Role of many-body interactions in the structure of coarse-grained polymers

K. Michael Salerno and Noam Bernstein

Phys. Rev. E **98**, 023310 — Published 27 August 2018

DOI: 10.1103/PhysRevE.98.023310

The role of many-body interactions in the structure of coarse-grained polymers

K. Michael Salerno[1] and Noam Bernstein[2]

[1] *NRC Research Associate, Resident at Center for Computational Materials Science, Naval Research Laboratory, Washington, DC 20375*

[2] *US Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20375**

(Dated: August 7, 2018)

In developing coarse-grained (CG) polymer models it is important to reproduce both local and molecule-scale structure. We develop a procedure for fast calculation of the bond-orientation correlation and the internal squared distance $\langle R^2(M)\rangle$ through evaluation of the probability distribution functions that represent a CG model. Different CG models inherently contain or omit correlations between CG variables. Here, we construct CG models that contain specific correlations between CG variables. The importance of different correlations is tested on CG models of polyethylene, polytetrafluoroethylene, and poly-L-lactic acid. The chain stiffness and $\langle R^2(M)\rangle$ are calculated using both analytic evaluation and Monte-Carlo sampling, and approximate model results are compared with exact results from all-atom simulations. For polymers with an exponential correlation decay, the bond-orientation correlation and $\langle R^2(M)\rangle$ indicate which CG variable correlations are most important to reproduce molecule-scale structure. Analysis of the bond-orientation correlation and internal-squared distance indicates that for poly-L-lactic acid the bond-orientation correlation requires qualitatively different additional terms in CG models, and quantifies the error in neglecting this important behavior.

I. INTRODUCTION

Modeling polymers is an inherently multi-scale problem, with important properties arising from both short and long length scales. Local structure, for example the local bond and angle stiffness, is a function of specific polymer chemistry, and small changes in structure can have important effects on the transferability of polymer models and important polymer properties, particularly near phase boundaries [1–4]. At larger lengths, the stiffness of a polymer is one of the most important properties of polymers in a melt or solution. Molecule-scale structural properties like the internal squared distance $\langle R^2(M)\rangle$, radius of gyration, and entanglement length all depend on the polymer stiffness. It also governs mechanical response of systems as diverse as polymer melts, polymer glasses, and biopolymers [5–9]. Given the fundamental importance of polymer stiffness it is critical that coarse-grained (CG) polymer models reproduce this intrinsic property, otherwise the substantial (up to $\times 10^4$) speedup realized by CG polymer models [10–12] is not useful.

Chemically specific CG models are often derived from underlying all-atom (AA) simulations, using various methods to determine effective interactions between coarse-graining sites. Often, these chemically specific CG models use either tabulated interaction potentials or simple analytic potentials to approximately reproduce the internal coordinates of the coarse-graining interaction sites, i.e. bond lengths, angles and dihedrals within the CG polymer chain. For tabulated potentials the local structure of bond lengths, angles and dihedrals can be an exact (statistical-distribution) match. However, CG polymer models often do not match molecule-scale structural properties of AA simulations, even when independent CG variable distributions match exactly, and different ad-hoc remedies have been used, including additional bonds between 1-4 neighbors along the CG polymer or phenomenological changes to the bond-angle or dihedral-angle interaction [13–15]. Previous work has found that for some CG polymer models, certain variables and the correlations between them are important. For example studies of polyethylene (PE) [15, 16] have found dihedral-angles significant while studies of poly(vinyl alcohol) [1, 17] did not. Researchers have long noted how some CG variables, for example the bond angle and dihedral angle in Boltzmann inverted CG models of PE [18] or polystyrene [14], can be correlated. Here we examine correlations between the variables in CG polymer representations and measure the influence of these correlations on molecule-scale structural properties.

Since the central goal is to create CG polymer models that reproduce features and properties of an underlying AA simulation, it is important to make an explicit distinction between a CG *representation*, by which we mean the CG variables yielded from applying a CG mapping to simulated AA configurations, and an approximate CG *model*, by which we mean approximations to the full probability distribution function (PDF) of all the CG variables in a CG representation [15]. Approximate models can be cast as either a PDF of CG variables or equivalently as a set of CG interaction potentials. Quantities can be calculated using the CG model through analytic evaluation using CG variable PDFs, or through Monte-Carlo (MC) sampling or molecular dynamics (MD) simulation of the equivalent set of CG interaction potentials.

As an example, a center-of-mass mapping scheme applied to AA simulation data produces CG site positions,

* noam.bernstein@nrl.navy.mil

and this CG representation contains all the CG site internal coordinates, i.e. all $(3N-6)$ bond-lengths, bond-angles, and dihedral-angles, where N is the number of CG bead sites in a single chain. From this representation an approximate CG model can be created, for example, by assuming independent and identically distributed bond lengths and bond angles and uniformly distributed dihedral angles.

An ideal CG model meets two criteria, computational feasibility and accuracy. The first, feasibility, requires that the model be fast both to create and to simulate. A CG model is constructed with explicit PDFs of CG variables from the full CG representation. The number of required samples increases exponentially with the dimensionality of a PDF, so to be fast requires that the dimensionality of the sampled PDFs should be small. For fast CG model simulations the CG interaction potential, which is simply k_BT times the log of the multi-dimensional PDF, should have as short a range as possible while maintaining accuracy. For example, if there are interactions between all the beads in a chain then the cost will grow with the number of CG beads N as N^2, and simulating the model will be slow.

The second criterion, accuracy, requires that the CG model at least match both the independent CG-variable distributions and molecule-scale structure of the CG representation. For CG models constructed using measured distributions of the underlying CG variables the independent distribution of each CG variable is matched by construction. In order to match the molecule-scale structure we aim to match the bond-orientation correlation $C_{bb}(m)$ for all bond-separations m. The full bond-orientation correlation decay determines the internal squared distance $\langle R^2(M)\rangle$, and, given an exponential correlation decay, the definition of a persistence length.

Analytic evaluation of $C_{bb}(m)$ for small m is orders of magnitude faster than MC sampling or MD simulation, and we exploit this speed by evaluating $C_{bb}(m)$ for $m = 1, 2, 3$ and then extrapolating an exponential decay curve to calculate $\langle R^2(M)\rangle$. As will be discussed in detail below, evaluating $C_{bb}(3)$ requires constructing a $(3 \times 5 - 6) = 9$ dimensional PDF from the CG variables and correlations that define the CG model.

An alternative is to use MC to sample from the potential equivalent to the $(3N-6)$ dimensional PDF for an N-bead chain. Several benefits of the method are that N is restricted only by computational limits, $C_{bb}(m)$ can be computed for any $m < N-1$, and $\langle R^2(M)\rangle$ can be computed for arbitrary separation M. We use these features of the MC sampling to check the analytic model calculations and to check the large m values and extrapolation of $C_{bb}(m)$.

We expect that there may be some polymers for which approximate CG models with local interactions will successfully match the local and molecule-scale structure, and others for which the models will not. For example, polymers with significant charge could have a more complicated form for the bond-orientation correlation decay that cannot be reproduced by local interactions. We construct seven different approximate models that include different correlations between CG variables and test which models reproduce molecule-scale structure for three different polymers, PE, polytetrafluoroethylene (PTFE) and poly-L-lactic acid (PLLA). For CG models of PE and PTFE analytic evaluation of the bond-orientation correlation and extrapolation to calculate $\langle R^2(M)\rangle$ gives a clear, effective metric for the CG models, and indicates which CG variable correlations are most important for reproducing long-range structure. For approximate models of PLLA, analytic evaluation of the bond-orientation correlation and $\langle R^2(M)\rangle$ indicates that our simple approximate models neglect important long-range interactions, which we confirm by results from MC simulations.

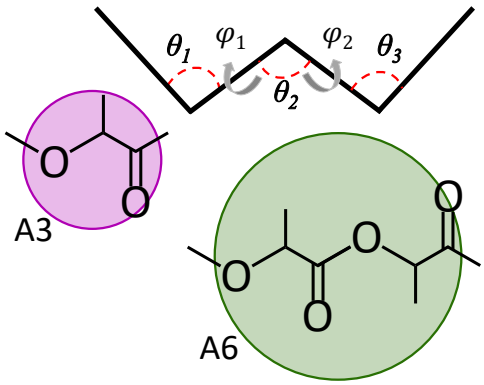

FIG. 1. (top) A diagram of the CG angle (θ) and dihedral (φ) variables studied in this work. (bottom) A diagram of the atoms included in the PLLA-A3 and A6 CG representation.

II. METHODS

A. All-atom simulation

All-atom (AA) simulations were performed to generate reference PE, PTFE and PLLA AA position data. Details of the PE and PTFE AA simulations have been presented previously [12, 15]. Like the PE and PTFE simulations, the reference AA PLLA simulations used parameters from the OPLS-AA force field. The AA PLLA simulation contained 1152 chains of 48 repeat units. The simulation temperature was 550 K, significantly above the glass transition and near the maximum of experimental temperatures [19]. The density was chosen to be 1.03 g/cm^3 based on a 1 ns NPT simulation with a target pressure of 30 atm and barostat damping constant $\tau_P = 0.1$ ps. A production simulation was run in the NVT ensemble for over 80 ns, with atom configurations sampled every 100 ps after 35 ns. An integration time step of 1 fs was used and a Langevin thermostat with time constant $\tau_T = 1$ ps was applied. A real-space cutoff of

1.0 nm was used. The precision for the reciprocal-space electrostatic interactions was 10^{-4}. All simulations were performed using LAMMPS [20]. Simulation times are of order 50-100 ns, representing greater than 500 times the longest decorrelation time for CG variables of 100-150 ps. Information about these CG variable decorrelation times is given in the Appendix.

B. Coarse-Grained Mapping/Reps

CG representations were created from AA simulations of PE, PTFE and PLLA. For PE and PTFE, the CG representations were constructed from AA simulations at 500K (PE) and 650K (PTFE) discussed in previous work [12, 15, 21]. While each CG representation is only exactly valid at single state point there has been work to explore CG model transferability [13, 22], a subject that is outside the scope of this work. The techniques and analysis that follow can be applied to a CG model developed at any temperature or pressure, provided that good sampling of the all-atom configurations may be obtained at that temperature. The CG representations used here were constructed by grouping λ CX_2 groups per CG bead, using either an averaging (Aλ) or decimation (Dλ) technique to determine the CG bead locations. In the averaging method the CG bead position is based on an average of atom positions, while the decimation technique places a CG bead at the location of every λth backbone atom. For PLLA, four representations were used. Two represent each $C_3O_2H_4$ repeat unit as a CG bead, while two representations group two repeat units in a single CG bead. Figure 1 illustrates the two averaging CG PLLA representations used in this paper. These representations are labeled PLLA-Xλ, where X represents averaging (A) or decimation (D) and the λ is the number of backbone atoms per bead.

C. Models

Evaluating observables using CG models requires constructing high-order PDFs from the product of low-order PDFs, an approach that has a long history in the physics literature [23–28]. We borrow from previous work both in some aspects of notation, as well as ideas [24, 27, 28]. Particularly relevant to the current context is work modeling small molecule conformations and estimating configurational entropy with reduced-dimensional models [27, 28].

Here we sample variables from CG representations of AA simulations to form low-order PDFs that include correlations between nearby CG variables. For example, the adjacent angles θ_1 and φ_1 shown in Fig. 1 are likely to contain non-trivial correlations. In this work we assume that for an N-bead chain the $(N-1)$ CG bond-length variables are independent of each other and of the other CG angle variables, leaving a $(2N-5)$-dimensional distribution of bond-angle and dihedral-angle variables. Including correlations between CG angles and bond-lengths does not influence the "stiffness" of the CG polymer model, which depends only on the bond-bond correlation, but it can change $\langle R^2 \rangle$. For the models studied here this effect is smaller than the error introduced in the angle approximations except in the decimation models with small λ (D2), which are of comparable (1-2%) size. (double check) Hence, we focus on the bond-angle and dihedral-angle variables and refer to them throughout the paper as the angle and dihedral respectively.

The number of histogram bins and the samples needed to build a PDF increases exponentially with the PDF dimensionality. Thus the computational cost of creating CG variable samples with AA MD simulations limits the dimensionality of the PDFs we can use. Based on CG variable decorrelation times discussed in the Appendix we are limited to one-, two-, and three-dimensional distributions of CG variables. In fact, sampling CG variables is only feasible because, excluding chain ends, the CG angle and dihedral variables are assumed to be independent of location along a chain and PDF histograms can be generated by averaging together CG variables at different positions along every chain.

The correlations between CG variables that are included in each different model are embodied in the "Minimal Set of PDFs" listed in Table I. In Table I the distributions use the shorthand notation $P(\beta_j, \beta_{j+\delta_1}, \beta_{j+\delta_2}, ...) \equiv [\beta_j, \beta_{j+\delta_1}, \beta_{j+\delta_2}, ...]$ adapted from recent work on reduced-dimensional models [27, 28]. In this notation the CG variable β represents either the CG angle (θ) or dihedral (φ) variable, and the PDF is defined by the sequence of CG variables and their relative indices δ_1, δ_2, etc., rather than the absolute index or location j along the chain. We also include the Jacobian factor $\sin\theta$ within the bracket definition $[-]$ when β represents the bond-angle θ. For example, the $\Theta\Phi\Theta$ model contains the PDF $P(\theta_j, \varphi_j, \theta_{j+1}) \sin\theta_j \sin\theta_{j+1} = [\theta_j, \varphi_j, \theta_{j+1}]$, which is sampled over all chains and all values of j in the CG representation of the AA simulation.

As we combine the low-order PDFs to represent a model polymer chain, we indicate that these CG angle and dihedral variables represent a model chain by using numeric indices to specify an absolute location along the chain. These variables are distinct random variables drawn from a distribution that is defined by the relationship between the variable subscripts or equivalently, their *relative* configuration along the chain.

The high-dimensional PDFs needed to analytically evaluate the bond-orientation correlation terms up to $m = 3$ require modeling $N = 5$ beads or constructing a $(2N-5) = 5$-dimensional PDF for all the CG angle and dihedral variables. For a MC simulation the value of N is not fixed, but simply represents the chain length.

Generically, if a single bead is added to a polymer chain the dimensionality of the distribution increases by two, since an additional φ and θ variable are each added. The approximate distribution for an $(N+1)$-bead chain can

TABLE I. The approximate CG models used in this work, the Minimal Set of PDFs that embody the correlations between CG variables within each CG model, and the Minimal-Chain Distribution that can be used build a polymer chain of arbitrary length using the approximate CG model.

Model	Minimal Set of PDFs	Minimal-Chain Distribution (MCD)
Θ	$[\theta]$	$[\theta]$
Θ-Φ	$[\theta], [\varphi]$	$[\theta_1][\varphi_1][\theta_2]$
ΘΦ-ΦΘ	$[\theta_1, \varphi_1], [\varphi_1, \theta_2]$	$\frac{[\theta_1, \varphi_1][\varphi_1, \theta_2]}{[\varphi_1]}$
ΘΦΘ	$[\theta_1, \varphi_1, \theta_2]$	$[\theta_1, \varphi_1, \theta_2]$
ΘΦ-ΦΘ-ΦΦ	$[\theta_1, \varphi_1], [\varphi_1, \theta_2], [\varphi_1, \varphi_2]$	$\frac{[\theta_1, \varphi_1][\varphi_1, \theta_2][\varphi_1, \varphi_2][\theta_2, \varphi_2][\varphi_2, \theta_3]}{[\varphi_1]^2[\theta_2][\varphi_2]^2}$
ΘΦΘ-ΦΘΦ	$[\theta_1, \varphi_1, \theta_2], [\varphi_1, \theta_2, \varphi_2]$	$\frac{[\theta_1, \varphi_1, \theta_2][\varphi_1, \theta_2, \varphi_2][\theta_2, \varphi_2, \theta_3]}{[\varphi_1, \theta_2][\theta_2, \varphi_2]}$
ΘΦΘ-ΘΦΦ	$[\theta_1, \varphi_1, \theta_2], [\theta_1, \varphi_1, \varphi_2]$	$\frac{[\theta_1, \varphi_1, \theta_2][\theta_1, \varphi_1, \varphi_2][\theta_2, \varphi_2, \theta_3]}{[\theta_1, \varphi_1][\theta_2][\varphi_2]}$

be constructed by extending the distribution for an N-bead chain. For this procedure it is useful to define a Minimal-Chain Distribution (MCD) of dimension n, the lowest dimensionality that includes all the PDFs in the Minimal Set of PDFs for an approximate model and both starts and ends with an angle θ [29]. This feature allows the MCD to be used to construct an arbitrarily high-dimensional PDF or equivalently an arbitrarily long polymer chain.

In constructing the MCD, Bayes' rule is used to eliminate redundant variables. For example, the MCD for the ΘΦΘ-ΦΘΦ model listed in Table I contains consecutive triplets of angle and dihedral CG variables, $[\theta_1, \varphi_1, \theta_2]$ and $[\varphi_1, \theta_2, \varphi_2]$, and these PDFs both contain φ_i and θ_{i+1}. The MCD is constructed by multiplying the full $[\theta_1, \varphi_1, \theta_2]$ PDF by the distribution of the new variable conditioned on the redundant variables $[\varphi_2|\varphi_1, \theta_2]$. Bayes' rule is applied to replace the conditional distribution with the ratio of the three-dimensional distribution and the PDF of the redundant variables, as in

$$[\theta_1, \varphi_1, \theta_2][\varphi_2|\varphi_1, \theta_2] = \frac{[\theta_1, \varphi_1, \theta_2][\varphi_1, \theta_2, \varphi_2]}{[\varphi_1, \theta_2]}. \quad (1)$$

For some models this process is repeated to build up the PDF with more CG variables until the full MCD is constructed, for example

$$\frac{[\theta_1, \varphi_1, \theta_2][\varphi_1, \theta_2, \varphi_2][\theta_3|\theta_2, \varphi_2]}{[\varphi_1, \theta_2]} = \frac{[\theta_1, \varphi_1, \theta_2][\varphi_1, \theta_2, \varphi_2][\theta_2, \varphi_2, \theta_3]}{[\varphi_1, \theta_2][\theta_2, \varphi_2]}. \quad (2)$$

The MCD for each of the models is listed in the right column of Table I. The MCDs have dimensionality n, which can be as small as $n = 1$ for the Θ model (top section), is $n = 3$ for the middle section, and is $n = 5$ for the bottom section, growing with the spatial extent of the CG variable correlations included in the model.

In order to evaluate observables for a CG model, the MCD is used to construct the $(2N-5)$-dimensional (angle and dihedral) distribution required to model an N-bead chain. This construction has two steps: first, use the MCD to represent the angle and dihedral distribution for an $N = (n + 5)/2$-bead chain. Second, for arbitrary N extend the chain by 1 bead to $N + 1$ by using the MCD. For the Θ model the distribution is extended by multiplying the N-bead approximate distribution by the angle distribution

$$[\theta_1, ..., \theta_{N-2}]_A \times [\theta_{N-1}] \quad (3)$$

For models with $n = 3$, the N-bead approximate distribution is multiplied by the two-dimensional conditional distribution constructed from the 3-dimensional MCD:

$$[\theta_1, \varphi_1, ...\varphi_{N-3}, \theta_{N-2}]_A \times [\varphi_{N-2}, \theta_{N-1}|\theta_{N-2}]_A, \quad (4)$$

while for $n \geq 5$ the general form is

$$[\theta_1, \varphi_1, ...\varphi_{N-3}, \theta_{N-2}]_A \times [\varphi_{N-2}, \theta_{N-1}|\theta_{N-n}, \varphi_{N-n}, ...\varphi_{N-3}, \theta_{N-2}]_A \quad (5)$$

with the subscript $[-]_A$ on both the N-bead distribution and the MCD signifying that they are distributions using the approximate model rather than the full distribution. Bayes' rule is used to eliminate the conditional in Eq. 4 for $n = 3$

$$[\theta_1, \varphi_1, ...\varphi_{N-3}, \theta_{N-2}]_A \times [\theta_{N-2}, \varphi_{N-2}, \theta_{N-1}]_A/[\theta_{N-2}], \quad (6)$$

with the redundant variable θ_{N-2} or Eq. 5 for $n \geq 5$

$$[\theta_1, \varphi_1, ...\varphi_{N-3}, \theta_{N-2}]_A \times [\theta_{N-n}, \varphi_{N-n}, ...\varphi_{N-2}, \theta_{N-1}]_A / [\theta_{N-n}, \varphi_{N-n}, ...\varphi_{N-3}, \theta_{N-2}]_A, \quad (7)$$

using $[\theta_{N-n}, \varphi_{N-n}, ...\varphi_{N-3}, \theta_{N-2}]_A$, the approximate distribution that contains only the components of the model needed to construct the $(2n-3)$ distribution, i.e. the CG variables that are common between the first two terms in Eq. 7. The final result is the $(2N-3)$-dimensional PDF that represents an $(N+1)$-bead chain with a CG model. This procedure ensures that any set of variables β_i, $\beta_{i+\delta_1}$, $\beta_{i+\delta_2}$, with relative indices δ_1 and δ_2 that match a PDF in the model's Minimal Set will follow that PDF.

For the CG models discussed in this paper we take care to approximate the infinite-chain case by discarding data from the bead nearest each chain end in the CG representations. This limits all our CG-variable PDFs to representing the interior chain beads, rather than the chain ends. We apply this practice to the MC simulations by simulating longer chains e.g. $N = 50$ beads, and measuring properties such as CG-variable PDFs or distances for the interior $N = 48$ beads. Throughout this paper values of the number of bonds M or beads N represent an internal segment of the chain and exclude chain ends. This ensures that measurements of the internal-squared distance $\langle R^2(M) \rangle$ match predictions based on extrapolations of the bond-orientation correlation.

D. Analytic Evaluation

It is straightforward to evaluate expectation values from an approximate CG model PDF. For example, computing the expectation value of a function is an integral over the approximate PDF

$$\langle f(\theta_1, \varphi_1, \theta_2...) \rangle = \int\int\int f(\theta_1, \varphi_1, \theta_2...)[\theta_1, \varphi_1, \theta_2, ...]_A d\theta_1 d\varphi_1 d\theta_2... \quad (8)$$

where the subscript $[-]_A$ signifies the approximate model PDF. In principle, any explicit function $f(\cdot)$ can be evaluated.

E. Monte-Carlo Simulations

One simple and effective method for testing the structural properties for approximate CG models is to use MC techniques to sample model bond, angle and dihedral interaction potentials for a single polymer chain. From the $(3N-6)$ dimensional PDF that represents an N-bead chain the interaction potential is constructed in the usual way as

$$U(X_1, X_2, ...) = -k_b T \log([X_1, X_2, ...]), \quad (9)$$

where the 1, 2, or 3 variable PDFs become 1, 2, or 3 variable potential terms. The full-chain potential is sampled by single particle-moves with a displacement Δ and using the Metropolis algorithm. The temperature for the MC sampling is the same as that used for the AA simulations, 550K for PLA, 500K for PE & 653K for PTFE. The initial displacement Δ_i is chosen to be 1, and for an initial, pre-production period of 10,000-100,000 whole-chain MC sweeps Δ is increased or reduced by 10% if it falls outside a 5-95% acceptance range over 100 sweeps. A MC sweep is defined as N randomly chosen particle displacements, where N is the chain length. An entire production MC run is typically of order 10^7 MC sweeps, and results quoted below use data from either the last half or quarter of the MC run.

III. RESULTS AND DISCUSSION

A. Physically Motivated Correlations

Our approximate CG models explicitly include correlations between the CG angle and dihedral variables. It is not clear *a priori* which of these correlations may be important, and here we visually compare the joint PDFs and their independent approximations for an intuitive sense of their importance. Below we compare the PLLA-A3, PTFE-A2 and PTFE-A4 representation. The PTFE-A2 and A4 representations are essentially similar to the representations of PE and illustrate the trends with increasing coarseness. The qualitatively different nature of the correlations between the CG variables in PLLA is also clear.

Figure 2 compares the full two-variable PDF $[\theta_1, \varphi_1]$ (left) for different CG polymer representations with the $\Theta - \Phi$ model approximation (right) for the same representation. The comparison for the PLLA model (top) is particularly poor, with symmetry in the independent approximation that is clearly missing in the full PDF. For the PTFE CG2 model (middle) the peak at $\theta = 160°$ is localized at $\varphi = 0°$, but is smeared out over a much wider range of φ in the independent approximation. Similarly, the peak at $\theta = 130°$ and $\varphi = \pm 140°$ is spread over a wide range of φ and is centered at the wrong location in the independent approximation. Finally, for the PTFE CG4 model (bottom) only a minor asymmetry in θ is neglected in the independent model. These visual cues are

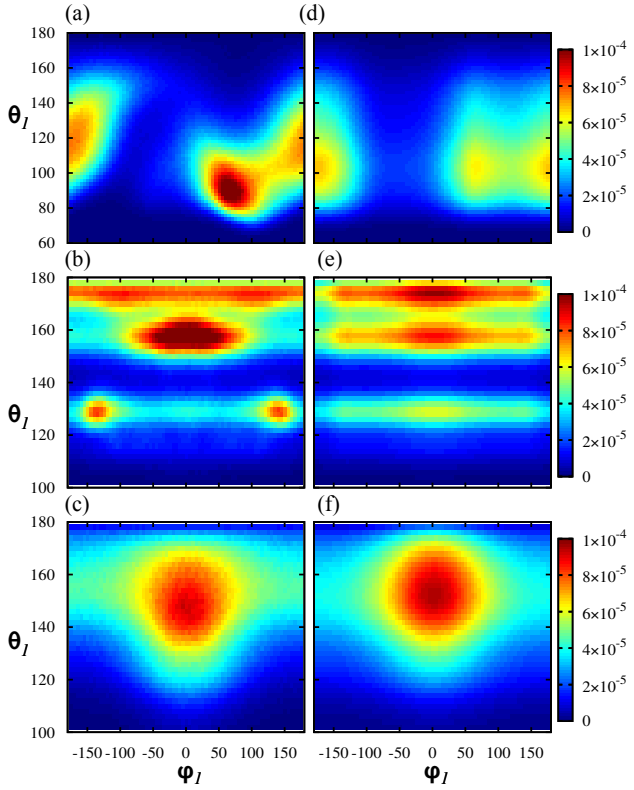

FIG. 2. The joint PDF $[\varphi_1, \theta_1]$ for the (a) PLLA A3, (b) PTFE A2, and (c) PTFE A4 CG representations. (d-f) The corresponding independent approximation $[\varphi_1][\theta_1]$ for the same CG representations.

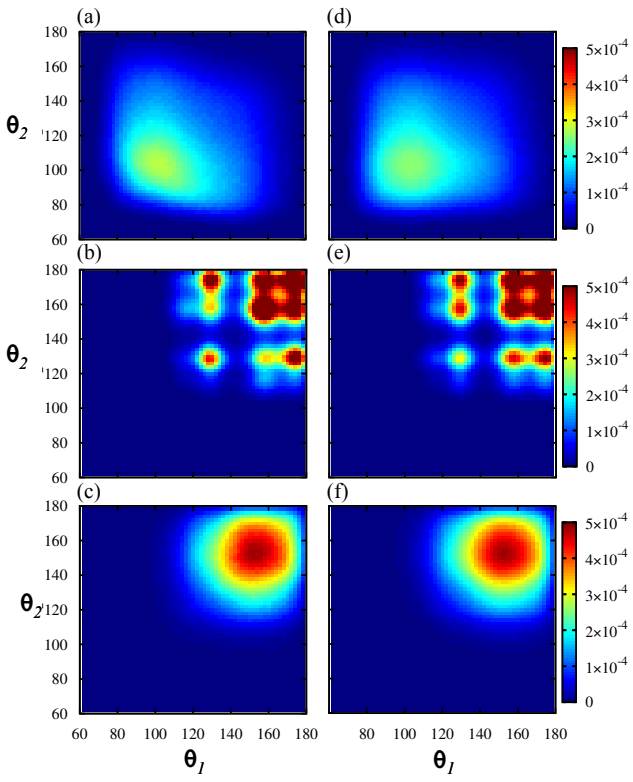

FIG. 3. The joint PDF $[\theta_1, \theta_2]$ for the (a) PLLA A3, (b) PTFE A2, and (c) PTFE A4 CG representations. (d-f) The corresponding independent approximation $[\theta_1][\theta_2]$ for the same CG representations.

good indicators that the independent angle approximations will become increasingly accurate for the models in order from top to bottom.

Calculations of the absolute value of the difference between the full two-dimensional PDF and the independent approximations confirm the visual evaluation of the agreement between the two figures. We use the simple measure

$$\Sigma_{\beta_i,\beta_j} = \int d\beta_i \int d\beta_j \, |[\beta_i, \beta_j] - [\beta_i][\beta_j]| \qquad (10)$$

where β_i and β_j are two CG variables. For example, the error introduced by assuming independent adjacent dihedrals and angles is $\Sigma_{\theta_1,\varphi_1} = \int d\theta_1 \int d\varphi_1 \, |[\theta_1, \varphi_1] - [\theta_1][\varphi_1]|$. The numerical value of the error $\Sigma_{\theta_1,\varphi_1} = 0.422$, 0.185, and 0.109 for the PLLA-A3, PTFE-A2, and PTFE-A4 models, respectively, indicates that the independent angle and dihedral approximation should improve from top to bottom in Fig. 2.

Figure 3 shows the joint PDF of adjacent angles θ_1, θ_2 for PLLA-A3 (top), PTFE-A2 (middle), and PTFE-A4 (bottom). As in Fig. 2 the left side represents the full joint PDF $[\theta_1, \theta_2]$, while the right column shows the product of two independent PDFs $[\theta_1][\theta_2]$. For PTFE, the good match between the right and left PDFs indicates that the $\theta_1 - \theta_2$ correlations may be less important for CG models to capture. Hence, the ΘΦ-ΦΘ and ΘΦ-ΦΘ-ΦΦ models, which omit the $\theta_1 - \theta_2$ correlations, could be effective at modeling the full distribution. The absolute value of the differences between the full two-dimensional $\theta_1 - \theta_2$ PDF and the independent approximation are $\Sigma_{\theta_1,\theta_2} = 0.122$, 0.134, and 0.035 for the PLLA-A3, PTFE-A2, and PTFE-A4 models, respectively, indicating that the PTFE-A4 $\theta_1 - \theta_2$ angles contain the least structure or correlation between adjacent angles, with the value of $\Sigma_{\theta_1,\theta_2}$ 3 times less than the other two models shown.

Figure 4 shows the joint PDF of adjacent dihedrals φ_1, φ_2 for PLLA-A3 (top), PTFE-A2 (middle), and PTFE-A4 (bottom). As in Figs. 2 and 3 the left side represents the full joint PDF $[\varphi_1, \varphi_2]$, while the right column represents the independent-variable approximation. In all cases there is substantial asymmetry in the joint distribution that is not contained in the product PDFs shown

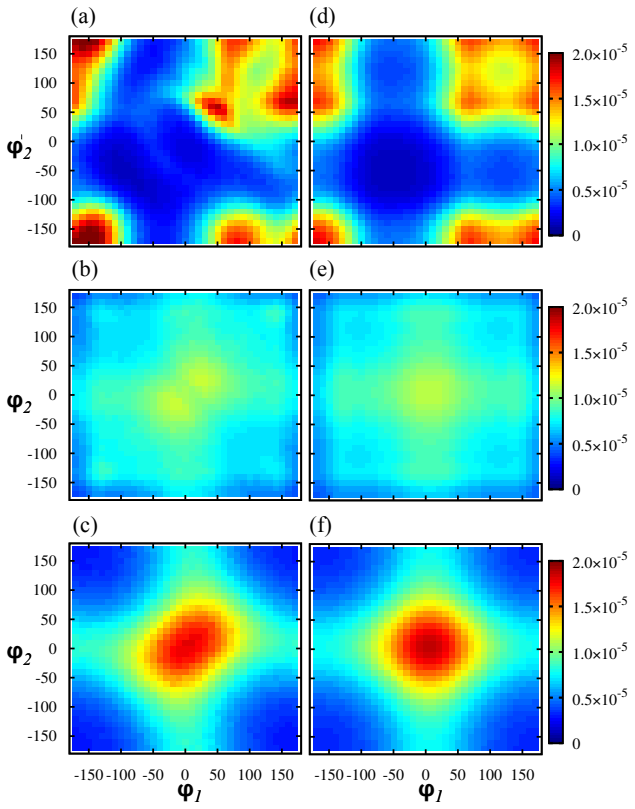

FIG. 4. The joint PDF $[\varphi_1, \varphi_2]$ for the (a) PLLA A3, (b) PTFE A2, and (c) PTFE A4 CG representations. (d-f) The corresponding independent approximation $[\varphi_1][\varphi_2]$ for the same CG representations.

at right. Given this asymmetry one would predict that for all CG representations shown, an approximate model that includes a $[\varphi_1, \varphi_2]$ term would perform better than a model that assumes independent dihedrals. Another notable feature is the overall flatness of the PTFE-A2 representation (middle). The absolute value of the error in the Θ-Φ approximation $\Sigma_{\varphi_1,\varphi_2} = 0.179$, 0.046, and 0.073, for PLLA-A3, PTFE-A2, and PTFE-A4, respectively; 2-3 times smaller for the PTFE models than for PLLA .

Above we have calculated a mismatch between pairwise joint PDFs in the CG angle and dihedral variables and their independent angle and dihedral approximations. It is important to note that while a value of 0 for Σ indicates a perfect match, the measure does not indicate how the mismatch affects observables like the bond-orientation correlation. That is, the quantity Σ says little or nothing about the expectation value of any particular expression evaluated using the PDFs. Instead it is something like an upper bound on the error introduced by the approximation.

B. Bond-Orientation correlation

The bond-orientation correlation links the local CG variables to large-scale structural properties. In this section we discuss how it is computed analytically using an approximate model PDF and numerically using MC sampling of the equivalent potential. We evaluate the bond-orientation correlation for the approximate models and compare those values with results from CG representations of AA simulations.

The bond-orientation correlation

$$C_{bb}(m) \equiv \left\langle \frac{\mathbf{b}_i \cdot \mathbf{b}_{i+m}}{|\mathbf{b}_i||\mathbf{b}_{i+m}|} \right\rangle \tag{11}$$

between two bonds separated by $m-1$ bonds along a polymer chain is an important quantity, since a sum over C_{bb} gives $\langle R^2(M)\rangle$ and related quantities. It is straightforward to write the $m = 1$, 2, and 3 terms in terms of CG angle variables θ_1, φ_1, θ_2, etc. The explicit expressions are

$$C_{bb}(1) = \langle \cos\theta_1 \rangle \tag{12}$$

for the $m = 1$ term,

$$\begin{aligned} C_{bb}(2) =& \langle \cos\theta_1 \cos\theta_2 \rangle \\ & - \langle \sin\theta_1 \sin\theta_2 \cos\varphi_1 \rangle \end{aligned} \tag{13}$$

for the second, and we derive the expression

$$\begin{aligned} C_{bb}(3) =& \langle \cos\theta_1 \cos\theta_2 \cos\theta_3 \rangle \\ & - \langle \cos\theta_1 \sin\theta_2 \sin\theta_3 \cos\varphi_2 \rangle \\ & - \langle \sin\theta_1 \sin\theta_2 \cos\theta_3 \cos\varphi_1 \rangle \\ & + \langle \sin\theta_1 \sin\theta_3 \sin\varphi_1 \sin\varphi_2 \rangle \\ & - \langle \sin\theta_1 \cos\theta_2 \sin\theta_3 \cos\varphi_1 \cos\varphi_2 \rangle \end{aligned} \tag{14}$$

for the third term. Given the expressions in Eqs. 11-14 it is possible to evaluate the bond-orientation correlation for the approximate models proposed using both analytic and MC methods. To analytically evaluate Eqs. 12-14 one first constructs the 5−dimensional PDF $[\theta_1, \varphi_1, \theta_2, \varphi_2, \theta_3]_A$ in the approximate model, and then explicitly evaluates the expectation values. The values of $C_{bb}(m)$ for $m = 1, 2, 3$ can be evaluated via Eqs. 12-14 using MC sampling of a single chain, however with MC it is also possible to calculate the bond-orientation correlation directly for all m using Eq. 11.

For polymers with only local interactions the bond-orientation correlation at large separations m follows an exponential decay,

$$C_{bb}(m) \approx C_0 e^{-m/L_p} \tag{15}$$

with a characteristic decay length L_p known as the persistence length [30]. We previously showed that CG representations of AA PE and PTFE simulation data followed the predicted relationship, $L_p^a = \lambda L_p^{A\lambda}$, where L_p^a is the

backbone AA persistence length and $\lambda L_p^{A\lambda}$ is the persistence length of the corresponding CG representation [15].

The bond-orientation correlation terms $m = 1 - 3$ for the PE-A2 and PE-A4 representation are shown in the upper two curves in Fig. 5 (a). Values from the full CG representation are labeled as $[\theta_1, \varphi_1, \theta_2, \varphi_2, \theta_3]$, and are represented as $*$. The log-linear slope between terms 2 and 3, an estimate of the persistence length L_p, is calculated from the full CG representation and is shown as a dotted line. The bond-orientation correlation values from the ΘΦ-ΦΘ ($\bigtriangledown$) and ΘΦ-ΦΘ-ΦΦ ($\bigtriangleup$) models are also shown. For both PE representations these two models are significantly better than the Θ-Φ model. For both PE-A2 and A4 the ΘΦ-ΦΘ-ΦΦ model best reproduces the 4th and 5th terms in Eq. 14, which contain a product of adjacent dihedrals. As we show below, the ΘΦΘ and ΘΦΘ-ΦΘΦ models (not shown for PE) perform about as well as the ΘΦ-ΦΘ and ΘΦ-ΦΘ-ΦΦ models, respectively. This indicates that for PE the correlations in $[\varphi_1, \varphi_2]$ are more important than those in $[\theta_1, \theta_2]$.

The corresponding values for PLLA are shown in the lower curves in Fig. 5 (a). The different approximate models show similar features to those for PE, with more sophisticated models that take correlations into account improving on the Θ-Φ models. Here it is interesting to note that including the dihedral-dihedral correlation, for example in the ΘΦΘ-ΦΘΦ model versus the ΘΦΘ model, does not necessarily improve the model. The ΘΦΘ-ΘΦΦ model, which was chosen for its ability to reproduce $C_{bb}(3)$, is also shown. As for PE, the log-linear slope between $C_{bb}(2)$ and $C_{bb}(3)$ is shown as a dashed line, and it shows a striking mismatch with the value of $C_{bb}(1)$. We investigate the PLLA data in more detail below.

The bond-orientation correlation calculated from the PE-A4 and PLLA-A3 CG representations using Eq. 11 is shown as a function of separation in Fig. 5(b) to large separation on log-log scales. Values of the bond-orientation correlation measured from CG representations are shown as open symbols, while the exponential form given in Eq. 15 is shown as a dotted (dashed-dot) line for PLLA-A3 (PE-A4). Crossed symbols show results from MC sampling of approximate CG models. The decay length of the exponential form L_p and the prefactor C_0 are estimated from a fit to the values of $C_{bb}(2)$ and $C_{bb}(3)$. It is clear that the PLLA data deviate from a simple exponential form and may follow either a power-law or a second exponential decay length for $\lambda m \geq 15$. Long-range electrostatic interactions are known to modify the simple exponential decay, introducing different correlations [31–33], and in this context the relatively strong polarity of PLLA monomers could modify the bond-orientation correlation. The enhanced correalation at large separations leads to errors of order 3% in sums (discussed below) over the bond-orientation correlation out to separation $m = 45$. In contrast, the exponential form with parameters fit from $C_{bb}(2)$ and $C_{bb}(3)$ is a good match for the PE-A4 data, capturing the persistence length quite well.

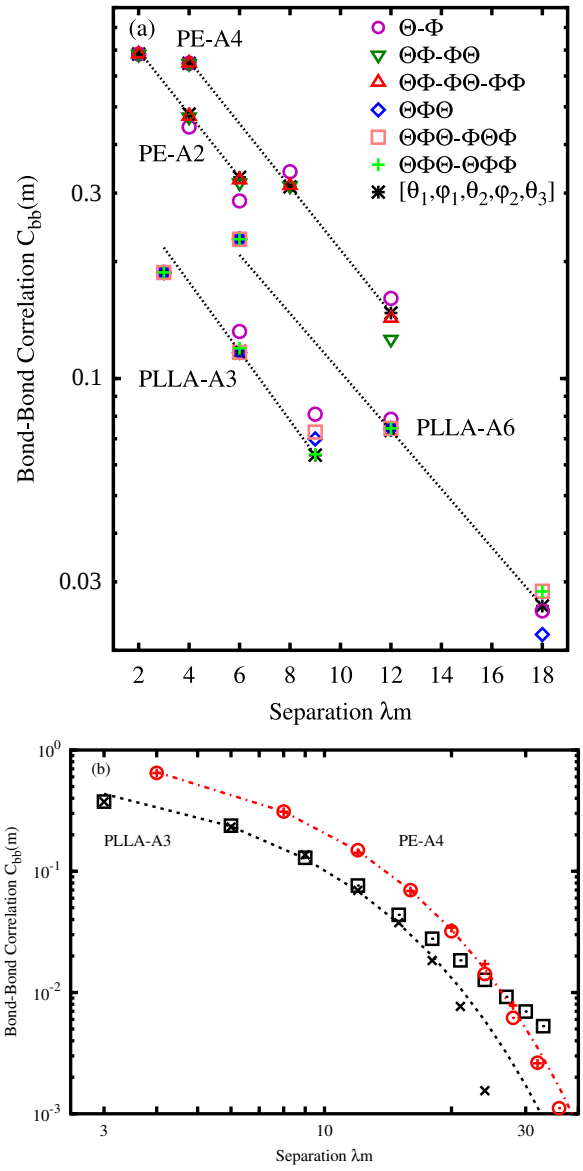

FIG. 5. (a) The bond-orientation correlation $C_{bb}(m)$ for PE and PLLA on log-linear scales. Symbols represent different approximate CG models and dotted lines represent the persistence length determined from the $m = 2$ and $m = 3$ terms of $C_{bb}(m)$ evaluated using the full five-dimensional distribution of CG variables $[\theta_1, \varphi_1, \theta_2, \varphi_2, \theta_3]$. Data for PLLA have been shifted down by a factor of 2. (b) The bond-orientation correlation of the PE-A4 (◯) and PLLA-A3 (□) representations of AA simulation data plotted on log-log axes. The dashed-dot (dotted) lines show the exponential form in Eq. 15 with the decay slope for PE-A4 (PLLA-A3) fit from $C_{bb}(2)$ and $C_{bb}(3)$ as in (a). Data from MC sampling of the PE-A4 ΘΦ-ΦΘ-ΦΦ model (+) and PLLA-A3 ΘΦΘ model (×) are also shown.

For a comparable sum over the PE-A4 representation values, the error is 0.5%.

C. Internal-squared $\langle R^2(M)\rangle$ Distance

The utility of the bond-orientation correlation is its direct link, through summation, to the squared distance between polymer backbone atoms or CG beads a distance M bonds away $\langle R^2(M)\rangle$. As shown above, the first three terms of the bond-orientation correlation can be rapidly and easily calculated for any CG approximate model or in the full CG representation by analytic evaluation of the CG variable distribution. Unfortunately, the internal squared distance $\langle R^2(M)\rangle$ depends on M terms in the bond-orientation correlation, which are not simple to compute analytically. In order to rank the different models using an analytically calculated internal squared distance we must extrapolate the bond-orientation correlation. It is clear from the bond-orientation correlation shown above for PLLA that this process introduces some systematic error. Here we examine the terms in the internal squared distance summation and estimate the error from this extrapolation.

For polymers like PE and PTFE that exhibit purely exponential bond-orientation correlation decay, the molecule-scale structure is dictated by the bond length and the correlation decay length L_p, known as the persistence length. The squared distance between beads separated by M bonds, a fundamental molecule-scale structural measure, is related to a sum over the bond-orientation correlation

$$\langle R^2(M)\rangle = \sum_{i=1}^{M}\sum_{j=1}^{M}\langle \mathbf{b}_i\cdot\mathbf{b}_j\rangle \approx \langle b^2\rangle\left(M+2\sum_{m=1}^{M}(M-m)C_{bb}(m)\right). \quad (16)$$

However, since the $m\geq 4$ values of the bond-orientation correlation $C_{bb}(m)$ become increasingly difficult to compute analytically we would like to determine the importance of the $m\geq 4$ terms in the sum, and to know whether a persistence length extracted from terms $m=2$ and $m=3$ is a good approximation for the rest of the curve. Here we evaluate quantitatively the assumption exponential decay with a constant slope, which appears reasonable based on a visual analysis of the PE-A4 representation data and fit in Fig. 5(b).

Extrapolation of an exponential fit from the $m=2$ and $m=3$ values to the $m\geq 4$ portion bond-orientation correlation curve is equivalent to computing the internal-squared distance as

$$\langle R^2(M)\rangle \approx \langle b^2\rangle\Bigg(M+2(M-1)C_{bb}(1) + 2\sum_{m=2}^{M}(M-m)C_0e^{-m/L_p}\Bigg), \quad (17)$$

where the value of $C_{bb}(1)=\langle\cos\theta\rangle$ is explicitly calculated. Values of C_0 and L_p are determined from a fit to the full CG representation values of $C_{bb}(2)$ and $C_{bb}(3)$. For PE, the value of $\langle R^2(M)\rangle$ predicted by Eq. 17 can be compared with the measured $\langle R^2(M)\rangle$ value computed by summing all the $C_{bb}(m)$ terms as in Eq. 16 for each CG representation. The difference between the actual value and the extrapolated value is less than 1%, except for the A6, and D6 representations which have errors of 3.6%, and 1.6%, respectively. Similarly, for PTFE, the difference between the actual and extrapolated $\langle R^2(M)\rangle$ values is less than 1% except for the D2 and A6 representations which have error of 3% and 1.3%, respectively.

For PLLA, Fig. 5 (b) shows that the exponential form for the bond-orientation correlation decay does not fit the data. Instead a power-law decay or double exponential better describes the data for separation $m\geq 3$. The different form leads to an error of 2.5% between the predicted value in Eq. 17 and the actual squared distance, for the A3 representation shown in Fig. 5 (b). If instead $\langle R^2(M)\rangle$ is predicted using a power-law fit to the data for $m\geq 4$, the extrapolation prediction and full sum values agree to better than 0.1%. In what follows we do not use this power-law fit since it cannot be reproduced by the short-range models that are our focus.

For each of the CG models a similar comparison can be made between the value of $\langle R^2(M)\rangle$ predicted by extrapolating the bond-orientation correlation and that measured using a single-chain MC simulation. The value $\langle R^2(M)\rangle$ measured from MC simulations differs from the predicted values by less than 1% for each of the approximate CG models. This discrepancy indicates the small error in using the exponential fit from $m=2$ and $m=3$ and extending it to the entire chain length.

To confirm that the CG model probability distributions and interaction potentials are equivalent, as they should be by construction, we compare the $C_{bb}(m)$ $m=1,2,3$ values explicitly. These predicted values vary by of order 0.1%, indicating that the MC simulations and analytic calculations are in excellent agreement. For computational speed we use analytic evaluations of $C_{bb}(m)$ and report values of $\langle R^2(M)\rangle$ from extrapolations of these values below.

Overall, these tests indicate error introduced by extrapolating the decay slope of the bond-orientation correlation rather than computing and summing every value is of order 1%. This is true both for CG representations of AA simulations and for approximate CG models sampled with MC. Extrapolation of the bond-orientation from analytic evaluation of the approximate CG models or MC sampling are essentially equivalent. In the following comparison of the different models we show the error introduced by extrapolation as an approximate measure of the systematic error.

D. Approximate Model Error

Our ultimate goal is to predict structural properties using CG polymer models. Since the CG mod-

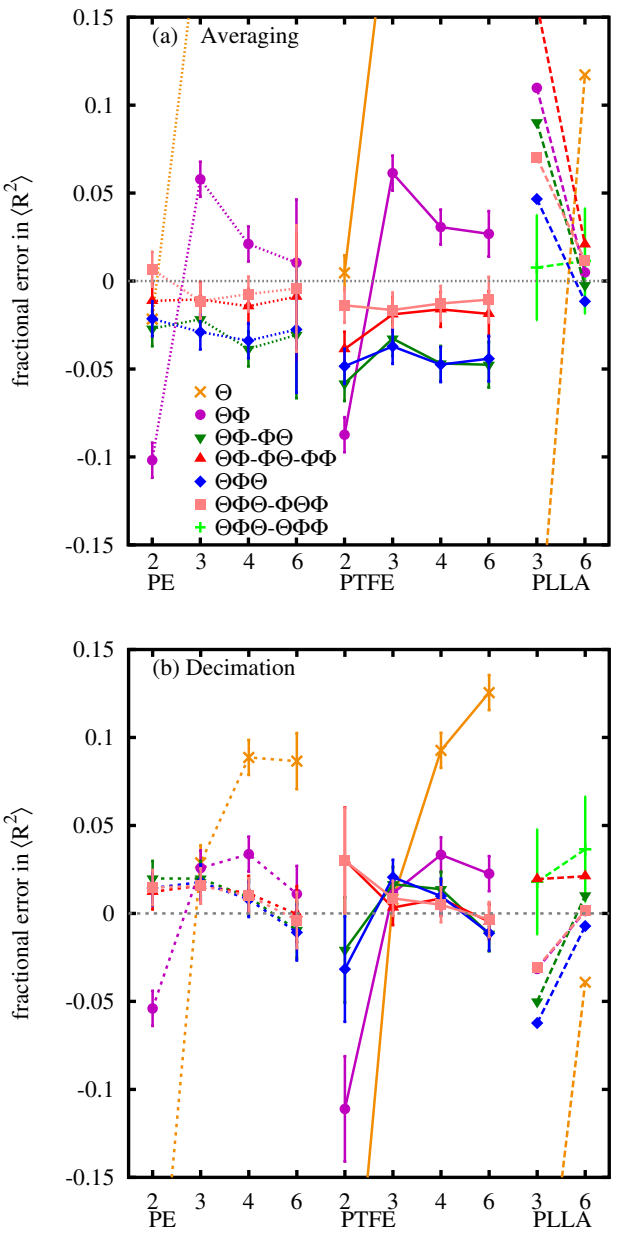

FIG. 6. The fractional error in the squared internal-squared distance $\langle R^2(M)\rangle$ for the different CG polymer approximate models in Table I relative to the true value of $\langle R^2(M)\rangle$. Both the true and predicted values of $\langle R^2(M)\rangle$ omit end effects. Error bars represent the systematic error in the calculation as described in Section IIIc, which are of order 1% for PE and PTFE, and of order 3% for PLLA.

els are constructed from CG variable PDFs the short-range structural properties are correct by construction, but molecule-scale structural properties are only approximately matched. Models can be evaluated on this large-scale structure by comparing the internal-squared distance values predicted from each of the approximate CG models with those measured from CG representations of AA simulations. Results from this comparison clearly identify which correlations are important to include in approximate CG models of PE and PTFE, and indicate the difficulties in modeling a polymer like PLLA that contains long-range correlations using CG models that do not include such interactions.

Figure 6 shows the fractional error between the extrapolated squared distance $\langle R^2(M)\rangle$ computed from Eq. 17 using approximate CG models and the measured value from CG representations of AA simulations. The top panel shows the models created by averaging, while those from decimation are shown in the bottom panel. For PE and PTFE both the Aλ and Dλ approximate models have similar features. The Θ models do not reproduce the chain stiffness correctly, except at a representation-dependent value of A$\lambda^* = 2$ and D$\lambda^* = 3$ as found previously [15]. The Θ-Φ model approximation is poor for $\lambda = 2$ and $\lambda = 3$ for the Aλ models, becoming better for coarser beads. This has an intuitive physical picture, that angle-dihedral correlations become less important for coarser models. For the Aλ approximate models it is clear that including the dihedral-dihedral $(\varphi_1 - \varphi_2)$ correlation is important. The models that include this correlation ($\triangle$ and $\square$) perform better than the models which omit it ($\triangledown$ and $\diamond$). In contrast, it is not clear that including the angle-angle correlations $(\theta_1 - \theta_2)$ is necessary. In contrast, for models based on decimation representations the importance of the dihedral-dihedral correlations appears diminished relative to averaging, except for the PTFE D2 model.

The approximate CG PLLA models are less successful than the PE and PTFE models. The CG variable correlations are clearly very strong, as can be seen in the $[\theta_1, \varphi_1]$ and $[\varphi_1, \varphi_2]$ PDFs in Figs. 2 and 4. Not only are these correlations important, but so are the correlations included in the $\Theta\Phi\Theta$-$\Theta\Phi\Phi$ model, which includes the term $[\theta_1, \varphi_1, \varphi_2]$. Unfortunately, all the approximate PLLA models perform poorly because they are constructed with an assumption of exponential bond-orientation correlation decay, which is not observed in this polymer. At the coarser $\lambda = 6$ level it appears that the models all match $\langle R^2(M)\rangle$ better than $\lambda = 3$, which is due to the long-range correlations being better represented by an exponential form with coarser beads.

IV. CONCLUSION

We have developed seven approximate CG models of PE, PTFE, and PLLA from CG representations of AA MD simulations. The models are implemented as either high-dimensional PDFs of CG angle and dihedral variables or as an equivalent interaction potential. Approximate CG models are constructed from low-dimensional CG variable PDFs that contain correlations between different CG variables including dihedral-dihedral, dihedral-angle, and angle-angle. We show that these CG-variable combinations contain significant correlations.

The $m = 1, 2, 3$ terms of the bond-orientation correlation $C_{bb}(m)$ are calculated with each approximate model analytically from the unique 5-dimensional PDF of CG angles and dihedrals. All $(N-2)$ terms of the bond-orientation correlation are computed for an N-bead chain using single-chain MC sampling for each model. Results for the internal squared distance $\langle R^2(M)\rangle$ from MC sampling and from an exponential extrapolation of analytical bond-orientation correlation decay are in agreement, confirming the accuracy of the analytical approximation.

The internal squared distance for the approximate models is compared with the value from the CG representation of the AA polymer simulations. For PE and PTFE, models generated from an averaging procedure that include dihedral-dihedral correlations perform best. Angle-dihedral correlations also have considerable importance, while angle-angle correlations have negligible effect. For models generated from a decimation procedure the angle-dihedral correlations are again important and models with dihedral-dihedral correlations give only a slight improvement. Evaluation of the model performance based on $\langle R^2(M)\rangle$ values from extrapolations of the bond-orientation correlation gives a very clear ranking of the relative importance of the different CG-variable correlations.

For PLLA, extrapolation of the bond-orientation correlation and comparison between the AA CG representation and CG model $\langle R^2(M)\rangle$ values gives a different type of insight. Unlike the case for PE and PTFE, for PLLA there is a mismatch between model and representation $\langle R^2(M)\rangle$ values even when the first three terms are a good match because the electrostatic interactions present for PLLA break the assumption of short-range interactions in our approximate CG models. Physically, these long-range interactions manifest as an enhancement of the bond-orientation correlation, which is not reproduced by either CG PDFs or local interaction potentials of CG variables.

Our overall goal was to use the bond-orientation correlation to evaluate approximate CG models for three common, different polymers. This approach proved remarkably robust. From our analysis it is clear that PLLA's electrostatic interactions, which manifest in the bond-orientation correlation, are not captured by the simple approximate models proposed. Further work will be needed to include long-range (electrostatic) interactions in CG polymer models [34] and to understand how to quickly evaluate the structural properties of polymers within those models. In contrast, the analysis was very successful for PE and PTFE. For these polymers the most important CG variable correlations were easily identified and the error in the large-scale structure for each CG model was quantified. We showed that more sophisticated CG models with dihedral-angle and dihedral-dihedral correlations can reproduce the structure of PE and PTFE with high fidelity. This success shows that for a wide class of polymers approximate CG models can be quickly evaluated for the ability to reproduce large-scale structure accurately.

[1] H. Meyer and F. Müller-Plathe, The Journal of Chemical Physics **115**, 7807 (2001), http://dx.doi.org/10.1063/1.1415456.

[2] P. Gao and H. Guo, Phys. Chem. Chem. Phys. **17**, 31693 (2015).

[3] I. A. Strelnikov, E. A. Zubova, M. A. Mazo, and L. I. Manevich, Polymer Science, Series A **59**, 242 (2017).

[4] S. W. I. Siu, K. Pluhackova, and R. A. Böckmann, J. Chem. Theory Comput. **8**, 1459 (2012).

[5] A. M. Donald and E. J. Kramer, Journal of Polymer Science: Polymer Physics Edition **20**, 899 (1982), https://onlinelibrary.wiley.com/doi/pdf/10.1002/pol.1982.180200505.

[6] F. Gittes and F. C. MacKintosh, Phys. Rev. E **58**, R1241 (1998).

[7] C. Bouchiat, M. Wang, J.-F. Allemand, T. Strick, S. Block, and V. Croquette, Biophysical Journal **76**, 409 (1999).

[8] J. Rottler, S. Barsky, and M. O. Robbins, Phys. Rev. Lett. **89**, 148304 (2002).

[9] R. Everaers, S. K. Sukumaran, G. S. Grest, C. Svaneborg, A. Sivasubramanian, and K. Kremer, Science **303**, 823 (2004), http://science.sciencemag.org/content/303/5659/823.full.pdf.

[10] D. Curcó and C. Alemán, Chem. Phys. Lett. **436**, 189 (2007).

[11] D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, and K. Kremer, Phys. Chem. Chem. Phys. **13**, 10412 (2011).

[12] K. M. Salerno, A. Agrawal, D. Perahia, and G. S. Grest, Phys. Rev. Lett. **116**, 058302 (2016).

[13] D. D. Hsu, W. Xia, S. G. Arturo, and S. Keten, Macromolecules **48**, 3057 (2015), http://dx.doi.org/10.1021/acs.macromol.5b00259.

[14] V. A. Harmandaris, D. Reith, N. F. A. van der Vegt, and K. Kremer, Macromolecular Chemistry and Physics **208**, 2109 (2007).

[15] K. M. Salerno and N. Bernstein, Journal of Chemical Theory and Computation **14**, 2219 (2018), pMID: 29229375, https://doi.org/10.1021/acs.jctc.7b01229.

[16] M. Dallavalle and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. , (2017).

[17] D. Reith, H. Meyer, and F. Mller-Plathe, Macromolecules **34**, 2335 (2001), http://dx.doi.org/10.1021/ma001499k.

[18] H. Fukunaga, J.-i. Takimoto, and M. Doi, J. Chem. Phys. **116**, 8183 (2002).

[19] D. Garlotta, Journal of Polymers and the Environment **9**, 63 (2001).

[20] S. Plimpton, J. Comput. Phys. **117**, 1 (1995).

[21] K. M. Salerno, A. Agrawal, B. L. Peters, D. Perahia, and G. S. Grest, The European Physical Journal Special Topics **225**, 1707 (2016).

[22] B. L. Peters, K. M. Salerno, A. Agrawal, D. Perahia, and G. S. Grest, Journal of Chem-

ical Theory and Computation **0**, null (0), http://dx.doi.org/10.1021/acs.jctc.7b00241.
[23] J. G. Kirkwood, The Journal of Chemical Physics **3**, 300 (1935), https://doi.org/10.1063/1.1749657.
[24] J. G. Kirkwood and E. M. Boggs, The Journal of Chemical Physics **10**, 394 (1942), https://doi.org/10.1063/1.1723737.
[25] P. Attard, O. G. Jepps, and S. Marčelja, Phys. Rev. E **56**, 4052 (1997).
[26] A. Singer, The Journal of Chemical Physics **121**, 3657 (2004), https://doi.org/10.1063/1.1776552.
[27] S. Somani, B. J. Killian, and M. K. Gilson, The Journal of Chemical Physics **130**, 134102 (2009), https://doi.org/10.1063/1.3088434.
[28] B. J. Killian, J. Y. Kravitz, and M. K. Gilson, The Journal of Chemical Physics **127**, 024107 (2007), https://doi.org/10.1063/1.2746329.
[29] A lower-dimensional distribution could be used to *extend* an approximate distribution in one direction, but only the symmetric distribution can be used for both ends of the polymer chain.
[30] M. Rubinstein and R. H. Colby, *Polymer Physics*, Vol. 23 (Oxford University Press, New York, 2003).
[31] M. J. Stevens, D. B. McIntosh, and O. A. Saleh, Macromolecules **46**, 6369 (2013), https://doi.org/10.1021/ma401211w.
[32] N. M. Toan and D. Thirumalai, The Journal of Chemical Physics **136**, 235103 (2012), https://doi.org/10.1063/1.4729371.
[33] J.-L. Barrat and J.-F. Joanny, EPL (Europhysics Letters) **24**, 333 (1993).
[34] G. Deichmann and N. F. A. van der Vegt, Journal of Chemical Theory and Computation **13**, 6158 (2017), pMID: 29161026, https://doi.org/10.1021/acs.jctc.7b00611.

Appendix A: Autocorrelation/Independence of samples

Figure 7 shows the autocorrelation of individual CG bond-length, bond-angle, and dihedral-angle variables during an AA simulation. Our analysis indicates that the longest decorrelation time for CG variables is approximately 100-150 ps for the dihedral-angle variables.

Constructing low-order PDFs of CG-variables means generating histograms of the relevant variables, which requires enough independent samples to construct a multi-dimensional histogram. Simulation times are of order 50-100 ns, representing about 500 independent sampling times. Typical all-atom simulation sizes are about 50,000 CG bonds, for a total of about 2.5×10^7 independent CG variable samples. For a three-dimensional histogram this number of samples produces hundreds to thousands of counts per histogram bin, given bin widths that adequately resolve the probability distribution. Unfortunately, this is nearly the limit for resolution, since increasing the dimensionality increases the number of bins without increasing the number of samples, thereby decreasing the counts in each bin by more than one order of magnitude. For the CG-variables of interest here four and five-dimensional distributions may be possible with considerable effort, but for practical considerations we use one-, two-, and three-dimensional distributions as the Minimal Set of PDFs for CG models.

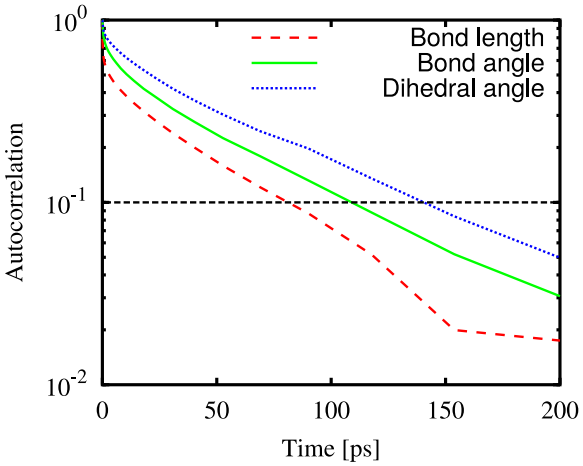

FIG. 7. The autocorrelation of individual CG bond-length, bond-angle and dihedral-angle variables over a 10 ns atomistic simulation. Curves are values averaged over 20 different bonds or angles in the middle of a chain.