore, the wavelet-based CG models explicitly link time and length scales. Finally, a straightforward method to introduce omitted DoFs is presented, which plays a major role in maintaining model fidelity in long-time simulations and capturing emergent behaviors.

15. SUBJECT TERMS
molecular dynamics, multiresolution, emergent physics, simulation, diffusion wavelets, polymers

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			B. Christopher Rinderspacher
Unclassified	Unclassified	Unclassified	UU	56	19b. TELEPHONE NUMBER (Include area code) 410-306-2811

Contents

List of Figures

Acknowledgments

Dr BC Rinderspacher would like to thank Dr Bryan Glaz for his insights and advice in determining the links between molecular dynamics and Koopman operators. He also wishes to thank Drs Bardhan and Ismail whose advice and dedication was indispensable.

INTENTIONALLY LEFT BLANK.

1. Introduction

Among the challenges faced in molecular simulations is balancing the competing needs of computational complexity and fidelity to the underlying chemical and physical phenomena being studied. 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 (DoFs) often leads to minor gains in the time steps that can be used in dynamic simulations[6], such improvements have not led to significant breakthroughs in simulation capabilities. More advanced coarse-graining techniques are required to enable simulations of systems exhibiting structure on the micron or submillimeter scales, such as semicrystalline materials.

In this report, we present a new approach toward that end, focusing on polymeric materials composed of homopolymers or block copolymers. Our approach is based on the concept of diffusion wavelets[7,8], which both automatically identifies chemical structures that can be reduced into coarse-grained (CG) units and also allows for repeated application, thus providing much greater customization of the level of simulation. Such a technique enables the potential reduction of the number of DoFs by orders of magnitude, as well as for a vastly increased time step, which, when taken together, permits vastly extended time scales to be simulated using currently available computational resources.

Among the shortcomings of typical coarse-graining approaches has been that they normally involve only 2 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 DoFs, which typically replace the atomistic groupings with a single "bead"-like entities[9–15]. While approaches such as the force-matching method of Voth et al.[16] and the reversible coarse-graining method for phenolic polymers by Kremer et al.[9,10,17] 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, because particle-like CG beads lose information about the particles they subsume, which has to be recaptured by other means (e.g., the use of dummy variables as in dissipative-particle dynamics with energy conservation[18]).

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 2-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[19]. However, coarse-graining need not be restricted only to polymers; even relatively "simple" fluids, such as water, have been the subject of CG models.

Many of the previously mentioned problems can be alleviated through the selection of 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 delta functions). A more sophisticated approach, suitable for analyzing certain average properties of the signal, might employ the Fourier basis, sines of varying frequency. However, the latter basis is global in nature, challenging processing and storage for signals that are infrequent or have sharply varying features only over 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[20,21]. 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.[22–24]. By analogy to the time-varying signal, purely atomistic models and standard 2-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 in-

finitely 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 $O(\log_2 N)$, 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[21]. Our application of wavelet ideas to *structural* representation extends the work of Ismail[25,26]. In particular, the earlier work employed Monte Carlo and did not capture dynamical information. The work here develops an approach suitable for MD simulation by interpreting the wavelet transform of Ismail 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 report; follow-on work will illustrate more concrete applications. 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 both to structures of arbitrary chemical complexity, including both relatively simple molecules such as water or benzene, as well as more complicated chain molecules such as polymers and biopolymers. Finally, it is also capable of capturing minor effects such as mass effects from chemical substitution (e.g., 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 hydrogen cyanide (HCN) before summarizing our conclusions.

2. 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

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

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

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

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.

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

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

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

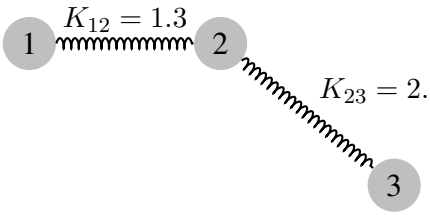
where $\mathbf{1}$ denotes a vector of all ones and $\mathrm{diag}(v)$ denotes a diagonal matrix defined

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 Figure 1.

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

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

where $V'(x) = \nabla V(x) - \Delta x$ 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 (Figure 1).



$\rightarrow \Delta =$

	1	2	3
1	1.3	-1.3	0
2	-1.3	3.4	-2.1
3	0	-2.1	2.1

Fig. 1 Example of a weighted graph Laplacian derived from a simple weighted graph

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 $\left(\|p\|^2 + \|q\|^2\right)/2$ and the equations of motion become

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

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,

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

where

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

It can be shown that $\tilde{\Delta}$ is positive semi-definite, and that the eigenvectors of the exponential operator in Eq. 7 are $(U_j, \pm iU_j)$, 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.

2.1 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[7]. 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 (Figure 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 is 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 concomittantly, the minimum time step size roughly doubles.

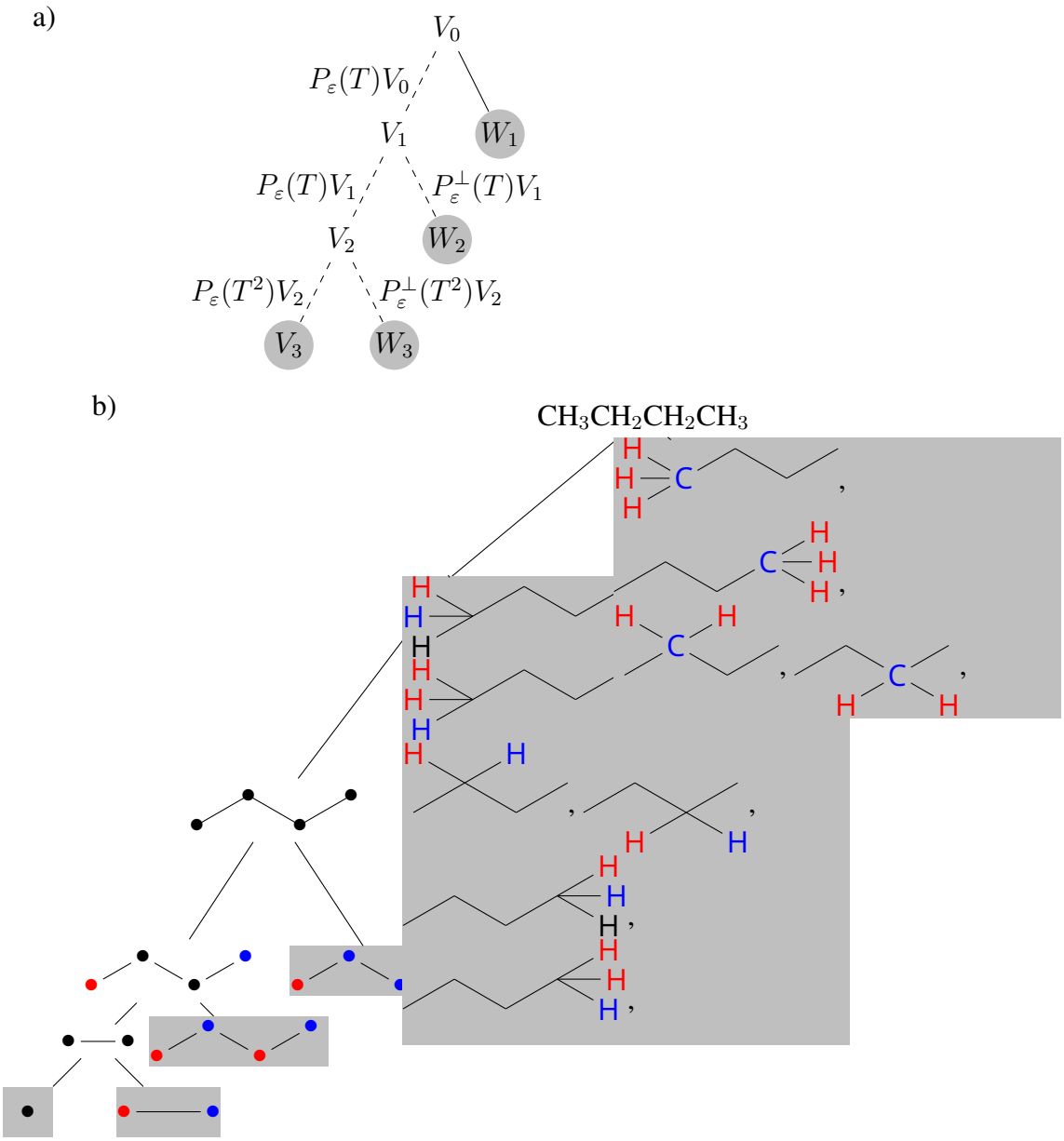


Fig. 2 Dyadic trees generated by multiresolution formalism using a filter T. Shaded spaces are final subspaces of the multiresolution. All other spaces are intermediates of the construction. a) Generic scheme; and b) an example using butane. In the final spaces in gray, red and blue denote opposite signs of weights in the construction from the finer scale.

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

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

and

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

are iteratively associated with orthonormal bases via QR-decompositions,

$$T^{2^n} = Q_n R_n, \tag{11}$$

where Q_n is unitary, R_n is upper triangular, and both are dependent on the basis used to express T^{2^n}.

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 $\|T\|_\infty \leq 1$. The wavelet space at scale n (W_n) is approximately spanned by eigenvectors of T whose eigenvalues obey

$$2^{-n+1} \ln \varepsilon < \ln \lambda \leq 2^{-n} \ln \varepsilon. \tag{12}$$

Only a limited degree of mixing with eigenvectors associated with eigenvalues outside of these bounds is possible (see Section 9 for details). Since T is positive semi-definite, 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 $\ln T$.

2.1.1 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 semi-definite. 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 $[0, 2\max_i \Delta_{ii}]$.[27] 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, sufficiently coarse scales ($n \gg m$) are quasi-static 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 us 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 $\log_2 N$ scales, where N is the dimensionality of $\tilde{\Delta}$. This leads to many DoFs per scale due to issues discussed in the following sections. 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 δ_{max}, $\varepsilon_{\text{effective}} = \varepsilon_{\text{machine}}^{2^{-m}} \approx 1 - \delta_{max}$.

2.1.2 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 methylene (CH_2) group has an associated medium-frequency, highly localized eigenvector of $\tilde{\Delta}$. This follows from the fact that the hydrogens are leaves on the graph (i.e., the weighted subgraph Laplacian

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

where o collects the contributions from further bonds with the carbon atom, shows

only contributions from CH_2 for the hydrogens). If the hydrogens 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 (Figure 3), which has an invariant subspace spanned by 2 independent eigenvectors. The repeat unit block of the weighted graph Laplacian $\tilde{\Delta}$ is

$$A = \begin{pmatrix} \frac{2X+Z_1}{12} & -\frac{X}{\sqrt{12\cdot 14}} & 0 & -\frac{X}{\sqrt{12\cdot 14}} & & \\ -\frac{X}{\sqrt{12\cdot 14}} & \frac{X+Y}{14} & -\frac{Y}{\sqrt{12\cdot 14}} & & & \\ 0 & -\frac{Y}{\sqrt{12\cdot 14}} & \frac{X+Y}{14} & & & -\frac{X}{\sqrt{12\cdot 14}} \\ -\frac{X}{\sqrt{12\cdot 14}} & & & \frac{X+Y}{14} & -\frac{Y}{\sqrt{12\cdot 14}} & 0 \\ & & & -\frac{Y}{\sqrt{12\cdot 14}} & \frac{X+Y}{14} & -\frac{X}{\sqrt{12\cdot 14}} \\ & & -\frac{X}{\sqrt{12\cdot 14}} & 0 & -\frac{X}{\sqrt{12\cdot 14}} & \frac{2X+Z_6}{12} \end{pmatrix}, \qquad (14)$$

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 2 vectors $(0, 1, 1, -1, -1, 0)$ and $(0, 1, -1, 1, -1, 0)$ span a subspace invariant under application of A but which 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 Section 10 for details of the general case).

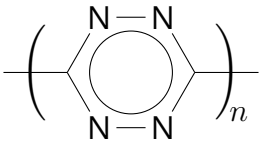


Fig. 3 Chemical structure of polytetrazine

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 because no choice of accuracy can be used to separate them into subscales. In such cases, it may be possible to incorporate more information from the potential $\tilde{V}$, but this is outside of the current scope.

As an example, butane ($H(CH_2)_4H$) has 14 atoms and, therefore, the graph Lapla-

cian associated with $H(CH_2)_4H$ is the 14-dimensional square matrix

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

where $\kappa = 44/31$ is the ratio of force constants for a CC-bond to a CH-bond in the polymer-consistent force field (PCFF)[28] and the units of kcal/mol/Å^2 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 6 scales. The first 4 applications $(T, T^2, T^4, T^8, T^{16})$ did not lead to any unit vectors below the threshold. At $n = 5$, the 4 highest frequency modes of Δ (λ = 517.1, 524.0, 556.3, and 556.3 kcal/mol/Å^2, respectively),

$$\begin{pmatrix}
H & C & H & H & C & H & H & C & H & H & C & H & H & H \\
\boxed{-0.50} & \boxed{0.46} & \boxed{-0.50} & \boxed{-0.50} & & & & & & & & & & \\
 & & & & \boxed{0.33} & \boxed{-0.53} & \boxed{-0.53} & \boxed{0.21} & \boxed{-0.36} & \boxed{-0.36} & & & & \\
 & & & & \boxed{-0.24} & \boxed{0.35} & \boxed{0.35} & \boxed{0.35} & \boxed{-0.52} & \boxed{-0.52} & & & & \\
 & & & & & & & & & & \boxed{-0.46} & \boxed{0.50} & \boxed{0.50} & \boxed{0.50}
\end{pmatrix}, \tag{16}$$

approximately map to zero for T^{32}, that is, the logarithm of the expectation values with T do not exceed $2^{-5} \ln \varepsilon \approx -1.13$ ($\log(1 - \lambda/660) < -1.53$) . These CAMs correspond to the symmetric stretches of the methyl groups.

The 6 second-highest frequencies (λ =440 kcal/mol/Å^2 for all, $\log(1 - \lambda/660) \approx$

-1.10),

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

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 2 rows cover the degeneracy between the 3 hydrogens within the respective methyl groups, whereas the third and fourth rows show isolated HH vibration modes.

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

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

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 (λ =42.1 kcal/mol/Å^2, $\log(1 - \lambda/660) \approx -0.066$),

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

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/Å^2, $\log(1 - \lambda/660) \approx -0.019$),

$$\left(\begin{array}{cccccccccccccc} H & C & H & H & C & H & H & C & H & H & C & H & H & H \\ \hline -0.17 & -0.59 & -0.17 & -0.17 & -0.24 & 0 & 0 & 0.24 & 0 & 0 & 0.59 & 0.17 & 0.17 & 0.17 \end{array}\right), \tag{20}$$

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/Å^2) is described by

$$\left(\begin{array}{cccccccccccccc} H & C & H & H & C & H & H & C & H & H & C & H & H & H \\ \hline 0.13 & 0.45 & 0.13 & 0.13 & 0.45 & 0.13 & 0.13 & 0.45 & 0.13 & 0.13 & 0.45 & 0.13 & 0.13 & 0.13 \end{array}\right). \tag{21}$$

This last CAM is just the center of mass.

2.1.3 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 composed of only 20 amino acids. Hence, it is instructive to characterize the effects of modifying bonds or substituting different atoms. Indeed, connecting fragments (see Section 11 for details) affect a small portion of 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.

2.1.4 Example: Linear Homopolymers

Linear homopolymers are an important class of materials, whose graph Laplacians exhibit convenient structures that we 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 $B_{1,m}$ (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 (Section 12).

2.2 Computational Methods

The following simulation exemplifies the effectiveness of the approach. All MD simulations were performed in the the Large-scale Atomistic/Molecular Massively Parallel Simulator (LAMMPS)[29]. A crystalline model of polyethylene (PE) consisting of 100,800 atoms in a single infinite chain was simulated in the NVT ensemble under periodic boundary conditions with 1-fs time steps, PCFF[28] at $500\,\mathrm{K}$. The monoclinic unit cell has edge vectors $\mathbf{a} = (103.432\,, 0, 0)$, $\mathbf{b} = (0, 147.87\,, 0)$, and $\mathbf{c} = (73.88\,, 0, 50.78\,)$. The initial structure is depicted in Figure 5.

2.3 Simulation Results and Analysis

We witnessed this behavior in computations of PE as well as polyglucose. Furthermore, when applied to the alanine dipeptide model, commonly used for testing coarse-graining methods for proteins, the wavelet decomposition correctly identified the various chemical groups at the finer scales, while the coarse scales captured the partitioning corresponding to single bond rotations. Figure 4 illustrates the decomposition for tetra-1,4-D-glucose and Figure 5 demonstrates the informational content in the wavelet DoFs for PE.

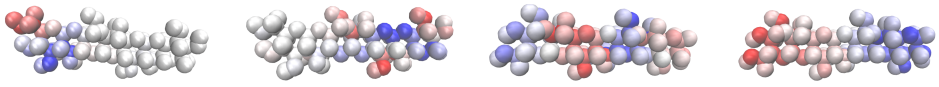


Fig. 4 Heat map of selected wavelet DoFs in 1,4-D-glucose tetramer. Blue denotes positive coefficients, red negative. From left to right: fine-grained wavelet DoF covering a single repeat unit; CG wavelet DoF covering one half of the oligomer; coarser wavelet DoF covering the oligomer isolating repeat units with sign changes; coarsest wavelet.

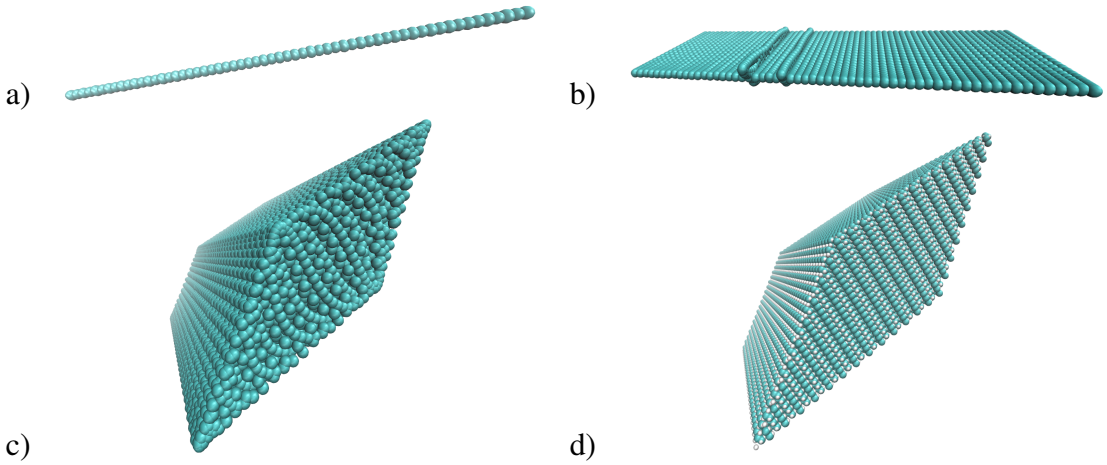


Fig. 5 Selected reduced wavelet representations of a PE crystal. From a) to d) wavelet information is successively added. a) Coarse representation highlights the anisotropy in 1-D chain averaged over all chain segments; b) Coarse representation of the 2-D chain plane; c) 3-D representation; d) full resolution.

Figure 6 shows the superimposed Fourier-transformed time series of 1,000 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 are active, most importantly around $0.45\,\mathrm{PHz}$, which limits the time step of atomistic simulations of PE to $1.2\,\mathrm{fs}$ or less.

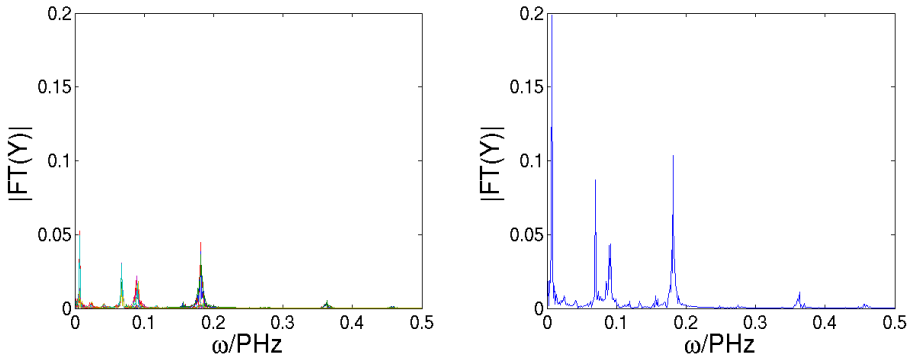


Fig. 6 Fourier transform of the y-component of 1,000 atoms in crystalline PE (100,800 atoms) excluding zero frequency to allow detail at other frequencies. MD at $500\,\mathrm{K}$ and $1\,\mathrm{atm}$. Left: Individual power spectra per atom. Right: Power spectrum of magnitude of optimal representation.

On the other hand, Figure 7 shows the effects of scales on time-series analysis. The finest scale out of 25 still retains the high-frequency components (top, left) 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 (top, right), no high-frequency components are found anymore. Therefore, medium scale and coarser DoFs are quasi-static 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 speed-up by reducing not only DoFs but also increasing the propagation time step.

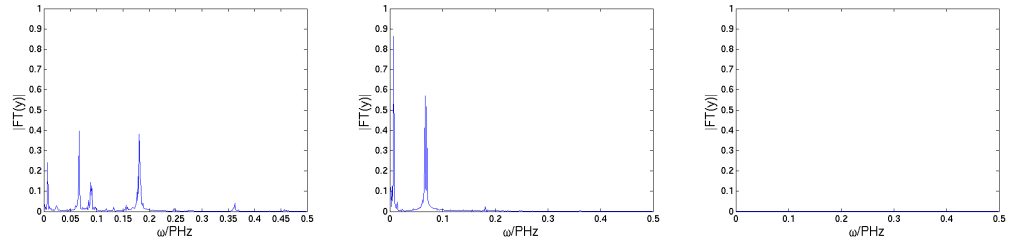


Fig. 7 Top row: Fourier transform of the y-component of a 100,800 atom crystalline PE sampled at $1\,\mathrm{fs}$. 3 scales are shown: 1st (left, finest scale), 5th (middle), and 12th scale (out of 25, right).

The same is witnessed for alanine dipeptide and polyglucose (not shown). Figure

8 shows the power spectra for all 22 DoFs. Clearly, a large number of frequencies are active for all atoms. The spectrum of optimal representation shows some clear peaks around 0.13 and 0.10 PHz. But in Figure 9, the 0.13-PHz peak is absent after the second scale and the 0.10-PHz peak including noise down to about 0.06 PHz vanishes from the fourth scale on through the remaining 4 scales.

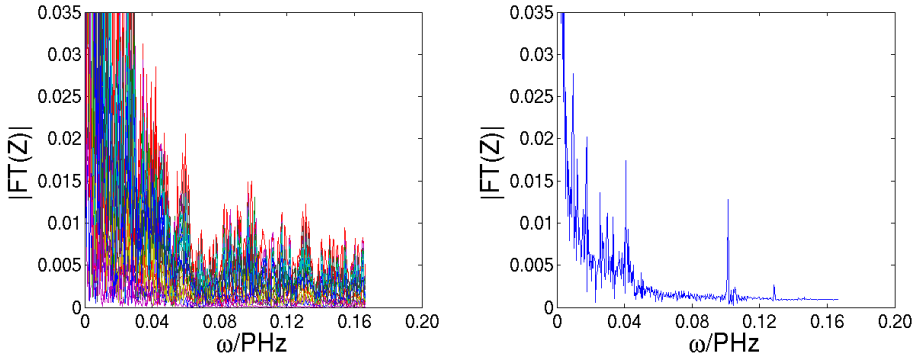


Fig. 8 Fourier transform of the z-component of alanine dipeptide in vacuum excluding zero frequency to allow detail at other frequencies. MD at $500\,\mathrm{K}$ and $1\,\mathrm{atm}$. Left: Individual power spectra per atom. Right: Power spectrum of magnitude of optimal representation.

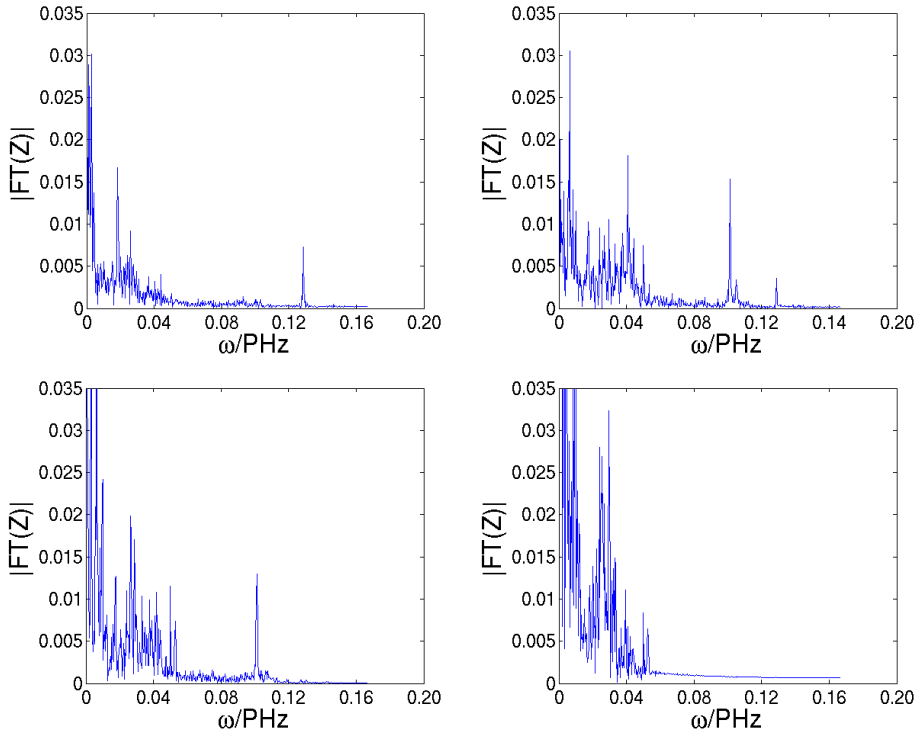


Fig. 9 Fourier transforms of alanine dipeptide DoF time series in vacuum excluding zero frequency to allow detail at other frequencies. MD at $500\,\mathrm{K}$ and $1\,\mathrm{atm}$. a) z-component of optimal representation; b) second finest optimal wavelet DoF z-component; c) third finest optimal wavelet DoF z-component; d) fourth finest optimal wavelet DoF z-component.

We investigated the correlation matrices of the various coordinate systems to quantify the extent of interdependence. Figure 10 shows the correlation matrices,

$$C_{ij} = \frac{\int x_i^*(t) x_j(s-t)\, ds\, dt}{\left(\int x_i^*(t) x_i(s-t)\, ds\, dt \int x_j^*(t) x_j(s-t)\, ds\, dt\right)^{\frac{1}{2}}}, \tag{23}$$

where $x_i(t)$ is a time-dependent coordinate, as heat maps. As may be expected, all particle coordinates are strongly correlated, but in both the water case as well as the alanine dipeptide case, much less correlation is witnessed for the wavelet DoFs (Figure 10, right), despite the few DoFs involved in these small systems. The extent of overall correlation can be assessed by the ℓ_1-norm of the correlation matrices, about 28 versus about 21 for water, and about 315 versus about 168 for alanine dipeptide.

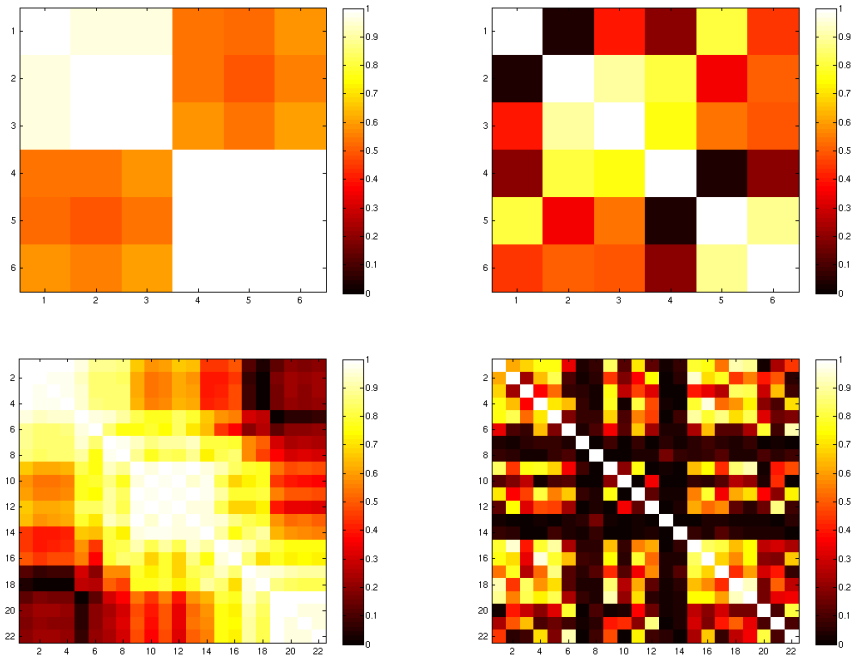


Fig. 10 Top: Heat maps of the DoF correlations of a water dimer in vacuum. Bottom: Heat maps of the DoF correlations of alanine dipeptide in vacuum. Left: Correlations of the cartesian DoFs. Right: Correlations of the wavelet DoFs.

A further indication of the appropriateness of the diffusion wavelet DoFs, is the ℓ_1-norm of structures and forces in a given representation basis. The ℓ_1-norm is bounded by the ℓ_2-norm for any orthonormal basis and there exists at least one coordinate system in which $\|x\|_1 = \|x\|_2$. These coordinate systems optimally describe a state x.

We compared the ℓ_1-norm in the Cartesian coordinate system, the nonlossy wavelet coordinate system, and the optimal coordinate system on each scale. For alanine dipeptide, a reduction by a factor of about 2 was achieved by switching to wavelets (3 times the optimum), while the optimal wavelet system managed to reduce to twice the optimum. For polyglucose, the full wavelet DoFs reduce the ℓ_1-norm to half of the Cartesian coordinate system, and the optimal wavelet system reduces it to within 20% of the optimum. For crystalline PE, the full wavelet DoFs improve the structural ℓ_1-norm by a factor of about 7, while the optimal wavelet representation comes in below 1% of the Cartesian coordinate system to a factor of roughly twice the optimal representation. Similar results were obtained for amorphous PE and

nanocellulose.

2.4 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 like 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 an iterative Boltzmann inversion) as well as analytical (where such exists, e.g., quadratic potentials) coarse-graining hierarchies.

2.4.1 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 2 topological state spaces α, the fine space, and β, the coarse space, which can be parameterized 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. 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 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. Com-

mon solutions in the MD community are probability measures from iterative Boltzmann inversions 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[30] 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).

2.4.2 Coarse-Graining Hierarchies

In the context of MD, mixed resolutions are necessary for the concurrent treatment of, for instance, 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 Section 13 and Figure 11 for details).

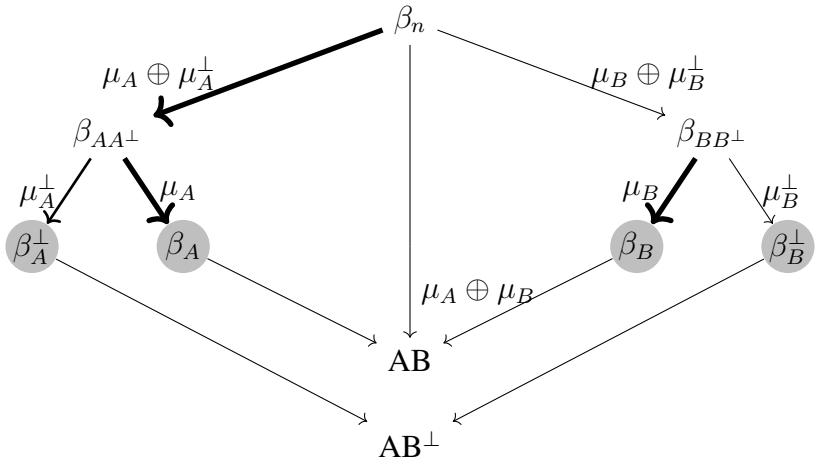


Fig. 11 Dyadic tree generated by separable coarsenings. Shaded spaces are final subspaces of the multiresolution. All other spaces are intermediates of the construction with $\gamma_1 = \mu_A \oplus \mu_A^\perp$ and $\gamma_2 = \mu_A \oplus \mu_B$.

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 previously.

2.4.3 The 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 $V_m = (\bigoplus_{n=1}^{\lfloor 2^{-m}N \rfloor} W_{2^m n-1})_{m=0}^{\log_2 N}$ is

$$\mu_{A_m} : \bigoplus_{n=1}^{\lfloor 2^{-m-1}N \rfloor} W_{2^m n-1} \to \bigoplus_{n=1}^{\lfloor 2^{-m-2}N \rfloor} W_{2^{m+1} n-1}, \tag{24}$$

$$\mu_{B_m} = \bigoplus_{n=\lfloor 2^{-m-1}N \rfloor+1}^{\lfloor 2^{-m}N \rfloor} W_{2^m n-1} \to \bigoplus_{n=\lfloor 2^{-m-2}N \rfloor+1}^{\lfloor 2^{-m-1}N \rfloor} W_{2^{m+1} n-1}, \tag{25}$$

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 $P_\varepsilon(T^{2^n})$).

2.4.4 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, both in the orginal and transformed basis of the potential,

$$V_{\text{bond}}(M^{-\frac{1}{2}}U^{\mathsf{T}}\tilde{r}) \approx \frac{1}{2}(\tilde{r} - \tilde{r}_0)^{\mathsf{T}} U \tilde{\Delta} U^{\mathsf{T}} (\tilde{r} - \tilde{r}_0), \tag{26}$$

where $\tilde{r} = UM^{1/2}x$ is the position vector in the wavelet basis, and $UM^{1/2}$ 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

$$V_{\text{bond}} \approx \frac{1}{8} \sum_{ij} \frac{K_{ij}}{r_{ij}^{(0)^2}} \left(\|x_i - x_j\|^2 - r_{ij}^{(0)^2} \right)^2 . \tag{27}$$

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

$$\min_{\tilde{r}_0} \sum C_{ij} \left(\|(M^{-\frac{1}{2}} U^T \tilde{r}_0)_i - (M^{-\frac{1}{2}} U^T \tilde{r}_0)_j\|^2 - r_{ij}^{(0)^2} \right)^2 , \tag{28}$$

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

For example, the 2 nonzero eigenvalues of H_2O correspond to a unique solution for reconstructing H_2O. The harmonic Laplacian for H_2O,

$$\tilde{\Delta}_{H_2O} = \begin{pmatrix} \frac{2K_{OH}}{m_O} & -\frac{K_{OH}}{m_O^{\frac{1}{2}}} & -\frac{K_{OH}}{m_O^{\frac{1}{2}}} \\ -\frac{K_{OH}}{m_O^{\frac{1}{2}}} & K_{OH} & 0 \\ -\frac{K_{OH}}{m_O^{\frac{1}{2}}} & 0 & K_{OH} \end{pmatrix}, \tag{29}$$

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

$$W = \begin{pmatrix} m_O^{-\frac{1}{2}} & & \\ & 1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} m_O^{\frac{1}{2}} & 0 & 2m_O^{-\frac{1}{2}} \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} (m_O + 2)^{-\frac{1}{2}} & & \\ & 2^{-\frac{1}{2}} & \\ & & (2 + 4/m_O)^{-\frac{1}{2}} \end{pmatrix}. \tag{30}$$

The corresponding harmonic potential for water is

$$V_{H_2O}(x_0, x_1, x_2) = \frac{1}{2} K_{OH} \left(\|x_0 - x_1\| - r_{OH} \right)^2 + K_{OH} \left(\|x_0 - x_2\| - r_{OH} \right)^2 . \tag{31}$$

This attains its minimum when

$$\|x_0 - x_1\|^2 = \|\tilde{r}_2\|^2 \left(\frac{1}{2} + m_O^{-1}\right) - \sqrt{1 + 2m_O^{-1}}\tilde{r}_1^T\tilde{r}_2 + \frac{1}{2}\|\tilde{r}_1\|^2 \quad = r_{OH}^2, \quad (32)$$

$$\|x_0 - x_2\|^2 = \|\tilde{r}_2\|^2 \left(\frac{1}{2} + m_O^{-1}\right) + \sqrt{1 + 2m_O^{-1}}\tilde{r}_1^T\tilde{r}_2 + \frac{1}{2}\|\tilde{r}_1\|^2 \quad = r_{OH}^2, \quad (33)$$

where we have used that $x_0 - x_1 = \tilde{r}_2\sqrt{1/2 + 1/m_O} - \tilde{r}_1/\sqrt{2}$, $x_0 - x_2 = \tilde{r}_2\sqrt{1/2 + 1/m_O} + \tilde{r}_1/\sqrt{2}$. Hence, $\tilde{r}_1$ and $\tilde{r}_2$ must be perpendicular to each other (subtracting Eqs. 32 and 33) and $\|\tilde{r}_2\|^2 (1/2 + 1/m_O) + \frac{1}{2}\|\tilde{r}_1\|^2 = r_{OH}^2$. 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 2 extreme solutions, the first being $\|\tilde{r}_2\| = 0$. In this case, the molecule is linear, while for $\|\tilde{r}_1\| = 0$ the 2 hydrogens overlap.

Another instructive example is HCN. Its harmonic Laplacian,

$$\tilde{\Delta}_{HCN} = \begin{pmatrix} \frac{K_{CH}+K_{CN}}{m_C} & -\frac{K_{CH}}{\sqrt{m_C m_H}} & -\frac{K_{CN}}{\sqrt{m_C m_N}} \\ -\frac{K_{CH}}{\sqrt{m_C m_H}} & \frac{K_{CH}}{m_H} & 0 \\ -\frac{K_{CN}}{\sqrt{m_C m_N}} & 0 & \frac{K_{CN}}{m_N} \end{pmatrix}, \quad (34)$$

is no longer as simple as for H_2O, nor are the eigenvalues except 0 simple functions of the variables in Eq. 34; for the generalized amber force field (GAFF), they are 14.7 and 42.8. Eq. 28 was numerically solved. The numerical GAFF CAM distances in one dimension are $(\tilde{r}_1^{(0)}, \tilde{r}_2^{(0)}) = (3.26, 0.66)$ and $\tilde{r}_1^{(0)} = 2.61, \tilde{r}_2^{(0)} = -1.32$. 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. 28 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 and under current investigation.

3. Operator Wavelets Theory

To perform efficient wavelet representations of DoFs, there also needs to be an appropriate effective operator that can exploit the wavelet structure of the CAMs. Ideally, this operator can itself be constructed in a hierarchical fashion. The following explores the properties of operator hierarchies that arise from doubling and shift operators. Repeated application of an invertible doubling operator $D : \mathscr{L}_1 \to \mathscr{L}_1$ and an invertible shift operator $\sigma : \mathscr{L}_1 \to \mathscr{L}_1$ on an operator space $\mathscr{L}_1$ generates a scaling operator Φ analogous to a scaling function in wavelet theory on functions. Then,

$$D(I + \sigma)\Phi = \Phi. \tag{35}$$

Since the identity is subject to D and $D\sigma$, there are associated operators $d \in \mathscr{L}_1$ and $ds \in \mathscr{L}_1$, respectively. If furthermore D and σ are endomorpisms on $\mathscr{L}_1$, then

$$(d + ds)\Phi\alpha = (DI + D\sigma I)(D\Phi + D\sigma\Phi)\alpha = \Phi\alpha \Rightarrow \tag{36}$$

$$(d + ds)\alpha = \alpha \tag{37}$$

for any eigenfunction α of Φ. Hence the eigenfunctions of Φ are scaling functions under the doubling operator d and shift operator s. The corresponding operator wavelets are

$$\Psi_{nk} = D^n\sigma^k D(I - \sigma)\Phi, \tag{38}$$

$$\psi_{\alpha,n,k} = d^n s^k (d - ds)\alpha. \tag{39}$$

An approximate procedure to develop a hierarchy that is not based on endomorphism on $\mathscr{L}_1$ is to use the doubling operator $\Phi \mapsto 2\Phi$ and use polynomial filtering on the spectrum (i.e., $\Phi \mapsto F(\Phi) = \sum_i F_i\Phi^i$ where the polynomial F approximates a low-pass filter). Under this construction any scaling operator is unitarily equivalent to any other (i.e., they share the same spectrum).

4. Trotter Factorization for Use with Operator Wavelets

A hierarchical decomposition of the infinitesimal generator of time propagation (Liouville operator) into operator wavelets enables the separation of time scales rigorously using Trotter factorizations for numerical propagation. We split the Liouville

operator into

$$\mathscr{H} = \sum_{i=0}^{N} H_i + \sum_{i,j\in\{0,\ldots,N\}} H_{ij}, \tag{40}$$

where H_i are individually solved subsystems and H_{ij} are the interactions between these, such as the operator wavelets introduced in Section 3. Preferably, these partitions coincide with inherent time scales t_i and t_{ij}. Let the time scales be in ascending order and readjusted to the closest integer multiple (i.e., $\tilde{t}_{I+1} = \left\lfloor t_{I+1}/\tilde{t}_I \right\rfloor \tilde{t}_I$) and $\tau = \max \tilde{t}_I$. If $\tilde{t}_{I+1} = \tilde{t}_I$, it is useless to separate H_{I+1} and H_I. So then,

$$e^{i\mathscr{H}\tau} \approx \prod_{I=0}^{N} \left(e^{iH_I\tilde{t}_I/2}\right)^{\tau/\tilde{t}_I} \prod_{I=N}^{0} \left(e^{iH_I\tilde{t}_I/2}\right)^{\tau/\tilde{t}_I}. \tag{41}$$

When the time propagator $e^{iH_I\tilde{t}_I}$ has analytical solutions, such as with harmonic oscillators, then the analytical solution $e^{iH_I\tau/2}$ can be used instead of the $\tau/\tilde{t}_I$ applications of the approximation. If $\tau/\tilde{t}_I$ is large, considerable savings may be expected.

In particular, consider a quadratic potential with matrix $\tilde{\Delta}$ (i.e., $i\mathscr{H}_q = -(x - \overline{x})^T\tilde{\Delta}\frac{d}{dp} + p^T M^{-1}\frac{d}{dx}$), then

$$e^{i\mathscr{H}\tau} \approx e^{i(\mathscr{H}-\mathscr{H}_q)\tau/2} e^{i\mathscr{H}_q\tau} e^{i(\mathscr{H}-\mathscr{H}_q)\tau/2}. \tag{42}$$

The middle term,

$$\begin{pmatrix} x(\tau) \\ p(\tau) \end{pmatrix} = e^{i\mathscr{H}_q\tau} \begin{pmatrix} x \\ p \end{pmatrix} = e^{\begin{pmatrix} & I \\ -\tilde{\Delta} & \end{pmatrix}\tau} \begin{pmatrix} x(0) - \overline{x} \\ p(0) \end{pmatrix} + \begin{pmatrix} \overline{x} \\ 0 \end{pmatrix}, \tag{43}$$

can be solved analytically, so any time interval τ can be computed exactly, whereas

$$\mathscr{H} - \mathscr{H}_q = \left(-\nabla^T V + (x - \overline{x})^T\tilde{\Delta}\right)\frac{d}{idp} \tag{44}$$

requires only a single force evaluation similar to other reversible reference system propagator algorithms.

To leverage this factorization, an appropriate origin $\overline{x}$ needs to be assessed either by running short bootstrapping trajectories or analytical expressions specific to the system. For high frequencies, the propagation by $\mathscr{H}_q$ may wildly oscillate through

the origin. Which origin is optimal may actually depend on the amount of energy in the quadratic term.

5. Solutions to "Harmonic" Oscillators

The Trotter factorization in Section 4 is most useful if analytical solutions exist for the involved operators. The Liouville operator $\mathscr{L}$ for bond oscillators in MD is

$$i\mathscr{L} = \sum_{ij} -k_{ij}\left(\|x_i - x_j\| - r_{ij}^{(0)}\right)\frac{x_i - x_j}{\|x_i - x_j\|}\frac{\partial}{\partial p_i} + \sum_i \frac{p_i}{m_i}\frac{\partial}{\partial x_i}. \tag{45}$$

We set $\eta = Ex$, where E is the incidence matrix (i.e., η are the edges of the graph spanned by the bonds). Furthermore, let $E^T\theta = p$. Then

$$\nabla_x = E^T\nabla_\eta, \tag{46}$$

$$E\nabla_p = \nabla_\theta. \tag{47}$$

Inserting these into Eq. 45, gives

$$i\mathscr{L} = \sum_i -\tilde{k}_I(\|\eta_I\| - r_I^{(0)})\frac{\eta_I}{\|\eta_I\|}\nabla_{\theta_I} + \theta^T E M^{-1} E^T \nabla_\eta. \tag{48}$$

Assuming a simple ansatz that

$$-\tilde{k}_I(\|\eta_I\| - r_I^{(0)})\frac{\eta_I}{\|\eta_I\|}\nabla_{\theta_I} f_I + \theta_I^T E_{Ij} M_{jj}^{-1} E_{Ij} \nabla_{\eta_I} f_I = \lambda_I f_I, \tag{49}$$

then

$$i\mathscr{L} g(\eta, \theta) = \theta^T (EM^{-1}E^T - D)\nabla_\eta g + \Lambda g, \tag{50}$$

where $g = \sum c_I f_I(\eta_I, \theta_I)$ and $\Lambda g = \sum c_I \lambda_I f_I(\eta_I, \theta_I)$.

The Liouville operator $\mathscr{L}$ is rotationally invariant (i.e., $\mathscr{L}$ commutes with any rotation $R(\phi)$). Hence, there are simultaneous eigenfunctions of $\mathscr{L}$ and $\eta^T\hat{R}\nabla_\eta + \theta^T\hat{R}\nabla_\theta$, where $\hat{R} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Eigenfunctions are compositions of $(1 \pm i\hat{R})\eta$ and $(1 \pm \hat{R})\theta$ with eigenvalues $\pm i$.

The following equations are some helpful identities that were used in the previous

derivations:

$$\eta = \eta_r \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} v \tag{51}$$

$$\theta = \theta_r \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} v \tag{52}$$

$$\nabla_v = \eta_r R(\phi)^T \nabla_\eta + \theta_r R(\phi) \nabla_\theta \tag{53}$$

$$\frac{d}{d\eta_r} = v^T R(\phi)^T \nabla_\eta \tag{54}$$

$$\frac{d}{d\theta_r} = v^T R(\phi) \nabla_\theta \tag{55}$$

$$\frac{d}{d\phi} = v^T \frac{dR}{d\phi}(\phi)^T \nabla_\eta + v^T \frac{dR}{d\phi}(\phi) \nabla_\theta \tag{56}$$

$$-k(\eta_r - r_0) v^T R^T(\phi) \nabla_\theta + \theta_r v^T R(\phi) M \nabla_\eta \tag{57}$$

6. Hermite Functions for Approximating Potentials

In the decomposition of the Liouville operator in Sections 3 and 4, a suitable basis of potential functions is needed. This basis is preferably sufficiently localized to avoid the need for global integrals. The Hermite functions,

$$\begin{aligned} \psi_n =& (-1)^n \left(2^n n! \sqrt{\pi}\right)^{-\frac{1}{2}} e^{\frac{1}{2}x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-x^2} \\ =& (-1)^n \left(2^n n! \sqrt{\pi}\right)^{-\frac{1}{2}} H_n(x) e^{-\frac{1}{2}x^2}, \end{aligned} \tag{58}$$

$$\begin{aligned} \psi_{2n} =& (-1)^{2n} \left(2^{2n} 2n! \sqrt{\pi}\right)^{-\frac{1}{2}} e^{\frac{1}{2}x^2} \frac{\mathrm{d}^{2n}}{\mathrm{d}x^{2n}} e^{-x^2} \\ =& (-1)^{2n} \left(2^{2n} 2n! \sqrt{\pi}\right)^{-\frac{1}{2}} H_{2n}(x) e^{-\frac{1}{2}x^2}, \end{aligned} \tag{59}$$

are an orthonormal complete set of basis functions that obey a 3-term recursion formula,

$$\psi_{n+1}(x) = \sqrt{\frac{2}{n+1}} x \psi_n(x) - \sqrt{\frac{n}{n+1}} \psi_{n-1}(x), \tag{60}$$

$$\psi_{2n+2} = \frac{(2x^2 - 4n - 1)}{\sqrt{(2n+2)(2n+1)}} \psi_{2n} - \sqrt{\frac{2n(2n-1)}{(2n+1)(2n+2)}} \psi_{2n-2}. \tag{61}$$

The recursion formula enables a fast transform of functions. Since "harmonic" po-

tentials are merely functions that capture 3 aspects of bonding, it is possible to use Hermite functions in their stead. The boundary conditions are

$$f(0) = r_0^2 - D, \tag{62}$$

$$\frac{\mathrm{d}f}{\mathrm{d}x}(r_0) = 0, \tag{63}$$

$$f(r_0) = -D, \tag{64}$$

$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = k \tag{65}$$

$$f(x) = f(-x), \tag{66}$$

where D is the dissociation energy, k is the force constant, and r_0 is the equilibrium distance. These conditions result in a linear system for coefficients in an expansion of Hermite functions,

$$f = \left(\alpha + \beta\left(\frac{x}{\sigma}\right)^2 + \gamma\left(\frac{x}{\sigma}\right)^4\right) e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}. \tag{67}$$

A simple decomposition converts this equation into the coefficients of the Hermite functions of scale σ! The Liouville operator for such a potential would be

$$\begin{aligned} i\mathscr{L} =& -\nabla_i^T f(\|x_i - x_j\|)\frac{\mathrm{d}}{\mathrm{d}p_i} - \nabla_j^T f(\|x_i - x_j\|)\frac{\mathrm{d}}{\mathrm{d}p_j} + m_i^{-1} p_i^T \frac{\mathrm{d}}{\mathrm{d}x_i} + m_j^{-1} p_j^T \frac{\mathrm{d}}{\mathrm{d}x_j} \\ =& \frac{x_i - x_j}{\sigma}\left(f + \left(2\beta + 4\gamma\left\|\frac{x_i - x_j}{\sigma}\right\|^2\right) e^{-\frac{1}{2}\left\|\frac{x_i - x_j}{\sigma}\right\|^2}\right)\left(\frac{\mathrm{d}}{\mathrm{d}p_i} - \frac{\mathrm{d}}{\mathrm{d}p_j}\right) + \\ & m_i^{-1} p_i^T \frac{\mathrm{d}}{\mathrm{d}x_i} + m_j^{-1} p_j^T \frac{\mathrm{d}}{\mathrm{d}x_j}. \end{aligned} \tag{68}$$

A further point of note is the fact that this expansion only includes quadratic terms with respect to the DoFs, which allows much simpler decomposition into component terms after linear transformation of the DoFs into CAMs. Finally, geometries may be defined by distances obviating the need for angular and dihedral potentials.

7. Koopman for MD

In the following, the relationship between operator wavelets and the Koopman modes of the Liouville operator is demonstrated. The Liouville operator $\mathscr{L}$ is a Koopman operator for MD systems. Let $\hat{x}(x,p) = \sum_{\{\lambda\}} x_\lambda K_\lambda(x,p)$, where K_λ is an eigenfunction of $\mathscr{L}$ and $\hat{x} : (x,p) \mapsto x$. Then,

$$x(t) = e^{-i\mathscr{L}t}\hat{x}(x_0,p_0) = \sum e^{i\lambda_j t} x_{\lambda_j} K_{\lambda_j}(x_0,p_0), \tag{69}$$

where λ_j is the corresponding eigenvalue to K_{λ_j}. A decomposition of $\mathscr{L}$ using an operator hierarchy as in Section 3 naturally separates the scales by the wavelet operators. Therefore, K_{λ_j} is in the kernel of all nontrivial operator wavelets save one. The eigenvalues λ_j are generally degenerate. Grouping eigenfunctions $K_{\lambda,k}$ with the same eigenvalue

$$x(t) = \sum e^{i\lambda_j} \sum x_{j,k} K_{j,k}(x_0,p_0) = \sum A_j e^{i\lambda_j}, \tag{70}$$

reduces the number of actually free variables, where A_j are the effective free variables. Assume 2 observables evolve in time

$$x(t) = e^{-i\mathscr{L}t}\hat{x}(x_0,p_0) = \sum e^{i\lambda_j t} x_{\lambda_j} K_{\lambda_j}(x_0,p_0), \tag{71}$$

$$y(t) = e^{-i\mathscr{L}t}\hat{y}(y_0,q_0) = \sum e^{i\lambda_j t} y_{\lambda_j} K_{\lambda_j}(y_0,q_0). \tag{72}$$

Then, cross-correlation indicates

$$\langle x|y\rangle = \sum x^*_{\lambda_j} y_{\lambda_j} K_{\lambda_j}(x_0,p_0) K_{\lambda_j}(y_0,q_0). \tag{73}$$

It follows that the eigenvectors of the correlation matrix of a set of observables is the optimal linear representation of same observables. This includes the original DoFs as well as CAMs.

We can further characterize the eigenfunctions K_λ of $\mathscr{L}$ for physical systems by its distributive property,

$$\mathscr{L}(f \cdot g) = g \cdot \mathscr{L} f + f \cdot \mathscr{L} g. \tag{74}$$

Hence, any power of an eigenfunction of $\mathscr{L}$ is itself an eigenfunction of $\mathscr{L}$ via

$$\mathscr{L}K_\lambda = \lambda K_\lambda \Rightarrow \mathscr{L}K_\lambda^n = \sum_{m=0}^{n-1} K_\lambda^m (\mathscr{L}K_\lambda) K_\lambda^{n-m-1} = n\lambda K_\lambda^n. \tag{75}$$

This motivates a simple shift operator K_λ^{-1} and a doubling operator $D : f(K_\lambda) \mapsto f(K_\lambda^2)$, where f is a formal series in polynomial powers of K_λ.

The effect of the Koopman operator can be captured by bootstrapping. With an initial trajectory, the effective frequencies can be found by solving

$$\min_{\{\tilde{x}_j\},\{w_j\}} \sum_{n=0}^{N} \left\| x(n\Delta t) - \sum_j^M e^{iw_j n\Delta t}\tilde{x}_j \right\|^2 + C\sum \|\tilde{x}_j\|, \tag{76}$$

where N is the number of time steps in the initial run, Δt is the time step, M is the number of frequencies to capture, and C is a cost factor to control overfitting with regularizer $\sum\|\tilde{x}_j\|$. By comparison with Eq. 71, $\tilde{x}_j \approx x_{w_j}$. Then $\Delta x := x(t) - \sum_j^M e^{iw_j t}\tilde{x}_j$ is the residual due to $\Delta\mathscr{L} := \mathscr{L} - \tilde{\mathscr{L}}$, where $\tilde{\mathscr{L}}$ is the linear operator with eigenfunctions and eigenvalues (K_{w_j}, w_j) such that $f \mapsto 0 \forall f \notin \text{span}\,\{K_{w_j}\}$. It follows that $\tilde{\mathscr{L}}$ and $\Delta\mathscr{L}$ are orthogonal operators and commute. Thus

$$e^{i\mathscr{L}t} = e^{i\Delta\mathscr{L}t + i\tilde{\mathscr{L}}t} = e^{i\Delta\mathscr{L}t}e^{i\tilde{\mathscr{L}}t} = e^{i\tilde{\mathscr{L}}t}e^{i\Delta\mathscr{L}t}. \tag{77}$$

At time t, the action of $\Delta\mathscr{L}$ on x is $\Delta p := p(t) - \sum iw_j\tilde{x}_j e^{iw_j t}$, whereas the action on Δp is $-\nabla V(x(t)) + \sum w_j^2\tilde{x}_j e^{iw_j t}$.

Using the usual Velocity-Verlet factorization going from time $t = 0$ to $t = \Delta t$,

$$\Delta x_1 = \Delta p_0 \Delta t/2 + \Delta x_0, \tag{78}$$

$$\Delta p_2 = \Delta p_0 - \left(\nabla V(\Delta x_1 + \tilde{x}(\Delta t/2)) - \dot{\tilde{p}}(\Delta t/2)\right) t, \tag{79}$$

$$\Delta x_3 = \Delta x_1 + \Delta p_2 t/2 = \Delta x_0 + (\Delta p_0 + \Delta p_2)\, t/2, \tag{80}$$

$$x(\Delta t) = \Delta x_3 + \tilde{x}(\Delta t), \tag{81}$$

$$p(\Delta t) = \Delta p_2 + \tilde{p}(\Delta t). \tag{82}$$

This approach is equivalent to introducing new variables α_j and $\beta_j := \dot{\alpha}_j$ and

changing the Liouville operator to

$$i\mathscr{L} = tr\left[(\nabla_x^T V)\left(x + \sum \alpha_j\right)\nabla_p\right] + tr\, p^T M^{-1}\nabla_x + \sum \beta_j^T \nabla_{\alpha_j} - \omega_j^2 \alpha_i^T \nabla_{\beta_j}. \tag{83}$$

8. Derivation of Wavelet Spaces

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

$$\left(\prod_{i=0}^{n} Q_i\right)^{\mathsf{T}} T^{2^{n+1}} \left(\prod_{i=0}^{n} Q_i\right) = Q_{n+1} R_{n+1}. \tag{84}$$

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,

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

9. 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

$$S_T^{(n)} := \frac{\varepsilon - \omega_{n,max}}{\omega_{n,min} - \varepsilon}, \tag{85}$$

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

can 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.

10. Invariant Subspaces

$\tilde{\Delta}$ is positive semi-definite 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 $j < k \forall j \in J, k \in K$. The rank of $\tilde{\Delta}_{JK}$ is generally low due to the low maximum degree of a vertex in $\tilde{\Delta}$. If $\Lambda = \operatorname{span} \tilde{\Delta}_{KJ}$ 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 Section 2.1.4 and Section 12, 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.

11. 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

$$\tilde{\Delta}' = (M^{-\frac{1}{2}} + \delta e_i e_i^T)\Delta(M^{-\frac{1}{2}} + \delta e_i e_i^T) \tag{86}$$

$$= \tilde{\Delta} + \delta m_i^{\frac{1}{2}}(e_i e_i^T \tilde{\Delta} + \tilde{\Delta} e_i e_i^T) + \delta^2 m_i \tilde{\Delta}_{ii} e_i e_i^T. \tag{87}$$

For an eigenvector ν of $\tilde{\Delta}$ contained in some wavelet space W_n with nondegenerate eigenvalue η_ν, the first-order correction to the eigenvalue is $m_i^{1/2}\delta|\nu_i|^2(2\eta_\nu + m_i^{1/2}\tilde{\Delta}_{ii}\delta)$, and the first-order correction to the eigenvector is $\sum_{\mu\neq\nu} \mu \cdot \delta\mu_i\nu_i m_i^{1/2}(\eta_\nu + \eta_\mu + \delta m_i\tilde{\Delta}_{ii})/(2\eta_\mu - 2\eta_\nu)$, where $\{\mu\}$ is an orthonormalized set of eigenvectors of $\tilde{\Delta}$. Hence the scale of ν is unaffected if $|\nu_i|^2$ is 0 or sufficiently negligible. Corrections to the eigenvector are only of significance if some eigenvector μ outside of W_n additionally has a significant component μ_i.

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

$$\tilde{\Delta}' = \tilde{\Delta} + \delta M^{-\frac{1}{2}}(e_i e_i^T + e_j e_j^T - e_i e_j^T - e_j e_i^T)M^{-\frac{1}{2}}, \tag{88}$$

$$\eta_\nu^{(1)} = \delta |m_i^{-\frac{1}{2}}\nu_i - m_j^{-\frac{1}{2}}\nu_j|^2, \tag{89}$$

$$\nu^{(1)} = \sum_{\mu\neq\nu} \delta \frac{(m_i^{-\frac{1}{2}}\mu_i - m_j^{-\frac{1}{2}}\mu_j)(m_i^{-\frac{1}{2}}\nu_i - m_j^{-\frac{1}{2}}\nu_j)}{\eta_\mu - \eta_\nu}\mu. \tag{90}$$

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

12. 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

$$\begin{pmatrix} A_{1,1}I_n & A_{1,2}I_n & \cdots & A_{1,m}I_n + B_{1,m}(\Sigma_n + e_1 e_n^T) \\ A_{2,1}I_n & A_{2,2}I_n & \cdots & A_{1,m}I_n \\ \vdots & & \ddots & \vdots \\ A_{m,1}I_n + B_{1,m}(\Sigma_n^T + e_n e_1^T) & \cdots & \cdots & A_{m,m}I_n \end{pmatrix}. \tag{91}$$

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

$$\begin{pmatrix} A_{1,1}I_n & A_{1,2}I_n & \cdots & D^* \\ A_{2,1}I_n & A_{2,2}I_n & \cdots & A_{1,m}I_n \\ \vdots & \vdots & \ddots & \vdots \\ D & A_{m,2}I_n & \cdots & A_{m,m}I_n \end{pmatrix} = \mathbb{U}^*\kappa^T\tilde{\Delta}\kappa\mathbb{U}, \tag{92}$$

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 $\tilde{A}_i = A - (A_{1,m} + D_{i,i})e_1e_m^T - (A_{1,m} + D_{i,i}^*)e_me_1^T$. 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 ν of A with $e_m^T\nu = e_1^T\nu = 0$ has an n-fold degenerate eigenvalue.

13. Derivation of Mixed Resolution from Separable Coarsenings

Let $\beta_n^\tau : \mathbb{R}^{m_n} \to \beta_n$ denote a parameterization 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 parameterizations $\beta_n^\tau : A_n \oplus B_n \to \beta_n$ 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

$$\beta_n^\tau(\mu_{A_n}^{-1}(t_A), \mu_{B_n}^{-1}(t_B)) = \gamma_n^{-1}(\beta_{n+1}^\tau(t_A, t_B)) \tag{93}$$

for $t_A \in A_{n+1}, t_B \in B_{n+1}$. μ_A and μ_B map a fine parametrization to a coarse parameterization. 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 Figure 11.

A separable coarsening for which neither μ_{A_n} nor μ_{B_n} are bijections 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.

14. Conclusions and Outlook

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 novel classes of molecules without requiring extensive expert insight and model parameterization. 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 parameterization) 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[31–33].

Future work will establish the relationship between time steps and simulation accuracy[34]. Currently, the bootstrapping procedure for optimal Koopman mode approximation requires embedding in the propagation scheme and linking to stochastic dynamics. The operator wavelet approach has implications for developing new operator decompositions, which require derivation and solution of analytical subsystems. Meanwhile, a proper effective operator decomposition using Hermite functions or conventional harmonic, angle, and dihedral functions as a basis for fitting can be achieved.

15. Publications and Presentations

The following are previously published presentations and publications that document work in this effort:

1. Rinderspacher BC, Bardhan J, Ismail AE. Diffusion wavelet-based decompositions for coarse-graining of polymer chains. Paper presented at: American Institute of Chemical Engineers. 2015 Nov10; Salt Lake City, UT.

2. Rinderspacher BC, Bardhan JP, Ismail AE. Diffusion wavelet decomposition for coarse-graining of polymer chains. MRS Proceedings. 2015;1753.

3. Rinderspacher BC, Bardhan J, Ismail AE. Multiresolution of molecular dynamics in the particle domain. Paper presented at: SIAM Mathematical Aspects of Materials Science Conference. 2016 May 8; Philadelphia, PA.

4. Rinderspacher BC, Bardhan JP, Ismail AE. Diffusion wavelet decomposition for coarse-graining of polymer chains. Paper presented at: Materials Research Society Fall Meeting. 2014 Dec 1; Boston, MA.

5. Rinderspacher BC, Bardhan J, Ismail AE. Diffusion wavelet-based decompositions for coarse-graining of polymer chains. Paper presented at: University of Delaware, Applied Mathematics. 2015 Oct 5; Newark, DE.

6. Rinderspacher BC, Bardhan J, Ismail AE. Wavelet analysis for molecular dynamics. Paper presented at: MEDE. 2014 Oct 16; Towson, MD.

7. Rinderspacher BC, Bardhan J, Ismail AE. Diffusion wavelet decomposition for coarse-graining of polymer chains. Paper presented at: American Chemical Society National Meeting. 2015 Aug 16; Boston, MA.

8. Rinderspacher BC, Bardhan J, Ismail AE. Diffusion wavelet decomposition for coarse-graining of polymer chains. Paper presented at: Hopkins Extreme Materials Institute. 2015 Jul 30; Baltimore, MD.

9. Rinderspacher BC, Bardhan JP, Ismail AE. Theory of wavelet-based coarse-graining hierarchies for molecular dynamics. Physical Review E. ; submitted.

16. Bibliography

1. Akkermans RLC, Briels WJ. Coarse-grained interactions in polymer melts: a variational approach. J Chem Phys. 2001;115:6210-6219.

2. Baschnagel J, Binder K, Doruker P, Gusev AA, Hahn O, Kremer K, Mattice WL, Müller-Plathe F, Murat M, Paul W, Santos S, Suter UW, Tries V. Bridging the gap between atomistic and coarse-grained models of polymers: Status and Perspectives. 2000;152.

3. Lyubartsev AP, Karttunen M, Vattulainen I, Laaksonen A. On coarse-graining by the inverse Monte Carlo method: dissipative particle dynamics simulations made to a precise tool in soft matter modeling. Soft Materials. 2003;1:121-137.

4. Tan RKZ, Petrov AS, Harvey SC. YUP: a molecular simulation program for coarse-grained and multiscaled models. J Chem Theory Comput. 2006;2:529-540.

5. Clementi C. Coarse-grained models of protein folding: toy models or predictive tools? Current Opinion in Structural Biology. 2008;18:10-15.

6. Depa PK, Maranas JK. Speed up of dynamic observables in coarse-grained molecular-dynamics simulations of unentangled polymers. J Phys Chem. 2005;123(9):094901.

7. Coifman R, Maggioni M. Diffusion wavelets. Appl Comput Harm Anal. 2006;21(1):53–94.

8. Bremer J, Coifman R, Maggioni M, Szlam A. Diffusion wavelet packets. Appl Comput Harm Anal. 2006;21(1):95–112.

9. Abrams C, Delle Site L, Kremer K. Dual-resolution coarse-grained simulation of the bisphenol-A-polycarbonate/nickel interface. Phys Rev E. 2003;67:021807.

10. Abrams C, Kremer K. Combined coarse-grained and atomistic simulation of liquid bisphenol A-polycarbonate: Liquid packing and intramolecular structure. Macromolecules. 2003;36:260.

11. Marrink SJ, de Vries AH, Mark AE. Coarse grained model for semiquantitative lipid simulations. J Phys Chem B. 2004;108(2):750-760.

12. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The martini force field:âĂL̓ coarse grained model for biomolecular simulations. J Phys Chem B. 2007;111(27):7812-7824.

13. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI coarse-grained force field: extension to proteins. J Chem Theory Comput. 2008;4:819-834.

14. Gautieri A, Russo A, Vesentini S, Redaelli A, Buehler MJ. Coarse-grained model of collagen molecules using an extended MARTINI force field. J Chem Theory Comput. 2010;6:1210-1218.

15. Fan B, Maranas J. Coarse-grained simulation of cellulose iÎš with application to long fibrils. Cellulose. 2015;22(1):31-44.

16. Lu L, Izvekov S, Das A, Andersen HC, Voth GA. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. J Chem Theory Comput. 2010;6(3):954–965.

17. Harmandaris VA, Reith D, Van der Vegt NFA, Kremer K. Comparison between coarse-graining models for polymer systems: Two mapping schemes for polystyrene. Macromolecular Chem Phys. 2007;208(2109).

18. Espanol P. Dissipative particle dynamics with energy conservation. Europhys. Lett. 1997;40(6):631-636.

19. Binder K, editor. Monte Carlo and molecular dynamics simulations in polymer science. Oxford (UK): Oxford University Press; 1995.

20. Daubechies I. Ten lectures on wavelets. Philadelphia: SIAM; 1992. Proceedings of the CBMS-NSF Regional Conference Series in Applied Mathematics; vol. 61.

21. Strang G, Nguyen T. Wavelets and filter banks. Cambridge (MA): Wellesley-Cambridge Press; 1996.

22. Heyden A, Truhlar DG. Conservative algorithm for an adaptive change of resolution in mixed atomistic/coarse-grained multiscale simulations. J Chem Theory Comput. 2008;4:217-221.

23. Praprotnik M, Delle Site L, Kremer K. Adaptive resolution molecular-dynamics simulation: changing the degrees of freedom on the fly. J Chem Phys. 2005;123(224106).

24. Peter C, Delle Site L, Kremer K. Classical simulations from the atomistic to the mesoscale and back: coarse graining an azobenzene liquid crystal. Soft Matter. 2008;4:859-869.

25. Ismail AE, Rutledge GC, Stephanopoulos G. Topological coarse graining of polymer chains using wavelet-accelerated Monte Carlo. I. freely jointed chains. J Chem Phys. 2005;122:234901.

26. Ismail AE, Stephanopoulos G, Rutledge GC. Topological coarse graining of polymer chains using wavelet-accelerated Monte Carlo. II. Self-avoiding chains. J Chem Phys. 2005;122:234902.

27. Chung FRK. Spectral graph theory. American Mathematical Society; 1997.

28. Sun H, Mumby SJ, Maple JR, Hagler AT. An *ab initio* cff93 all-atom force field for polycarbonates. J Am Chem Soc. 1994;116(7):2978–2987.

29. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. J Am Chem Soc. 1995;117(1):1–19; .

30. Vazquez E, Bect J. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. Journal of Statistical Planning and Inference. 2010;140(11):3088 - 3095; .

31. Katsoulakis MA, Plecháč P, Rey-Bellet L, Tsagkarogiannis DK. Mathematical strategies in the coarse-graining of extensive systems: error quantification and adaptivity. J Non-Newtonian Fluid Mech. 2008;152:101-112.

32. Shell MS. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J Chem Phys. 2008;129(14).

33. Katsoulakis MA, Plecháč P. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. J Chem Phys. 2013;139(7).

34. Mansour AA, Ortoleva PJ. Implicit time integration for multiscale molecular dynamics using transcendental Padé approximants. J Chem Theory Comput. 2016;(in press).

INTENTIONALLY LEFT BLANK.

List of Symbols, Abbreviations, and Acronyms

CAM collective action mode

CG coarse-grained

DoF degree of freedom

GAFF generalized amber force field

LAMMPS Large-scale Atomistic/Molecular Massively Parallel Simulator

MD molecular dynamics

PCFF polymer-consistent force field

PE polyethylene

INTENTIONALLY LEFT BLANK.

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
4 (PDF)	DIR USARL RDRL WML B B RICE RDRL WMM G J ANDZELM T SIRK J LENHART

INTENTIONALLY LEFT BLANK.