In the eye of the beholder: Inhomogeneous distribution of high-resolution shapes within the random-walk ensemble

Christian L. Müller, Ivo F. Sbalzarini, Wilfred F. van Gunsteren, Bojan Žagrović, and Philippe H. Hünenberger

Citation: J. Chem. Phys. **130**, 214904 (2009); doi: 10.1063/1.3140090
View online: http://dx.doi.org/10.1063/1.3140090
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v130/i21
Published by the American Institute of Physics.

In the eye of the beholder: Inhomogeneous distribution of high-resolution shapes within the random-walk ensemble

Christian L. Müller,[1] Ivo F. Sbalzarini,[1,2] Wilfred F. van Gunsteren,[3] Bojan Žagrović,[2] and Philippe H. Hünenberger[3,a)]
[1]*Institute of Computational Science and Swiss Institute of Bioinformatics, ETH Zürich, CH-8092 Zürich, Switzerland*
[2]*Laboratory of Computational Biophysics, Mediterranean Institute for Life Sciences, HR-21000 Split, Croatia*
[3]*Laboratory of Physical Chemistry, ETH Zürich, CH-8093 Zürich, Switzerland*

(Received 1 October 2008; accepted 17 April 2009; published online 1 June 2009)

The concept of high-resolution shapes (also referred to as folds or states, depending on the context) of a polymer chain plays a central role in polymer science, structural biology, bioinformatics, and biopolymer dynamics. However, although the idea of shape is intuitively very useful, there is no unambiguous mathematical definition for this concept. In the present work, the distributions of high-resolution shapes within the ideal random-walk ensembles with $N=3,\ldots,6$ beads (or up to $N=10$ for some properties) are investigated using a systematic (grid-based) approach based on a simple working definition of shapes relying on the root-mean-square atomic positional deviation as a metric (i.e., to define the distance between pairs of structures) and a single cutoff criterion for the shape assignment. Although the random-walk ensemble appears to represent the paramount of homogeneity and randomness, this analysis reveals that the distribution of shapes within this ensemble, i.e., in the total absence of interatomic interactions characteristic of a specific polymer (beyond the generic connectivity constraint), is significantly inhomogeneous. In particular, a specific (densest) shape occurs with a local probability that is 1.28, 1.79, 2.94, and 10.05 times ($N=3,\ldots,6$) higher than the corresponding average over all possible shapes (these results can tentatively be extrapolated to a factor as large as about 10^{28} for $N=100$). The qualitative results of this analysis lead to a few rather counterintuitive suggestions, namely, that, e.g., (i) a fold classification analysis applied to the random-walk ensemble would lead to the identification of random-walk "folds;" (ii) a clustering analysis applied to the random-walk ensemble would also lead to the identification random-walk "states" and associated relative free energies; and (iii) a random-walk ensemble of polymer chains could lead to well-defined diffraction patterns in hypothetical fiber or crystal diffraction experiments. The inhomogeneous nature of the shape probability distribution identified here for random walks may represent a significant underlying baseline effect in the analysis of real polymer chain ensembles (i.e., in the presence of specific interatomic interactions). As a consequence, a part of what is called a polymer shape may actually reside just "in the eye of the beholder" rather than in the nature of the interactions between the constituting atoms, and the corresponding observation-related bias should be taken into account when drawing conclusions from shape analyses as applied to real structural ensembles.
 [DOI: 10.1063/1.3140090]

I. INTRODUCTION

From a broad perspective, the ideal random-walk model is a mathematical device to emulate ensemble statistics on sequences of events (defined by successive points in a given space) by assuming that the displacements between two successive events along the sequence are (i) of constant length (a common variant considers a Gaussian length distribution instead) and (ii) entirely uncorrelated in their directions. In spite of its apparent simplicity, this model already provides a very reasonable approximate description of many phenomena involving weakly correlated successive displacements, such as diffusion of molecules in gases, liquids, or solids,[1] bacterial chemotaxis,[2] population dynamics,[3] fixational eye movements,[4] behavior of economic markets,[5] gambling,[6] and structure of polymer chains.[7–11] The present article exclusively focuses on the latter (structural) type of application. In this case, events correspond to the occurrence of specific polymer sites (e.g., atom or reference point of a monomer) in three-dimensional Cartesian space, while the sequence of events is determined by the covalent connectivity of the polymer chain and the constant displacement by the (pseudo)bond length between successive sites.

The mathematical characterization of the structural random-walk ensemble has led to a number of classical analytical results concerning the distributions or averages of

a)Author to whom correspondence should be addressed. Laboratory of Physical Chemistry, ETH Hönggerberg, HCI, CH-8093 Zürich, Switzerland. Tel.: +41-44-632-55-03. FAX: +41-44-632-10-39. Electronic mail: phil@igc.phys.chem.ethz.ch.

low-resolution properties, such as the end-to-end distance and the radius of gyration.[7,8,10] More recently, the full radius of gyration tensor was also investigated through a combination of analytical, computational, and experimental approaches.[11–18] These studies revealed a seemingly counterintuitive feature: Although all individual steps of the walk are taken in random (i.e., isotropically distributed) directions, anisotropic shapes (nonequal eigenvalues of the radius of gyration tensor) are actually more probable than isotropic ones within the random-walk ensemble. As a result, the average shape of a walk over the entire ensemble is not spherical but a prolate ellipsoid with the ratio of eigenvalues (mean-square principal radii of gyration) of approximately 9:2.3:1.[11,13]

In spite of the general interest of the above results, the end-to-end distance and the radius of gyration (scalar or tensor) only provide a very low-resolution characterization of polymer shapes. In contrast, much less is known concerning the distributions of high-resolution properties (i.e., detailed three-dimensional shapes) within the random-walk ensemble.

The concept of high-resolution shape (also referred to as fold or state, depending on the context) of a polymer chain plays a central role in structural biology and bioinformatics (e.g., protein structure classification[19–22] and prediction[23–26]), as well as in biopolymer dynamics (e.g., definition of states based on clustering[27–33] of computer simulation trajectories and evaluation of their thermodynamical properties based on the corresponding populations[34–38]). However, although the idea of shape is intuitively very useful, there is no unambiguous mathematical definition for this concept, and many alternative formulations have been used in the field, involving, in particular, different comparison metrics.[11,39–46]

A possible way to define polymer shapes based on the ensemble of all possible structures with a specified chain length and composition (i.e., all possible sets of coordinates for the constituting atoms) is based on the specification of (i) a pairwise metric for the determination of the difference (or distance) between two structures and (ii) a cutoff distance applied to this metric for the assignment of individual structures to a common shape. More specifically, given these two choices, the shape of a polymer chain associated with a specific (central) structure may be defined as the collection of all structures for which the distance to this central structure is below the cutoff value. This definition implies, in particular, that (i) every structure can be used to define a corresponding shape (of which it is the central structure) and (ii) that different shapes may be overlapping in terms of the structures they encompass (i.e., individual structures are not associated to a single shape). This definition has been adopted in the present work due to its simplicity, but alternative definitions are also possible (and commonly used). For example, a nonoverlapping set of shapes may be constructed by tesselation of the structural space around a finite number of central structures (e.g., by clustering,[27–34] the definition is then more complex and must include a generalization of the cutoff parameter for determining the clustering resolution, i.e., the approximate number and sizes of the shapes to be produced).

Ideally, the pairwise metric for evaluating structural differences should match our intuitive feeling of how dissimilar two structures "look like." In particular, it should be (i) independent of the (rigid-body) relative positioning and orientation of the two compared structures and (ii) unaffected by performing a mirror symmetry, a central inversion, or an atom renumbering on the two compared structures. This metric can characterize the structural difference either partially (i.e., at low resolution, e.g., difference in the end-to-end distance,[47] radius of gyration,[48] radius of gyration tensor anisotropy,[13] number of hydrogen bonds,[48,49] or number of interatomic contacts[50]) or globally (i.e., at high resolution, e.g., root-mean-square atomic positional deviation[39,51–55] as well as its so-called unit-vector[41] and universal[43] variants, distance-matrix root-mean-square difference,[36,56] ρ measure,[40] or TM-align parameter[46]). In general, the high-resolution comparisons will be more meaningful when performed in terms of Cartesian (as opposed to internal) coordinates because our intuitive feeling concerning structural similarity refers to our vision in three-dimensional Cartesian space.

The most commonly used comparison metric in structural biology and biopolymer dynamics is by and large the atom-positional root-mean-square deviation (RMSD) after least-squares rototranslational fitting.[39,51–55] The RMSD is easy to calculate and corresponds well to our intuitive notion of structural dissimilarity. It can be applied to entire biomolecules (e.g., comparison of protein structures) as well as to subunits thereof (e.g., comparison of structural motifs adopted by the same amino-acid sequence in different proteins). The RMSD has been used in many different contexts. These include, for example, evaluating the quality of structure prediction schemes,[23–26] monitoring structural changes during protein folding,[27,28,34,36,48] comparing the diversity of model structures derived from experiments,[57,58] comparing the properties of modeling approaches at different levels of resolution,[59,60] evaluating the extent of conformational space accessible to a polymer via N-cube analysis,[59] or constructing cumulative distribution functions for polymer conformational ensembles.[60] For these reasons, the RMSD was adopted in the present work for the definition of high-resolution shapes.

A common (implicit) assumption in the analysis of polymer chain structures in terms of shapes (as defined above, i.e., using a common metric and cutoff value for all shapes) is that the relative populations of the different shapes in an ensemble of structures at equilibrium (e.g., a protein in solution) only depend on the energetics of the structures they encompass. As will be shown in the present study, this assumption is incorrect. According to this assumption, one would expect that in the absence of any energetics (interatomic interactions restricted to a mere connectivity constraint, that is, in the random-walk ensemble), the probability distribution of shapes is homogeneous, i.e., all shapes are equally probable. This, however, need not be the case: The neighborhood of specific structures may be intrinsically more populated than the neighborhood of others. In particular, for a given metric and cutoff, there is generally a most probable shape (or a set thereof) in the random-walk ensemble, i.e., structures taken at random from the ensemble fall with the

highest probability within this shape (or these shapes) compared to any other one. This inhomogeneity in the probability distribution of shapes in the random-walk ensemble will depend on the chosen pairwise metric and cutoff value, i.e., on two choices that characterize the way we have decided to "look" at structures rather than the physics of the system itself (i.e., the interactions between the constituting atoms).

The goal of the present work is to investigate the nature and magnitude of this intrinsic bias in high-resolution shape analysis. This is done by considering the ideal random-walk ensemble, along with the above simple definition of high-resolution shapes, the commonly used RMSD metric, and various cutoff distances (both infinitesimal and finite). For completeness, the distributions of various low-resolution shape parameters (end-to-end distance, radius of gyration, radius of gyration tensor anisotropy, and number of contacts) over the random-walk ensemble are also evaluated and compared to analytical results whenever possible. This investigation is carried out using a systematic (grid-based) sampling of the random-walk ensemble in terms of internal coordinates. Due to the exponential growth in the computational cost of any systematic approach with the system size (at fixed grid spacing), the analysis is restricted to relatively short (up to at most 6–10 beads depending on the analysis), oriented, unbranched, and non-self-avoiding random walks. These can be viewed as a highly simplified model for polypeptides (where each bead would represent e.g., the C_α of an amino-acid residue from the N-terminus to the C-terminus with a pseudobond distance of about[61] 0.38 nm). A few of these calculations are also performed for self-avoiding walks (i.e., introducing an excluded volume for the beads).

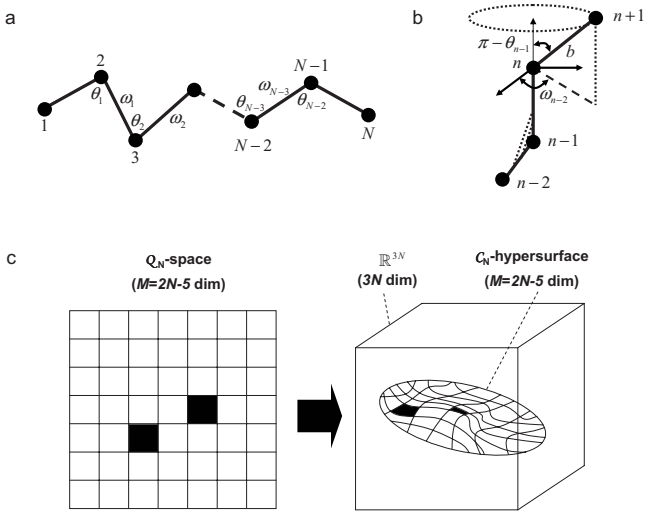

FIG. 1. (a) Definition of the angles θ_n and dihedral angles ω_n characterizing an anchored walk of length N (Sec. II A). (b) Illustration of the conversion $\mathbf{q}\rightarrow\mathbf{r}(\mathbf{q})$ from an internal coordinate vector $\mathbf{q}$ to the corresponding Cartesian coordinate vector $\mathbf{r}$ via trigonometry (Sec. II A). (c) Schematic of the coordinate transformation from $\mathcal{Q}_N$ (M-dimensional internal coordinate space, with $M=2N-5$) to $\mathcal{C}_N$ (M-dimensional hypersurface within the entire Cartesian coordinate space $\mathbb{R}^{3N}$) for anchored walks of length N (Sec. II A). $\mathcal{Q}_N$ is bounded and of finite volume $V_{\mathcal{Q}_N}$ [Eq. (2)]. $\mathcal{C}_N$ is also bounded and of finite area $A_{\mathcal{C}_N}$. Note that the $N-3$ dihedral-angle coordinates within $\mathbf{q}$ are actually periodic (i.e., bounded only by the definition of a reference interval). For simplicity, this periodicity (i.e., the "folding" of a part of the boundary of $\mathcal{Q}_N$ or $\mathcal{C}_N$ on itself) is not represented in the figure. The same drawing could also illustrate the coordinate transformation from $\mathcal{Q}_N$ to $\tilde{\mathcal{R}}_N$ (M dimensions within $\mathbb{R}^{3N}$), the hypersurface associated with the local RMSD metric $\tilde{R}$ of Eq. (7) (Sec. II B). In this case, the fact that two infinitesimal volumes of $\mathcal{Q}_N$ transform to patches of different areas of $\tilde{\mathcal{R}}_N$ indicates that the corresponding shapes have different local probabilities $P_N(\mathbf{q},0)$ [Sec. II E; Eqs. (18) and (19)]. In the present case, the patch on the right is representative of a (locally) more probable shape compared to the one on the left.

II. THEORY

A. Walks

A walk of length N and step size b corresponds to the path obtained (in three-dimensional Cartesian space) by starting at some origin P_1 and taking $N-1$ successive (straight) steps of common length b in arbitrary directions, as illustrated in Fig. 1(a). The N points $\{P_n|n=1,\ldots,N\}$ along such a path will be referred to here as beads and the corresponding steps as bonds (although they are most often rather pseudobonds). Note that walks defined in this way are (i) unbranched (linear topology), (ii) non-self-avoiding (beads may be positioned arbitrarily close to each other except for consecutive ones), and (iii) oriented (a walk is distinct from its reverse walk, as defined by taking the beads in a reverse order). In the following, it is also assumed that $N\geq 3$.

A walk can be entirely specified by the $3N$-dimensional vector $\mathbf{r}\doteq\{\mathbf{r}_n|n=1,\ldots,N\}=\{r_\alpha|\alpha=1,\ldots,3N\}$, where $\mathbf{r}_n$ is the Cartesian coordinate vector of bead n and r_α a single Cartesian coordinate within $\mathbf{r}$. To avoid the redundancy of walks that can be superimposed by a trivial rigid-body translation and rotation, it is convenient to define anchored walks as the walks satisfying the six additional constraints $r_\alpha=0$ for $\alpha=1,2,3,5,6,9$, along with $r_4=b$ and $(r_7-b)^2+r_8^2=b^2$. In other words, for an anchored walk, the P_1 bead is placed at the origin, the P_1-P_2 bond aligned along the x-axis, and the P_2-P_3 bond contained in the xy-plane, which uniquely defines the overall (rigid-body) position and orientation of the walk. Due to the $N-1$ bond plus six rigid-body constraints, the space $\mathcal{C}_N$ spanned by the Cartesian coordinate vectors $\mathbf{r}$ associated with all anchored walks of length N represents a M-dimensional hypersurface within $\mathbb{R}^{3N}$, where $M\doteq 2N-5$.

Alternatively, an anchored walk can also be entirely specified by a M-dimensional internal coordinate vector $\mathbf{q}\doteq\{q_n|n=1,\ldots,M\}$, where q_n $(n=1,\ldots,N-2)$ is the cosine of the angle θ_n formed by $P_n-P_{n+1}-P_{n+2}$ and q_{n+N-2} $(n=1,\ldots,N-3)$ is the dihedral angle ω_n formed by $P_n-P_{n+1}-P_{n+2}-P_{n+3}$ (oriented and measured in radians) [see Fig. 1(a)]. This can be written as

$$\mathbf{q}\doteq\{\{\cos(\theta_n)|n=1,\ldots,N-2\},\{\omega_n|n=1,\ldots,N-3\}\}. \tag{1}$$

The $N-2$ angle-cosine coordinates are nonperiodic and bounded to the range $]-1;1]$. The $N-3$ dihedral-angle coordinates are periodic and chosen here (by convention) within the range $[-\pi;\pi]$. The M-dimensional space $\mathcal{Q}_N$ spanned by the internal coordinate vectors $\mathbf{q}$ corresponding to all anchored walks of length N is thus compact (no "holes") and bounded, with a finite volume $V_{\mathcal{Q}_N}$ given by

$$V_{\mathcal{Q}_N} \doteq \int_{\mathcal{Q}_N} d^M\mathbf{q} = 2^{N-2}(2\pi)^{N-3} = 2^M\pi^{(M-1)/2}. \tag{2}$$

Note that if any bond angle θ_n of the walk is equal to 0 or π, the preceding and succeeding dihedral angles (ω_{n-1} and ω_n) are undefined and should be replaced by a single dihedral angle $P_{n-1}-P_n-P_{n+2}-P_{n+3}$ (assuming that θ_{n-1} and θ_{n+1} themselves differ from 0 and π). Because these special walks belong to the boundary of $\mathcal{Q}_N$, such a special handling turns out, however, to be unnecessary for the present purposes (Secs. II E and III).

The mapping $\mathbf{q}\rightarrow\mathbf{r}(\mathbf{q})$ of an anchored walk from $\mathcal{Q}_N$ to $\mathcal{C}_N$, as well as the reverse mapping $\mathbf{r}\rightarrow\mathbf{q}(\mathbf{r})$ from $\mathcal{C}_N$ to $\mathcal{Q}_N$, is defined and unique (except for walks involving one or more bond angles equal to 0 or π), as well as continuous. Both transformations can be performed using straightforward trigonometry, as illustrated in Fig. 1(b). Because $\mathcal{Q}_N$ is compact and bounded (with a finite volume $V_{\mathcal{Q}_N}$), the uniqueness and continuity of the transformation implies that the hypersurface $\mathcal{C}_N$ (within $\mathbb{R}^{3N}$) is also compact and bounded, with a finite area $A_{\mathcal{C}_N}$. The relationship between the two spaces is schematically illustrated in Fig. 1(c).

B. Pairwise metric

The pairwise metric used in the present work for structure comparison is the root-mean-square atomic positional deviation R after least-squares rototranslational fitting (RMSD[39,51–55]). For a given reference structure $\mathbf{r}$ and a given compared structure $\mathbf{r}'$, R is the metric representing the scalar distance between two associated $3N$-dimensional vectors $\mathbf{s}(\mathbf{r})$ and $\mathbf{s}'(\mathbf{r},\mathbf{r}')$, defined by $\mathbf{s}(\mathbf{r})\doteq N^{-1/2}\mathbf{r}$ and

$$\mathbf{s}'_n(\mathbf{r},\mathbf{r}') \doteq N^{-1/2}(\underline{\mathbf{T}}\mathbf{r}'_n + \mathbf{t}), \quad n = 1,\ldots,N, \tag{3}$$

namely,

$$R(\mathbf{r},\mathbf{r}') \doteq |\mathbf{s}'(\mathbf{r},\mathbf{r}') - \mathbf{s}(\mathbf{r})| = \left\{ N^{-1}\sum_{n=1}^{N} [\underline{\mathbf{T}}\mathbf{r}'_n + \mathbf{t} - \mathbf{r}_n]^2 \right\}^{1/2}. \tag{4}$$

Here, $\underline{\mathbf{T}}\doteq\underline{\mathbf{T}}(\mathbf{r},\mathbf{r}')$ and $\mathbf{t}\doteq\mathbf{t}(\mathbf{r},\mathbf{r}')$ denote, respectively, the three-dimensional rotation matrix (three degrees of freedom) and translation vector (three degrees of freedom) leading to the minimum value of R for the given structure pair. It can be shown that R satisfies all properties of a metric in the mathematical sense[62–64] [positivity: $R(\mathbf{r},\mathbf{r})=0$ and $R(\mathbf{r},\mathbf{r}')>0\,\forall\,\mathbf{r}'\neq\mathbf{r}$; symmetry: $R(\mathbf{r},\mathbf{r}')=R(\mathbf{r}',\mathbf{r})\,\forall\,\mathbf{r},\mathbf{r}'$; triangle inequality: $R(\mathbf{r},\mathbf{r}'')\leq R(\mathbf{r},\mathbf{r}')+R(\mathbf{r}',\mathbf{r}'')\,\forall\,\mathbf{r},\mathbf{r}',\mathbf{r}''$). A number of alternative (equivalent) procedures for determining $\underline{\mathbf{T}}$ and $\mathbf{t}$ from $\mathbf{r}$ and $\mathbf{r}'$ have been proposed in the literature.[51,52,65,66]

For a given reference structure $\mathbf{r}$, the M-dimensional hypersurface (within $\mathbb{R}^{3N}$) containing the vectors $\mathbf{s}'(\mathbf{r},\mathbf{r}')$ associated with all anchored walks $\mathbf{r}'$ of length N will be noted $\mathcal{R}_N(\mathbf{r})$. This hypersurface contains the $3N$-dimensional Cartesian coordinate vectors (amplified by $N^{-1/2}$) of all anchored walks after least-squares rototranslational fitting onto $\mathbf{r}$. Because $\mathcal{Q}_N$ is compact and bounded (with a finite volume $V_{\mathcal{Q}_N}$), for any $\mathbf{r}$, the hypersurface $\mathcal{R}_N(\mathbf{r})$ (within $\mathbb{R}^{3N}$) is also compact and bounded, with a finite area $A_{\mathcal{R}_N(\mathbf{r})}$.

The hypersurface $\mathcal{R}_N(\mathbf{r})$ associated with the R metric depends on the choice of the reference structure $\mathbf{r}$. However, if only distances between very close structures are of interest, it is possible to piece the $\mathcal{R}_N(\mathbf{r})$ hypersurfaces together into a single hypersurface $\tilde{\mathcal{R}}_N$ with a metric $\tilde{R}$ that is locally equivalent to R, i.e., satisfying

$$\lim_{R(\mathbf{r},\mathbf{r}')\rightarrow 0} [R(\mathbf{r},\mathbf{r}') - \tilde{R}(\mathbf{r},\mathbf{r}')] = 0. \tag{5}$$

This can be done by introducing a regular paving of $\mathcal{Q}_N$ using G grid cells centered at grid points $\{\mathbf{q}_k | k=1,\ldots,G\}$. The hypersurface $\tilde{\mathcal{R}}_N$ is then defined as

$$\tilde{\mathcal{R}}_N \doteq \lim_{G\rightarrow\infty} \cup_{k=1}^{G} \mathcal{R}_{N,k}, \tag{6}$$

where $\mathcal{R}_{N,k}$ denotes the portion of $\mathcal{R}_N(\mathbf{r}(\mathbf{q}_k))$ corresponding to all structures contained within the grid cell k. The metric $\tilde{R}$ in $\tilde{\mathcal{R}}_N$ is the scalar distance

$$\tilde{R}(\mathbf{r},\mathbf{r}') \doteq \lim_{G\rightarrow\infty} |\tilde{\mathbf{s}}'_G(\mathbf{r}') - \tilde{\mathbf{s}}_G(\mathbf{r})|, \tag{7}$$

where $\tilde{\mathbf{s}}_G(\mathbf{r})$ and $\tilde{\mathbf{s}}'_G(\mathbf{r}')$ represent the $3N$-dimensional Cartesian coordinate vectors (amplified by $N^{-1/2}$) of the two anchored walks after least-squares rototranslational fitting onto the structures associated with the two respective closest grid points (for a given G). Note that unlike $\mathbf{s}'$, $\tilde{\mathbf{s}}'_G$ only depends on $\mathbf{r}'$ (not on $\mathbf{r}$).

On a nonlocal level (i.e., when comparing structures at a finite distance R), $\tilde{\mathcal{R}}_N$ is essentially equivalent to $\mathcal{C}_N$ amplified by $N^{-1/2}$, and $\tilde{R}$ represents the RMSD-like distance between two anchored walks without any rototranslational fitting. However, at the local level, $\tilde{\mathcal{R}}_N$ is not equivalent to $\mathcal{C}_N$ scaled by $N^{-1/2}$. For any finite G, the patched hypersurface (union of the $\mathcal{R}_{N,k}$) is discontinuous at the grid-cell boundaries, and this discontinuity survives in $\tilde{\mathcal{R}}_N$ at the infinitesimal (local) level when taking the limit $G\rightarrow\infty$. It is easily seen that Eq. (5) holds provided that the limit $R(\mathbf{r},\mathbf{r}')\rightarrow 0$ in this equation is taken before the limit $G\rightarrow\infty$ in Eq. (7), i.e., provided that the distance between the two compared structures remains infinitesimal compared to the grid spacing, even when taking the latter toward zero.

Because $\mathcal{Q}_N$ is compact and bounded (with a finite volume $V_{\mathcal{Q}_N}$), the hypersurface $\tilde{\mathcal{R}}_N$ (within $\mathbb{R}^{3N}$) is also compact and bounded with a finite area $A_{\tilde{\mathcal{R}}_N}$. For the above-mentioned reasons, however, $A_{\tilde{\mathcal{R}}_N}$ is not equal to $N^{-1/2}A_{\mathcal{C}_N}$. The drawing in Fig. 1(c) could thus also apply to the relationship between $\mathcal{Q}_N$ and $\tilde{\mathcal{R}}_N$ (keeping in mind the peculiar local properties of the latter hypersurface).

The RMSD metric, as a measure of structural dissimilarity, satisfies the intuitively expected conditions that it is (i) independent of the (rigid-body) relative positioning and orientation of the two compared structures and (ii) unaffected by performing a mirror symmetry, a central inversion, or an atom renumbering on the two compared structures. Note, however, that since the present walks are oriented, the distance between a walk and its reverse walk is, in general, not zero. Although the RMSD metric is probably the most ap-

propriate one to match our visual intuition concerning structural difference, it is not unique. For example, a distance-matrix root-mean-square difference[36,56] could be more appropriate to match our expectations concerning structure-related energetical differences (for systems where the dominant interactions correlate with pairwise interatomic distances). In contrast, a RMSD without rototranslational fitting would represent a poor measure, in the intuitive sense, of the structural dissimilarity between two walks because the anchoring of the walks in the $\mathcal{C}_N$ space (performed here on the first three beads) is arbitrary. This would mean, in particular, that (i) differences in the first angles and dihedral angles along the walk will have more influence on the metric compared to corresponding differences at the end of the walk and (ii) the distance between two walks would not be equal to the distance between the two corresponding reverse walks. A root-mean-square difference in internal coordinates would also represent a poor metric for structure comparison, in particular, because (i) differences in the angles and dihedral angles would be equally weighed along the chain, although the central ones are intuitively expected to have more impact on the overall shape compared to the terminal ones and (ii) dihedral angles are periodic variables, so that the resulting measure would depend on the arbitrary choice of a reference interval for the dihedral angles.

C. Random-walk ensemble

The random-walk ensemble $\mathcal{W}_N$ is defined as an infinite ensemble of anchored walks of length N with a homogeneous (normalized) probability distribution p_N over $\mathcal{Q}_N$, i.e.,

$$p_N(\mathbf{q}) = V_{\mathcal{Q}_N}^{-1}, \quad \text{so that} \quad \int_{\mathcal{Q}_N} d^M\mathbf{q}\, p_N(\mathbf{q}) = 1. \tag{8}$$

It is easily seen that $\mathcal{W}_N$ can be generated by taking (an infinite number of) walks in $\mathcal{C}_N$ for which each successive step of length b is taken in a random (i.e., isotropically distributed) direction, keeping in mind the six constraints imposed to the Cartesian coordinate components of the first three beads (as all hikers know this is, however, a fairly inefficient way of walking). For the last statement to be true, it is essential that the first $N-2$ coordinates of $\mathbf{q}$ (corresponding to the bond angles) are defined as angle cosines and not as angles.

D. Low-resolution shape parameters

Although the internal coordinates are, by definition, distributed homogeneously in $\mathcal{W}_N$, this is not the case for the corresponding distributions of specific low-resolution shape parameters, which may evidence a significant extent of "apparent" structure and anisotropy[7–11] (induced by the corresponding coordinate transformation).

The low-resolution observables considered in the present work are the end-to-end distance, radius of gyration, radius of gyration tensor anisotropy factors, and number of contacts between bead pairs. For simplicity, these parameters are defined below in a reduced form, so as to enforce (i) independence of the value chosen for the bond length b and (ii) a possible convergence to a unique (i.e., size independent) distribution in the limit $N \to \infty$. Weighing by bead masses is also left out from the definitions (i.e., it is assumed that all bead masses are unity).

The reduced end-to-end distance r_E is defined as the distance between beads 1 and N, scaled by $(N-1)^{1/2}b$, i.e.,

$$r_E(\mathbf{r}) \doteq (N-1)^{-1/2} b^{-1} |\mathbf{r}_N - \mathbf{r}_1|. \tag{9}$$

The reduced radius of gyration r_G is defined as the root-mean-square distance between all beads and the center of geometry of the walk, scaled by $(N-1)^{1/2}b$, i.e.,

$$r_G(\mathbf{r}) \doteq (N-1)^{-1/2} b^{-1} \left\{ N^{-1} \sum_{n=1}^{N} \left[\mathbf{r}_n - \left(N^{-1} \sum_{m=1}^{N} \mathbf{r}_m \right) \right]^2 \right\}^{1/2}. \tag{10}$$

The radius of gyration tensor anisotropy factors $a_1 \geq a_2 \geq a_3$ are defined as

$$a_\mu(\mathbf{r}) \doteq \frac{I_\mu}{I_1 + I_2 + I_3}, \quad \mu = 1,2,3, \tag{11}$$

where $I_1 \geq I_2 \geq I_3$ are the eigenvalues of the radius of gyration tensor $\underline{\mathbf{I}}$, the components of which are

$$I_{\mu\nu}(\mathbf{r}) \doteq \sum_{n=1}^{N} \left[\mathbf{r}_{n,\mu} - \left(N^{-1} \sum_{m=1}^{N} \mathbf{r}_{m,\mu} \right) \right] \times \left[\mathbf{r}_{n,\nu} - \left(N^{-1} \sum_{m=1}^{N} \mathbf{r}_{m,\nu} \right) \right], \quad \mu,\nu = 1,2,3. \tag{12}$$

Note that

$$r_G(\mathbf{r}) = [N(N-1)]^{-1/2} b^{-1} (I_1 + I_2 + I_3)^{1/2}. \tag{13}$$

Finally, the reduced number of contacts c is defined here as the number of (unique, distinct, and nonconsecutive) bead pairs located at a distance smaller or equal to b, scaled by $(N-1)(N-2)/2$.

Analytical expressions for the distributions and mean values of the reduced end-to-end distance [$P_N^E(r_E)$ and $\langle r_E \rangle_N$] and for the root-mean-square values of the radius of gyration ($\langle r_G^2 \rangle_N^{1/2}$) over the N-beads (non-self-avoiding) random-walk ensembles $\mathcal{W}_N$ can easily be derived [see the Appendix, Eqs. (A15), (A17), and (A23), respectively]. They converge to N-independent expressions in the limit $N \to \infty$ [see the Appendix, Eqs. (A19), (A20), and (A24), respectively].

E. High-resolution shape parameters

The shape of a polymer chain associated with a given (central) structure is defined in the present work (Sec. I) as the collection of all structures for which the distance to this central structure [based on a specified metric, here the RMSD (Sec. II B)] is below a given cutoff value R_c. This definition implies, in particular, that (i) every structure can be used to define a corresponding shape (of which it is the central structure) and (ii) different shapes may be overlapping in terms of the structures they encompass (i.e., individual structures are not associated to a single shape). For simplicity, the central structure of a shape will be noted $\mathbf{q}$, i.e., as an internal coordinate vector of $\mathcal{Q}_N$ [with the corresponding Carte-

sian coordinate vector in $\mathcal{C}_N$ noted $\mathbf{r} \doteq \mathbf{r}(\mathbf{q})$], and the shape of which $\mathbf{q}$ is the central structure will be loosely referred to as the shape $\mathbf{q}$.

The central problem considered here is to determine how the homogeneous internal-coordinate probability distribution $p_N(\mathbf{q})$ [Eq. (8)] associated with individual structures in the random-walk ensemble $\mathcal{W}_N$ (Sec. II C) transforms to a corresponding shape probability distribution $P_N(\mathbf{q},R_c)$ associated with shapes (based on the RMSD metric R and for a given cutoff distance R_c). This shape probability distribution will be normalized to $V_{\mathcal{Q}_N}$ [Eq. (2)] rather than to unity, i.e.,

$$\int_{\mathcal{Q}_N} d^M\mathbf{q} P_N(\mathbf{q},R_c) = V_{\mathcal{Q}_N}. \tag{14}$$

This permits an immediate interpretation of $P_N(\mathbf{q},R_c)$ as the probability that an arbitrary random walk from $\mathcal{W}_N$ belongs to the specific shape $\mathbf{q}$, relative to the average of this probability over all possible shapes. For example, a value of 1.1 for $P_N(\mathbf{q},R_c)$ indicates that for the given cutoff R_c, shape $\mathbf{q}$ is 10% more likely to encompass an arbitrary random walk, compared to any shape of $\mathcal{Q}_N$ taken at random. The probability $P_N(\mathbf{q},R_c)$ can also be interpreted as the subvolume of $\mathcal{Q}_N$ spanned by the specific shape $\mathbf{q}$, relative to the average of this subvolume over all possible shapes. For example, a value of 1.1 for $P_N(\mathbf{q},R_c)$ also indicates that for the given cutoff R_c, the neighborhood of structure $\mathbf{q}$ spans a 10% larger subvolume of $\mathcal{Q}_N$ compared to the neighborhood of any shape of $\mathcal{Q}_N$ taken at random. P_N is thus a measure of the average density of random walks in the neighborhood of structure $\mathbf{q}$ (i.e., within the shape $\mathbf{q}$). When comparing two shapes, the ratio of the corresponding P_N values indicates accordingly how much more likely one of them is compared to the other. Finally, the chosen normalization implies that if all shapes were equiprobable, P_N would be uniformly one over $\mathcal{Q}_N$. In the following discussion, the cases of a finite versus an infinitesimal cutoff R_c are handled consecutively.

Consider first the case of a finite cutoff. For the RMSD metric, a given shape $\mathbf{q}$ with $\mathbf{r} \doteq \mathbf{r}(\mathbf{q})$ can be given a weight $\Omega_N(\mathbf{q},R_c)$ defined by the subvolume of $\mathcal{Q}_N$ mapping to the region of the hypersurface $\mathcal{R}_N(\mathbf{r})$ (Sec. II B) enclosed within a $3N$-dimensional hypersphere of radius R_c centered at $\mathbf{s}(\mathbf{r})$, i.e.,

$$\Omega_N(\mathbf{q},R_c) \doteq \int_{\mathcal{Q}_N} d^M\mathbf{q}'\Theta(R_c - |\mathbf{s}(\mathbf{r}(\mathbf{q}),\mathbf{r}'(\mathbf{q}')) - \mathbf{s}(\mathbf{r}(\mathbf{q}))|), \tag{15}$$

where Θ is the Heaviside function. The corresponding shape probability density [normalized as defined by Eq. (14)] may then be written as

$$P_N(\mathbf{q},R_c) \doteq \frac{V_{\mathcal{Q}_N}\Omega_N(\mathbf{q},R_c)}{\int_{\mathcal{Q}_N} d^M\mathbf{q}\Omega_N(\mathbf{q},R_c)}. \tag{16}$$

Another quantity of interest is the fractional coverage function $f_N(\mathbf{q},R_c)$, defined as the fraction of $\mathcal{Q}_N$ covered by $\Omega_N(\mathbf{q},R_c)$, i.e.,

$$f_N(\mathbf{q},R_c) \doteq V_{\mathcal{Q}_N}^{-1}\Omega_N(\mathbf{q},R_c). \tag{17}$$

For a given value of N, the function $f_N(\mathbf{q},R_c)$ is expected to present three regimes depending on the choice of R_c: (i) for R_c below some threshold value R_N^*, all shapes will only encompass part of $\mathcal{Q}_N$, i.e., $f_N<1$ for all $\mathbf{q}$; (ii) for R_c above some threshold value $R_N^{**}>R_N^*$, all shapes will encompass the entire extent of $\mathcal{Q}_N$, i.e., $f_N=1$ (and, consequently, $P_N=1$) for all $\mathbf{q}$; and (iii) for intermediate values $R_N^* \le R_c \le R_N^{**}$, a single shape ($R_c=R_N^*$), and then an increasingly large set of shapes ($R_c>R_N^*$), will extend over the entire $\mathcal{Q}_N$ (i.e., have $f_N=1$, the other shapes still being characterized by $f_N<1$). The single shape $\mathbf{q}_N^*$ (it can also be a few symmetry-related shapes) for which $f_N(\mathbf{q}_N^*,R_N^*)=1$ has a special meaning. It is the shape that can encompass the entire extent of its $\mathcal{R}_N(\mathbf{r})$ hypersurface for the smallest possible value of the cutoff distance. For this reason, $\mathbf{q}_N^*$ will be referred to as the "barycentric" shape of $\mathcal{Q}_N$. Note also that the second threshold R_N^{**} represents the maximum possible distance between any two walks of $\mathcal{Q}_N$ (i.e., the distance between the two most different walks). A more detailed analysis of the properties of R_N^* and R_N^{**} will be presented in a following article.[67]

Consider next the case of an infinitesimal cutoff, i.e., the limiting case $R_c \to 0$ (which will loosely be written as $R_c=0$). In this case, $P_N(\mathbf{q},0)$ probes the local density of random walks within a shape of infinitesimal size centered at $\mathbf{q}$. Because distances within an infinitesimal shape are infinitesimal and due to Eq. (5), it is possible to work here in the patched hypersurface $\tilde{\mathcal{R}}_N$ [Eq. (6)] rather than in the individual hypersurfaces $\mathcal{R}_N(\mathbf{r})$ (Sec. II B). For the RMSD metric, if an infinitesimal volume element $d^M\mathbf{q}$ around $\mathbf{q}$ in $\mathcal{Q}_N$ maps to a corresponding infinitesimal hypersurface element $d^M\tilde{\Sigma}_N$ of $\tilde{\mathcal{R}}_N$ around $\mathbf{s}(\mathbf{r})$, the shape can be given a weight $\Gamma_N(\mathbf{q})$ defined by

$$\Gamma_N(\mathbf{q}) \doteq \frac{d^M\mathbf{q}}{d^M\tilde{\Sigma}_N}. \tag{18}$$

Intuitively, for a given $d^M\mathbf{q}$ around a central structure $\mathbf{q}$, a large Γ_N (small $d^M\tilde{\Sigma}_N$) indicates that the random walks within $d^M\mathbf{q}$ are more densely packed in $\tilde{\mathcal{R}}_N$ around the central structure (i.e., that the corresponding shape is more probable), while a small Γ_N indicates that these walks are more widely spread (i.e., that this shape is less probable). This correspondence is schematically illustrated in Fig. 1(c). The ratio Γ_N represents the inverse of (the absolute value of) a Jacobian determinant of a special kind, which associates infinitesimal variations in a M-dimensional space ($\mathcal{Q}_N$) to corresponding variations in a $3N$-dimensional space ($\mathbb{R}^{3N}$) that are constrained to a M-dimensional hypersurface ($\tilde{\mathcal{R}}_N$). This Jacobian determinant is that of a $3N$-dimensional matrix containing in its first M lines the variations ds_α/dq_n with $n=1,\ldots,M$ and $\alpha=1,\ldots,3N$, and in its $3N-M=N+5$ last lines, the coefficients of a set of $3N$-dimensional unit vectors that are orthogonal to those in the first M lines as well as to each other. The corresponding local shape probability density [normalized as defined by Eq. (14)] may then be written as

$$P_N(\mathbf{q},0) \doteq I_N^{-1} V_{\mathcal{Q}_N}^{-1} \Gamma_N(\mathbf{q}), \tag{19}$$

where I_N is the average of Γ_N over all shapes of $\mathcal{Q}_N$ divided by the volume $V_{\mathcal{Q}_N}$, i.e.,

$$I_N \doteq V_{\mathcal{Q}_N}^{-2} \int_{\mathcal{Q}_N} d^M\mathbf{q} \Gamma_N(\mathbf{q}). \tag{20}$$

The single shape $\mathbf{q}_N^{\#}$ (it can also be a few symmetry-related shapes) that maximizes $P_N(\mathbf{q},0)$ over $\mathcal{Q}_N$ has a special meaning. It corresponds to the structure that has the highest density of random walks in its local neighborhood. For this reason, $\mathbf{q}_N^{\#}$ will be referred to as the "densest" shape of $\mathcal{W}_N$. Finally, based on Eq. (18), the area $A_{\tilde{\mathcal{R}}_N}$ of $\tilde{\mathcal{R}}_N$ can be evaluated as

$$A_{\tilde{\mathcal{R}}_N} = \int_{\mathcal{Q}_N} d^M\mathbf{q} \Gamma_N^{-1}(\mathbf{q}). \tag{21}$$

Note that the above approach is not applicable to walks containing one or more bond angles equal to 0 or π since these cannot be unambiguously represented in $\mathcal{Q}_N$. However, because they possess fewer than M degrees of freedom, it is easily seen that they are characterized by a vanishing local shape probability density $P_N(\mathbf{q},0)=0$.

It is important to realize that the weight Γ_N [$R_c \to 0$; Eq. (18)] differs from the weight Ω_N [$R_c \neq 0$; Eq. (15)] in that the former one is a local surface density on $\tilde{\mathcal{R}}_N$ (infinitesimal volume of $\mathcal{Q}_N$ divided by the associated infinitesimal area of $\tilde{\mathcal{R}}_N$) while the latter one is a volume in $\mathcal{Q}_N$ [finite subvolume of $\mathcal{Q}_N$ associated with a finite area of $\mathcal{R}_N(\mathbf{r})$], i.e., Γ_N is not the limit of Ω_N when $R_c \to 0$. An approximate relationship between the two quantities for small R_c values can be obtained by assuming that (i) the surface density of random walks is approximately constant over $\mathcal{R}_N(\mathbf{r}(\mathbf{q}))$ in the neighborhood of $\mathbf{q}$, (ii) this surface density is approximately equal to $\Gamma(\mathbf{q})$, and (iii) the effect of the curvature of $\mathcal{R}_N(\mathbf{r}(\mathbf{q}))$ in $\mathbb{R}^{3N}$ can be neglected. In this case, $\Omega_N(\mathbf{q},R_c)$ should be approximately equal (for a given $\mathbf{q}$) to $\Gamma_N(\mathbf{q})$ multiplied by the area of a M-dimensional hyperdisk of radius R_c, i.e.,

$$\Omega_N(\mathbf{q},R_c) \approx \pi^{M/2} R_c^M \gamma^{-1}(M/2+1) \Gamma_N(\mathbf{q}) \quad \text{for small } R_c, \tag{22}$$

where γ is the Euler gamma function, with $\gamma(M/2+1) = 2^{-(M+1)/2}\pi^{1/2} M!!$ for M odd (always the case here). This can be rewritten in terms of the fractional coverage function f_N [Eq. (17)] as [using Eq. (2) and for M odd]

$$f_N(\mathbf{q},R_c) \approx [2^{(M-1)/2} M!!]^{-1} \Gamma_N(\mathbf{q}) R_c^M \quad \text{for small } R_c. \tag{23}$$

Note that due to curvature effects, one should not expect this equation to be exactly satisfied even in the limit $R_c \to 0$ [i.e., in the sense of evaluating $\lim_{R_c \to 0} R_c^{-M} f_N(\mathbf{q},R_c)$].

To better appreciate the extent of inhomogeneity in the distribution of random walks into shapes, one may consider subvolumes of $\mathcal{Q}_N$ containing the most likely shapes based on a cutoff criterion P_c (in the range $[0,\ldots,P_N(\mathbf{q}_N^{\#},0)]$) applied to the local shape probability density $P_N(\mathbf{q},0)$. The fractional volume of such a region in terms of internal coordinates (relative to the entire accessible volume) is given by

$$v_N(P_c) \doteq V_{\mathcal{Q}_N}^{-1} \int_{\mathcal{Q}_N} d^M\mathbf{q} \Theta(P_N(\mathbf{q},0) - P_c). \tag{24}$$

This function varies between 0 ($P_c = P_N(\mathbf{q}_N^{\#},0)$) and 1 ($P_c = 0$) upon decreasing P_c. The integrated local shape probability over the same region relative to the corresponding integral $V_{\mathcal{Q}_N}$ over the entire $\mathcal{Q}_N$ is given by

$$P_{\text{in},N}(P_c) \doteq V_{\mathcal{Q}_N}^{-1} \int_{\mathcal{Q}_N} d^M\mathbf{q} \Theta(P_N(\mathbf{q},0) - P_c) P_N(\mathbf{q},0). \tag{25}$$

This function varies between 0 ($P_c = P_N(\mathbf{q}_N^{\#},0)$) and 1 ($P_c = 0$) upon decreasing P_c. Finally, the average of the local shape probability over this same region is given by

$$P_{\text{av},N}(P_c) \doteq v_N^{-1}(P_c) V_{\mathcal{Q}_N}^{-1} P_{\text{in},N}(P_c). \tag{26}$$

This function varies between $P_N(\mathbf{q}_N^{\#},0)$ [$P_c = P_N(\mathbf{q}_N^{\#},0)$; in a limiting sense] and 1 ($P_c = 0$) upon decreasing P_c. By comparing $P_{\text{in},N}(P_c)$ or $P_{\text{av},N}(P_c)$ to $v_N(P_c)$ for increasing values of P_c (i.e., by plotting $P_{\text{in},N}$ or $P_{\text{av},N}$ as a function of v_N), it is possible to assess the extent to which the most likely shapes dominate the less likely ones in the shape analysis of the $\mathcal{W}_N$ ensemble.

III. COMPUTATIONAL DETAILS

A systematic (grid-based) approach was used to sample the random-walk ensemble $\mathcal{W}_N$ for chain lengths ranging from $N=3$ to 10 beads. This approach involved the regular paving of $\mathcal{Q}_N$ using $G=g^M$ grid cells of volume $G^{-1}V_{\mathcal{Q}_N}$ centered at grid points $\{\mathbf{q}_k | k=1,\ldots,G\}$, with g being the number of cell subdivisions along one dimension (for simplicity, this number was chosen identical for all angle cosine as well as dihedral angle variables). Grid-based sampling is, in principle, the most appropriate method when sufficiently large g values are computationally affordable because it is deterministically reproducible and guarantees a rigorously homogeneous sampling throughout $\mathcal{Q}_N$. Note, however, that when used in combination with too small g values, it may introduce a systematic bias in the sampling (in which case a random sampling approach might be more adequate).

Apart from the number of beads N, the bond length b is the only free parameter in the considered random-walk ensembles. Because b has the dimension of a length (while N is dimensionless), all monitored properties scale in a predictable manner with b. This parameter was thus set to unity in all calculations for simplicity without affecting the generality of the results.

The probability distributions of the low-resolution shape parameters [r_E, r_G, a_1, a_2, a_3, and c, in reduced form (Sec. II D)] were monitored by sorting the corresponding values at the grid points into (normalized) occurrence histograms with a bin width of 0.005.

To evaluate the finite-cutoff shape probability density $P_N(\mathbf{q},R_c)$ (Sec. II E) at a given grid point $\mathbf{q}_k$, the volume $\Omega_N(\mathbf{q}_k,R_c)$ of Eq. (15) was estimated by the number of structures $\mathbf{q}_{k'}$ on the grid satisfying the involved cutoff condition

relative to $\mathbf{q}_k$, amplified by $G^{-1}V_{\mathcal{Q}_N}$. These estimates were then used to calculate the corresponding (gridded) finite-cutoff probability density $P_N(\mathbf{q}_k,R_c)$ via Eq. (16), where the integral in the denominator was replaced by a discrete sum over all grid points. The (gridded) fractional coverage function $f_N(\mathbf{q}_k,R_c)$ was evaluated similarly via Eq. (17). These two functions were computed for a discrete set of cutoff values R_c usually corresponding to $R_c/R_N^{**}=0.1$, 0.2, 0.4, and 0.6 (for $N=5$, the last two values were replaced by 0.25 and 0.45; for $N=6$, the first value was replaced by 0.15 and the last value omitted). The computational cost of the above calculation (one R calculation for each unique pair of distinct structures) is $\mathcal{O}[G(G-1)/2]\sim\mathcal{O}[g^{2M}]$, which is only tractable for reasonable values of g along with a rather small number of beads (e.g., $N\leq 6\rightarrow g^{2M}\leq g^{14}$). For this reason, the analysis using finite cutoff distances was not extended beyond 6 beads.

To evaluate the local shape probability density $P_N(\mathbf{q},0)$ (Sec. II E) at a given grid point $\mathbf{q}_k$, the surface density $\Gamma_N(\mathbf{q}_k)$ of Eq. (18) was estimated using a finite-difference approach. More precisely, for a grid point $\mathbf{q}_k$ (reference structure), the $\mathbf{q}$ vector was increased or decreased by half of the grid spacing along each of the M dimensions, resulting in $2M$ slightly altered structures (shifted structures). The reference and shifted structures were transformed to $\mathcal{C}_N$ and the latter ones rototranslationally fitted onto the former one. The corresponding $M\times 3N$ Cartesian displacements, divided by the grid spacing and scaled by $N^{1/2}$, provided finite-difference estimates for the elements of the first M rows of the Jacobian matrix (Sec. II E). The matrix was then completed by the $N+5$ orthogonal unit vectors (the construction of which required an equivalent number of matrix inversions). Finally, the (absolute value of the) inverse of this Jacobian determinant provided the required value for $\Gamma_N(\mathbf{q}_k)$. These estimates were then used to calculate the corresponding (gridded) local probability density $P_N(\mathbf{q}_k,0)$ via Eq. (19), where the integral involved in $I_N(\mathbf{q}_k)$ was replaced by a discrete sum over all grid points. The value of g employed for these calculations was chosen to be even, so that the reference walk (a grid-cell center) can never present angles equal to 0 or π. Whenever this situation occurred for a shifted walk, the corresponding angle cosine was simply set to ± 0.9999 instead of ± 1. This avoided the need of a special handling of this situation (the resulting error being essentially negligible). The computational cost of the above calculation (one Γ_N calculation for each structure) is $\mathcal{O}[G]=\mathcal{O}[g^M]$, which represents a more favorable scaling compared to the corresponding calculation at finite cutoff (see above), i.e., it remains tractable for reasonable values of g along with a larger number of beads (e.g., $N\leq 9\rightarrow g^M\leq g^{13}$). For this reason, the analysis using infinitesimal cutoffs was extended up to nine beads. In order to obtain more precise coordinates for the densest shape $\mathbf{q}_N^{\#}$ [the shapes maximizing $P_N(\mathbf{q},0)$], a grid-focusing approach was used, whereby the grid cell containing the best structure at a relatively low G value was subsequently rediscretized by a full grid of G points (iteratively if needed; this could be done reliably up to six beads only).

The computation of the finite-cutoff shape probability density was implemented into MATLAB 7.4 and carried out on an Apple Mac Pro (3 GHz Intel Xeon dual cores). The computation of the local shape probability density was implemented into a C program and carried out on Sun workstations. In both cases, the most expensive calculations ($n=9$ and $g=4$ for the local shape probability density or $n=6$ and $g=5$ for the finite-cutoff probability density) required about 3 weeks of computer time. Corresponding calculations with the same grid spacing but considering only one more bead in the walk would be prohibitively expensive (requiring about 1 or 40 years of computer time, respectively, on the same computers).

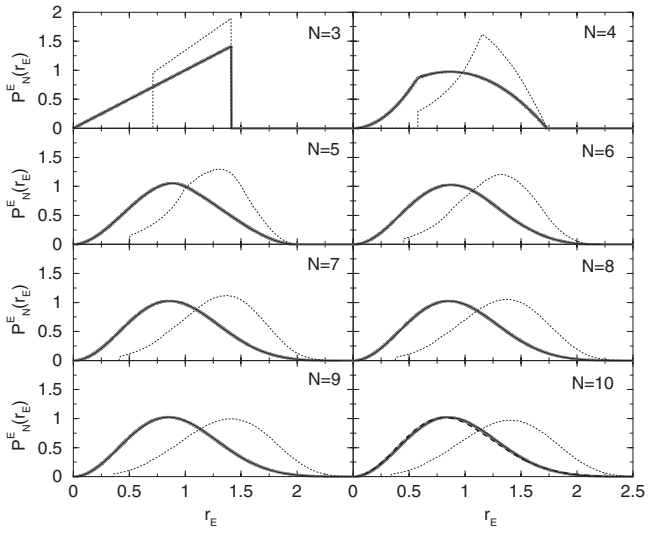

FIG. 2. Normalized probability distributions $P_N^E(r_E)$ of the reduced end-to-end distance r_E [Eq. (9)] calculated over the random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,10$. The data were evaluated using grid-based sampling in the internal-coordinate space (number of grid points per dimension set to $g=10^6$, 500, 40, 20, 10, 8, 6, and 4 for $N=3,\ldots,10$) and sorted into histograms (bin width 0.005). The solid and dotted lines refer to the ensembles of non-self-avoiding and self-avoiding random walks, respectively. The curves for the non-self-avoiding case are compared to corresponding analytical curves (bold gray lines) from Eq. (A15). The dashed line in the panel for $N=10$ represents the limiting analytical curve for $N\rightarrow\infty$ from Eq. (A19). The corresponding averages $\langle r_E\rangle_N$ are reported in Table I.

Unless otherwise specified, the calculations were performed as described above, considering non-self-avoiding random walks. However, a few additional calculations were also performed for self-avoiding walks (i.e., excluding any walk containing at least one pair of distinct and nonconsecutive beads at a distance smaller or equal to b). In this case, the sampling over angle-cosine coordinates was restricted to the range $[-1, 0.5]$ rather than $[-1, 1]$.

IV. RESULTS

A. Low-resolution shape parameters

The (normalized) probability distributions of the low-resolution shape parameters [r_E, r_G, a_1, a_2, a_3, and c, expressed in reduced form (Sec. II D)] over the random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,10$ are displayed in Figs. 2–5, while the corresponding averages reported in Table I, for both non-self-avoiding and self-avoiding walks. In the former case, the results for r_E and r_G are also compared to corresponding analytical expressions (see the Appendix).

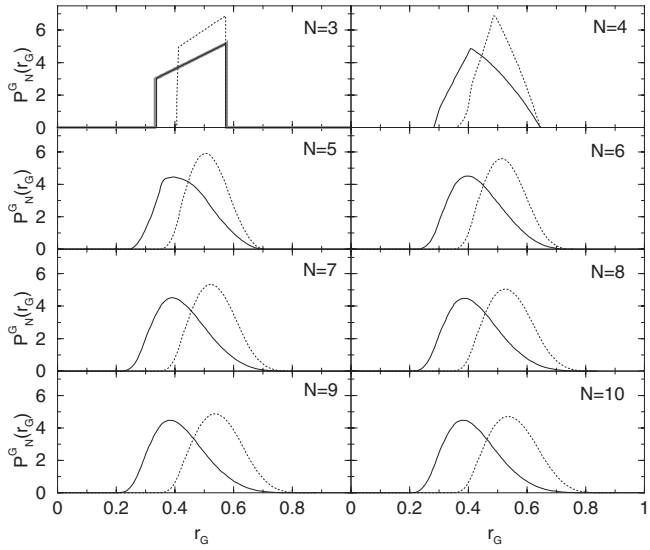

FIG. 3. Normalized probability distributions $P_N^G(r_G)$ of the reduced radius of gyration r_G [Eq. (10)] calculated over the random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,10$. The data were evaluated using grid-based sampling in the internal-coordinate space (number of grid points per dimension set to $g=10^6$, 500, 40, 20, 10, 8, 6, and 4 for $N=3,\ldots,10$) and sorted into histograms (bin width 0.005). The solid and dotted lines refer to the ensembles of non-self-avoiding and self-avoiding random walks, respectively. The curve for the non-self-avoiding case with $N=3$ is compared to the corresponding analytical curve (bold gray line) from Eq. (A30). The corresponding averages $\langle r_G\rangle_N$ and root-mean-square averages $\langle r_G^2\rangle_N^{1/2}$ are reported in Table I.

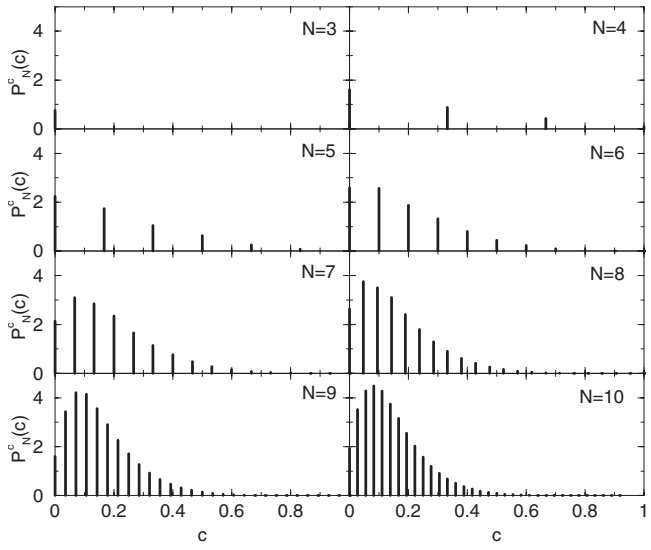

FIG. 5. Normalized probability distributions $P_N^c(c)$ of the reduced number of contacts c [number of unique, distinct, and nonconsecutive bead pairs located at a distance smaller or equal to b, scaled by $(N-1)(N-2)/2$ (Sec. II D)] calculated over the random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,10$. The data were evaluated using grid-based sampling in the internal-coordinate space (number of grid points per dimension set to $g=10^6$, 500, 40, 20, 10, 8, 6, and 4 for $N=3,\ldots,10$) for non-self-avoiding random walks. The corresponding distributions evaluated over the self-avoiding random-walk ensembles (not shown) are uninformative (single bar at $c=0$). Note that for ease of comparison the distributions are normalized as integrals, so that the sum of all bars is always equal to $(N-1)(N-2)/2$. The corresponding averages $\langle c\rangle_N$ are reported in Table I.

The distributions $P_N^E(r_E)$ of the reduced end-to-end distance r_E (Fig. 2) over the non-self-avoiding random-walk ensembles match very well the corresponding analytical curves [Eq. (A15)]. The resulting averages $\langle r_E\rangle_N$ (Table I) are also identical (within numerical accuracy) to the expected analytical values [Eq. (A17)]. These distributions are non-zero only within the interval $[0;(N-1)^{1/2}]$ because the end-to-end distance for a walk of length N can neither be negative nor exceed $(N-1)b$ and consist of piecewise-defined polynomials (involving $[N-O_N]/2+2$ successive definition intervals over $\mathbb{R}$, where O_N is one for N odd and zero otherwise). The resulting functions are either discontinuous ($N=3$), continuous to the first derivative ($N=4$), or continuous to the second derivative ($N\geq 5$). Upon increasing N, these distributions converge to an N-independent limiting form (Gaussian amplified by r_E^2 and truncated to the range $r_E\geq 0$ [Eq. (A19)]; see Fig. 2, panel for $N=10$), with a corresponding average value of $(3\pi/8)^{-1/2}\approx 0.9213$ [Eq. (A20)]. The convergence is relatively rapid, and these limiting expressions already represent good approximations for relatively low values of N. Although it may seem surprising at first sight that $\langle r_E\rangle_N$ decreases upon increasing N, it should be recalled that r_E is defined in a reduced form [scaled by $(N-1)^{1/2}b$]. The average of the (unscaled) end-to-end distance actually increases with increasing N, as expected.

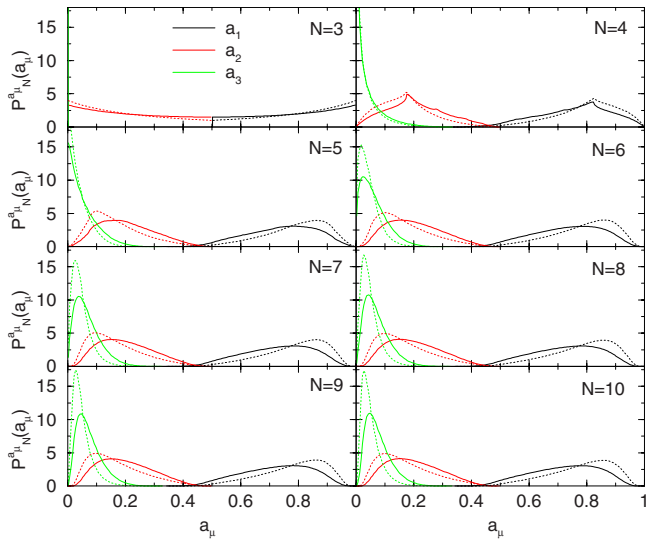

FIG. 4. (Color online) Normalized probability distribution $P_N^{a_\mu}(a_\mu)$ of the radius of gyration tensor anisotropy factors a_1, a_2, and a_3 [Eq. (11)] calculated over the random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,10$. The data were evaluated using grid-based sampling in the internal-coordinate space (number of grid points per dimension set to $g=10^6$, 500, 40, 20, 10, 8, 6, and 4 for $N=3,\ldots,10$) and sorted into histograms (bin width 0.005). The solid and dotted lines refer to the ensembles of non-self-avoiding and self-avoiding random walks, respectively. The corresponding averages $\langle a_\mu\rangle_N$ are reported in Table I.

The distributions $P_N^G(r_G)$ of the reduced radius of gyration r_G (Fig. 3) over the non-self-avoiding random-walk ensembles are nonzero only within the interval $[r_{G,N}^{\min};r_{G,N}^{\max}]$ [Eqs. (A25) and (A26)] and also appear to consist of piecewise-defined functions. The numerical distribution $P_3^G(r_G)$ for $N=3$ is in excellent agreement with the corresponding analytical curve [Eq. (A30); analytical forms could not be derived for $N\geq 4$]. The resulting root-mean-square averages $\langle r_G^2\rangle_N^{1/2}$ (Table I) are also identical (within numerical accuracy) to the expected analytical values [Eq. (A23)]. Upon increasing N, these root-mean-square averages con-

TABLE I. Average values of low-resolution shape parameters over the random-walk ensembles $\mathcal{W}_N$ with different numbers of beads $N=3,\ldots,10$. The parameters are the reduced end-to-end distance r_E [Eq. (9)], the reduced radius of gyration r_G [Eq. (10)], the radius of gyration tensor anisotropy factors $a_1 \geq a_2 \geq a_3$ [Eq. (11)], and the reduced number of contacts c [number of unique, distinct, and nonconsecutive bead pairs located at a distance smaller or equal to b, scaled by $(N-1)(N-2)/2$]. The data refer to the non-self-avoiding (top) and self-avoiding (bottom) random-walk ensembles. The values for $\langle r_E \rangle_N$ and $\langle r_G^2 \rangle_N^{1/2}$ over the non-self-avoiding random-walk ensembles were calculated from the analytical results of Eqs. (A17) and (A23), respectively (the corresponding numerical values were identical within numerical accuracy, i.e., with deviations of at most ± 0.001). The two corresponding $N \to \infty$ limits were calculated using Eqs. (A20) and (A24). All other values were calculated numerically using grid-based sampling in the internal-coordinate space (number of grid points per dimension set to $g=10^6$, 500, 40, 20, 10, 8, 6, and 4 for $N=3,\ldots,10$). The corresponding distributions are shown in Figs. 2–5.

N	$\langle r_E \rangle_N$	$\langle r_G \rangle_N$	$\langle r_G^2 \rangle_N^{1/2}$	$\langle a_1 \rangle_N$	$\langle a_2 \rangle_N$	$\langle a_3 \rangle_N$	$\langle c \rangle_N$
			Non-self-avoiding				
3	0.9428	0.4661	0.4714	0.7828	0.2166	0.0000	0.2500
4	0.9382	0.4497	0.4564	0.7555	0.2082	0.0441	0.2222
5	0.9333	0.4395	0.4472	0.7406	0.2049	0.0574	0.2001
6	0.9309	0.4326	0.4410	0.7329	0.2036	0.0647	0.1823
7	0.9293	0.4277	0.4364	0.7283	0.2029	0.0694	0.1815
8	0.9281	0.4239	0.4330	0.7256	0.2022	0.0725	0.1544
9	0.9272	0.4210	0.4303	0.7238	0.2017	0.0748	0.1536
10	0.9265	0.4188	0.4282	0.7221	0.2013	0.0770	0.1420
∞	0.9213	⋯	0.4082	⋯	⋯	⋯	⋯
			Self-avoiding				
3	1.0983	0.4980	0.5004	0.8053	0.1945	0.0000	0.0000
4	1.1716	0.5056	0.5088	0.7984	0.1718	0.0372	0.0000
5	1.2255	0.5151	0.5188	0.7920	0.1674	0.0435	0.0000
6	1.2681	0.5249	0.5291	0.7888	0.1666	0.0460	0.0000
7	1.3051	0.5349	0.5394	0.7870	0.1666	0.0472	0.0000
8	1.3237	0.5405	0.5455	0.7835	0.1680	0.0490	0.0000
9	1.3546	0.5504	0.5556	0.7827	0.1683	0.0495	0.0000
10	1.3638	0.5538	0.5594	0.7798	0.1697	0.0508	0.0000

verge to an N-independent limiting value of $6^{-1/2} \approx 0.4082$ [Eq. (A24)]. Although the convergence of the distributions toward a limiting form is less rapid than was the case for r_E, the limiting root-mean-square average still represents a good approximation for even relatively low values of N. Here again, $\langle r_G \rangle_N$ and $\langle r_G^2 \rangle_N^{1/2}$ decrease upon increasing N because r_G is defined in a reduced form [scaled by $(N-1)^{1/2}b$]. The average and root-mean-square average of the (unscaled) radius of gyration actually increase with increasing N, as expected.

The corresponding r_E and r_G distributions and averages evaluated over the self-avoiding random-walk ensembles (Figs. 2 and 3 and Table I) are systematically shifted (at identical N) toward higher values. This is due to the exclusion of walks presenting bead overlaps, which are on average more compact. In particular, the r_E distributions now vanish for $r_E < (N-1)^{-1/2}$ (because the first and last beads of a walk can no longer be closer than a distance b). In contrast to the non-self-avoiding case, the corresponding averages $\langle r_E \rangle_N$ and $\langle r_G^2 \rangle_N^{1/2}$ now increase upon increasing N. Furthermore, the corresponding dependence on N over the (limited) interval $N=3,\ldots,10$ suggests (without rigorously proving) the absence of convergence to a limiting value for large N. As a consequence, the scalings selected for r_E and r_G in Eqs. (9) and (10) may be inappropriate for the self-avoiding random-walk ensemble (in the sense of leading to an N-independent distribution in the limit $N \to \infty$).

The distributions $P_N^{a_\mu}(a_\mu)$ of the radius of gyration tensor anisotropy factors a_μ ($\mu=1,\ldots,3$; Fig. 4) over the non-self-avoiding random-walk ensembles clearly evidence the overall anisotropy effect revealed previously.[11–18] The a_1 and a_2 distributions are nearly nonoverlapping, while the a_2 and a_3 distributions, though partially overlapping, remain clearly distinct. This indicates that most non-self-avoiding random walks present an anisotropic overall shape, with an average that can be characterized as a prolate ellipsoid ("flattened-out cigar"). The corresponding average values of a_1, a_2, and a_3 for large N (e.g., about 0.72, 0.20, and 0.08 for $N=10$) are in a ratio of about 9.4:2.6:1. This ratio is already very close to the corresponding ratio of about 9:2.3:1 suggested previously[11] for the limit $N \to \infty$. The comparison of these distributions with the corresponding distributions evaluated over the self-avoiding random-walk ensembles reveal a significant anisotropy enhancement (at identical N), especially for a_1 versus a_2 and a_3. As a result, the average values of a_1, a_2, and a_3 for large N (e.g., about 0.78, 0.17, and 0.05 for $N=10$) are now in a ratio of about 15.3:3.3:1. Visual inspection suggests that the distributions of the anisotropy factors probably converge to unique functions in the limit of large N in the non-self-avoiding as well as in the self-avoiding case.

Finally, the distributions $P_N^c(c)$ of the reduced number of contacts c (Fig. 5) over the non-self-avoiding random-walk ensembles suggest that most such walks do not present a compact overall shape. The reduced number of contacts is,

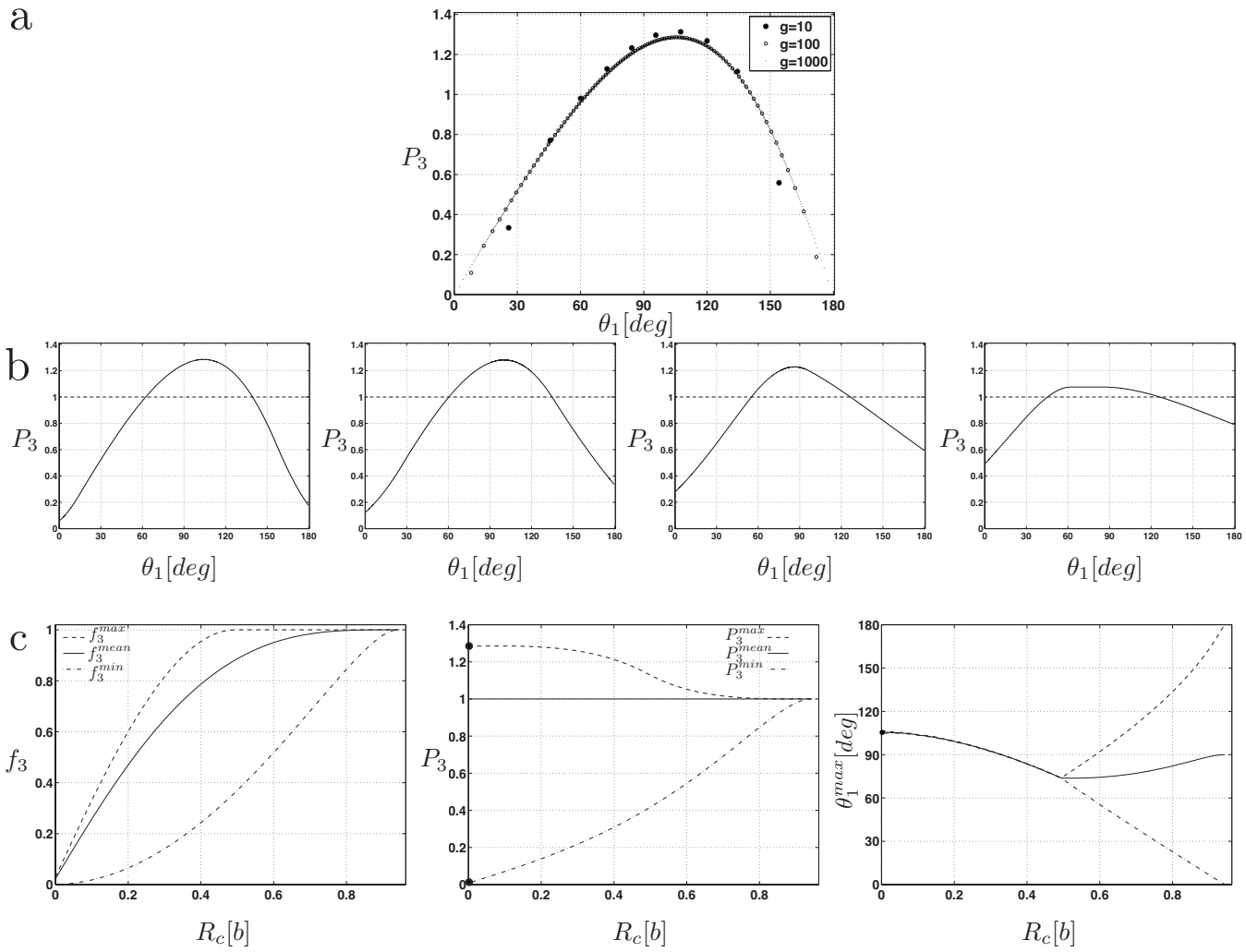

FIG. 6. (a) Normalized local shape probability distribution $P_3(\mathbf{q},0)$ [Eq. (19)], where $\mathbf{q}=\{\cos\theta_1\}$, displayed as a function of the single angle θ_1 of the walk. (b) Corresponding normalized finite-cutoff shape probability distribution $P_3(\mathbf{q},R_c)$ [Eq. (16)] displayed as a function of the single angle θ_1 of the walk; left to right: $R_c=0.09b$, $0.19b$, $0.38b$, and $0.57b$, with b being the bond length. (c) Left: maximum ($f_3^{\max}$), mean (f_3^{mean}), and minimum ($f_3^{\min}$) values of the fractional coverage function $f_3(\mathbf{q},R_c)$ [Eq. (17)] over $\mathcal{Q}_3$ displayed as a function of the cutoff R_c. Middle: maximum ($P_3^{\max}$), mean (P_3^{mean}), and minimum ($P_3^{\min}$) values of $P_3(\mathbf{q},R_c)$ over $\mathcal{Q}_3$, displayed as a function of the cutoff R_c. Right: maximum, mean, and minimum values of the angle θ_1 over the set of structure maximizing $P_3(\mathbf{q},R_c)$, displayed as a function of the cutoff R_c; points indicated at $R_c=0$ in the middle and right panels correspond to expected values based on the local probability analysis. Note that P_3 in (a) and (b) is normalized in terms of $\cos\theta_1$ (not θ_1) so that its average over the graph differs from one. The data in (a) were evaluated using three different numbers of grid points $g=10$, 100, or 1000. The data in (b) and (c) were evaluated using $g=10^5$ grid points.

on the average, relatively low ($c\leq 0.25$) compared to the theoretically possible maximum number of contacts (all beads in contact; $c=1$). The extent of compactness also appears to decrease upon increasing N, the average value $\langle c\rangle_N$ decreasing from 0.25 to 0.14 over the interval $N=3,\ldots,10$. Note that the corresponding distributions evaluated over the self-avoiding random-walk ensembles (not shown) are uninformative (single peak at $c=0$; contacts are by definition not allowed in this ensemble).

B. High-resolution shape parameters

The present discussion of high-resolution shape parameters is almost exclusively restricted to non-self-avoiding walks. The case of self-avoiding walks is briefly discussed at the end of the section.

The results for $N=3$ beads are displayed in Fig. 6, the corresponding key parameters being reported in Table II. Here, the internal coordinate vector $\mathbf{q}$ consists of the cosine of the single angle θ_1 defined by the three beads.

The local shape probability density $P_3(\mathbf{q},0)$ is displayed in Fig. 6(a) as a function of θ_1 for three different grid spacings g (10, 100, and 1000). The results for the three g values are essentially consistent, the curves corresponding to $g=100$ and 1000 being nearly indistinguishable (indicating a sufficient accuracy of the finite-difference approximation to the Jacobian). As expected, the shapes centered at $\theta_1=0,\pi$ are characterized by a vanishing probability. The distribution shows a single maximum and is slightly biased toward open angles ($>90°$). The maximum (densest shape $\mathbf{q}_3^{\#}$) is located at $\theta_1^{\#}=105.5°$ and associated with a local shape probability density $P_3(\mathbf{q}_3^{\#},0)=1.28$. This indicates that the local neighborhood of this central structure is 28% more populated (in terms of random-walk density) compared to the corresponding average over all possible shapes or, equivalently, that the corresponding shape is 28% more likely (in a local sense)

TABLE II. Parameters characterizing the local shape probability density $P_N(\mathbf{q},0)$ (top) and the finite-cutoff shape probability density $P_N(\mathbf{q},R_c)$ (bottom) for the random-walk ensembles $\mathcal{W}_N$ with different numbers of beads $N=3,\ldots,9$. The parameters are the central structure $\mathbf{q}_N^{\#}$ of the densest shape (internal coordinates $\{\theta^{\#}\}$ and $\{\omega^{\#}\}$), the local probability density $P_N(\mathbf{q}_N^{\#},0)$ associated with this shape [Eq. (19)], the area $A_{\tilde{\mathcal{R}}_N}$ of the $\tilde{\mathcal{R}}_N$ hypersurface [Eq. (21)], the average I_N of the random-walk local surface density Γ_N divided by the volume $V_{\mathcal{Q}_N}$ [Eq. (20)], the central structure $\mathbf{q}_N^{*}$ of the barycentric shape (internal coordinates $\{\theta^{*}\}$ and $\{\omega^{*}\}$), the smallest cutoff radius R_N^{*} for which this shape encompasses the entire ensemble, the finite-cutoff probability density $P_N(\mathbf{q}_N^{*},R_N^{*})$ associated with these shape and cutoff [Eq. (16)], and the cutoff radius R_N^{**} above which all shapes encompass the entire ensemble (maximum distance between any two walks of $\mathcal{Q}_N$). Although the data mainly refer to the ensembles of non-self-avoiding random walks, the values of P_N, $A_{\tilde{\mathcal{R}}_N}$, and I_N for the self-avoiding random-walk ensemble are reported between parentheses [the $\mathbf{q}_N^{\#}$ structures are self-avoiding and thus identical for the two ensembles; this possibly also holds for the $\mathbf{q}_N^{*}$ structures and R_N^{*} values, but not for the R_N^{**} values (Ref. 67)]. The data were evaluated using grid-based sampling in the internal-coordinate space. The number of grid points per dimension was set to $g=10^4$, 200, 24, 12, 8, 4, and 4 for $N=3,\ldots,9$ (local density) or $g=10^5$, 51, 13, and 5 for $N=3,\ldots,6$ (finite-cutoff density). The parameters of $\mathbf{q}_N^{\#}$ for $N=3,\ldots,6$ were further refined to a precision of at least 1° using grid focusing.

Infinitesimal cutoff

N	$\{\theta^{\#}\}$ (deg)	$\{\omega^{\#}\}$ (deg)	$P_N(\mathbf{q}_N^{\#},0)$	$A_{\tilde{\mathcal{R}}_N}[b^M]$	$I_N[b^{-M}]$
3	105.5	⋯	1.28 (1.15)	1.02(0.82)	1.21(1.35)
4	137.7/137.7	0.0	1.79 (1.50)	0.77(0.52)	1.51(1.80)
5	154.5/143.7/154.5	0.0/0.0	2.94 (2.34)	0.71(0.38)	1.74(2.18)
6	163.5/151.8/151.8/163.5	0.0/0.0/0.0	10.05 (7.60)	0.63(0.26)	2.07(2.73)
7	⋯	⋯	⋯	0.52(0.17)	2.64(3.75)
8	⋯	⋯	⋯	0.43(0.09)	3.42(5.89)
9	⋯	⋯	⋯	0.27(0.05)	5.55(9.86)

Finite cutoff

N	$\{\theta^{*}\}$ (deg)	$\{\omega^{*}\}$ (deg)	$P_N(\mathbf{q}_N^{*},R_N^{*})$	$R_N^{*}[b]$	$R_N^{**}[b]$
3	74.2	⋯	1.13	0.486	0.943
4[a]	36.0/36.0	±162.4	1.11	0.712	1.236
5[b]	81/67/81	±166/∓83, ∓83/±166	1.07	0.873	1.600
6[c]	66/78/78/66	±144/(+72,−72)/∓144	1.06	0.998	1.899

[a]Two symmetry-related barycentric shapes (enantiomers $\omega_1^{*}\leftrightarrow-\omega_1^{*}$).
[b]Precisions of reported θ^{*} and ω^{*} parameters are only about 20 and 30°, respectively; four symmetry-related barycentric shapes (enantiomers $\omega_1^{*},\omega_2^{*}\leftrightarrow-\omega_1^{*},-\omega_2^{*}$ and bead-order inversion $\omega_1^{*}\leftrightarrow\omega_2^{*}$).
[c]Precisions of reported θ^{*} and ω^{*} parameters are only about 30 and 70°, respectively; two alternative ω_2^{*} values, each with four symmetry-related barycentric shapes (enantiomers $\omega_1^{*},\omega_2^{*},\omega_3^{*}\leftrightarrow-\omega_1^{*},-\omega_2^{*},-\omega_3^{*}$ and bead-order inversion $\omega_1^{*}\leftrightarrow\omega_3^{*}$).

than any shape taken at random. This difference may seem modest because it expresses a bias relative to the average over all shapes (i.e., including the most likely ones). However, pairwise comparisons can show more dramatic effects. For example, the densest shape is about 15 times more likely compared to the one centered at $\theta_1=5°$.

The finite-cutoff shape probability density $P_3(\mathbf{q},R_c)$ is displayed in Fig. 6(b) as a function of θ_1 for four different cutoff values R_c. As expected, the curve corresponding to the lowest R_c value is the closest to the limiting case $R_c\to0$ [Fig. 6(a)]. Upon increasing R_c, the bias in the distribution and the location of the maximum progressively shift in the direction of closed angles (<90°). For the largest R_c value considered here [$R_c=0.57b>R_3^{*}$ (see below)], the maximum no longer corresponds to a single θ_1 value, but to a range thereof. This arises because for such a large cutoff, a collection of shapes is now able to encompass the entire random-walk ensemble.

The dependence of $P_3(\mathbf{q},R_c)$ on the cutoff value is characterized in more details in Fig. 6(c). The left panel illustrates the maximum $f_3^{\max}(R_c)$, mean $f_3^{\mathrm{mean}}(R_c)$, and minimum $f_3^{\min}(R_c)$ values (over shapes centered at all points of $\mathcal{Q}_3$) of the fractional coverage function $f_3(\mathbf{q},R_c)$, displayed as a function of R_c. The central panel displays the associated maximum $P_3^{\max}(R_c)$, mean $P_3^{\mathrm{mean}}(R_c)$, and minimum $P_3^{\min}(R_c)$ values of the finite-cutoff shape probability density $P_3(\mathbf{q},R_c)$. Finally, the right panel displays the range of θ_1 values associated with the central structures of the most likely shapes (i.e., those corresponding to $f_3^{\max}$ and $P_3^{\max}$), as well as the average value of θ_1 over this set. The function $f_3^{\max}$, i.e., the fraction of the random-walk ensemble encompassed by the most likely shape for a given R_c, increases from zero at $R_c=0$ (infinitesimal shape) to 1 (the most probable shapes encompass the entire ensemble). This function reaches 1 at a cutoff value $R_3^{*}=0.486b$ for a specific shape (barycentric shape $\mathbf{q}_3^{*}$) characterized by $\theta_1^{*}=74.2°$ and $P_3(\mathbf{q}_3^{*},R_3^{*})=1.13$. This shape is the one that encompasses the entire ensemble for the smallest possible value of R_c. As could be anticipated from Fig. 6(b), the θ_1 angle associated with the most likely shape decreases upon increasing R_c from 0 to R_3^{*}. Over a sizable range of R_c values ($0\le R_c\le0.4b$), this most likely shape is consistently more probable by about 20% ($P_3^{\max}$) than any shape taken at random. The function

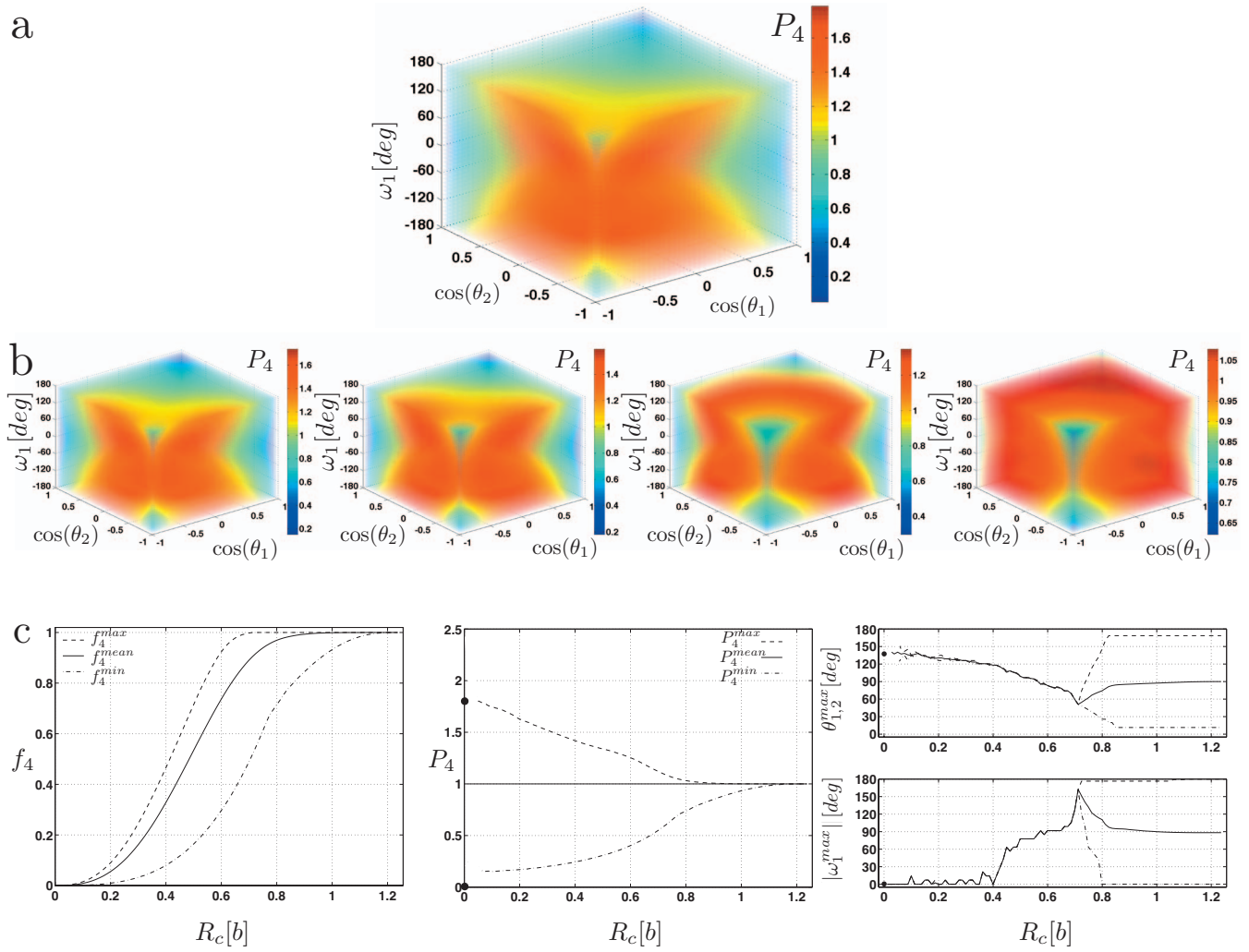

FIG. 7. (Color) (a) Normalized local shape probability distribution $P_4(\mathbf{q},0)$ [Eq. (19)], where $\mathbf{q}=\{\cos\theta_1,\cos\theta_2,\omega_1\}$, displayed as a function of the angle cosines $\cos\theta_1$ and $\cos\theta_2$ and the single dihedral angle ω_1 of the walk. (b) Corresponding normalized finite-cutoff shape probability distribution $P_4(\mathbf{q},R_c)$ [Eq. (16)] displayed as a function of the angle cosines $\cos\theta_1$ and $\cos\theta_2$ and the single dihedral angle ω_1 of the walk; left to right: $R_c=0.12b$, $0.25b$, $0.50b$, and $0.74b$, with b being the bond length. (c) Left: maximum ($f_4^{\max}$), mean (f_4^{mean}), and minimum ($f_4^{\min}$) values of the fractional coverage function $f_4(\mathbf{q},R_c)$ [Eq. (17)] over $\mathcal{Q}_4$, displayed as a function of the cutoff R_c. Middle: maximum ($P_4^{\max}$), mean (P_4^{mean}), and minimum ($P_4^{\min}$) values of $P_4(\mathbf{q},R_c)$ over $\mathcal{Q}_4$, displayed as a function of the cutoff R_c. Right: maximum, mean, and minimum values of the angles θ_1 and θ_2 as well as of the dihedral angle ω_1 (absolute value) over the set of structure maximizing $P_4(\mathbf{q},R_c)$, displayed as a function of the cutoff R_c; points indicated at $R_c=0$ in the middle and right panels correspond to expected values based on the local probability analysis. The data in (a) were evaluated using $g=200$ grid points per dimension, and the data in (b) and (c) using $g=51$ grid points per dimension.

$f_3^{\min}$, i.e., the fraction of the random-walk ensemble encompassed by the least probable shape for a given R_c, also increases from zero at $R_c=0$ to 1. This function reaches 1 at a cutoff value $R_3^{**}=0.943b$. Above this R_c value, all shapes encompass the entire ensemble. As expected, the range of θ_1 values satisfying $f_3(\mathbf{q},R_c)=1$ widens upon increasing R_c from R_3^{*} to R_3^{**}, while the average θ_1 value over this set slightly increases over this interval.

As a final note concerning the above results for $N=3$, it is important to stress that although three beads are always contained in a plane, the present results pertain to random walks in three dimensions. In this case, the probability distribution $p_3(\mathbf{q})$ in the random-walk ensemble [Eq. (8)] is homogeneous in $\cos\theta_1$ (with θ_1 in the range $[0;\pi]$), so that the corresponding average $\cos\theta_1$ value is 0. This is in qualitative agreement with the observation that the most likely shapes have θ_1 angles close to 90°. In two dimensions, however, the corresponding probability distribution would be homogeneous in θ_1 so that the corresponding average θ_1 value would be 0° (if θ_1 is chosen in the range $]-\pi;\pi]$). The results in terms of shape probability distributions would then look quite different.

The results for $N=4$ beads are displayed in Fig. 7, the corresponding key parameters being reported in Table II. Here, the internal coordinate vector $\mathbf{q}$ consists of the cosines of the two angles θ_1 and θ_2 along with the single dihedral angle ω_1 defined by the four beads.

The local shape probability density $P_4(\mathbf{q},0)$ is displayed in Fig. 7(a) as a function of $\cos\theta_1$, $\cos\theta_2$, and ω_1. As expected, the shapes centered at $\theta_1=0,\pi$ or $\theta_2=0,\pi$ are characterized by a vanishing probability. In addition, due to the symmetry properties of the RMSD metric (Sec. II B), the distribution is invariant with respect to the changes $\omega_1\leftrightarrow-\omega_1$ and $\theta_1\leftrightarrow\theta_2$. This distribution displays a single maxi-

mum and is significantly biased toward open θ_1 and θ_2 angles. The maximum (densest shape $\mathbf{q}_4^{\#}$) is located at $\theta_1^{\#}=\theta_2^{\#}=137.7°$ and $\omega_1^{\#}=0.0°$ and is associated with a local shape probability density $P_4(\mathbf{q}_4^{\#},0)=1.79$. Note that the presence of a single maximum is not a consequence of the above-mentioned symmetry properties (these merely imply that if the maximum is unique, it must satisfy $\theta_1^{\#}=\theta_2^{\#}$ and $\omega_1^{\#}=0°$). This specific shape is 79% more likely (in a local sense) than any shape taken at random. Here too, the bias may be much more dramatic when performing pairwise comparisons between shapes. For example, the densest shape is about 35 times more likely compared to the one centered at $\theta_1=\theta_2=5°$ and $\omega_1=0°$.

The finite-cutoff shape probability density $P_4(\mathbf{q},R_c)$ is displayed in Fig. 7(b) as a function of cos θ_1, cos θ_2, and ω_1 for four different cutoff values R_c. All graphs preserve the symmetry features described above for $P_4(\mathbf{q},0)$. The curve corresponding to the lowest R_c value is again the closest to the limiting case $R_c\to 0$ [Fig. 7(a)]. Upon increasing R_c, the bias in the distribution and the location of the maximum progressively shift in the direction of closed angles. For $R_c>0.4b$, the single maximum becomes split into two symmetry-related (enantiomeric) maxima with opposite $\omega_1\neq 0$ values. For the largest R_c value considered ($R_c=0.74b>R_4^*$, see below), these two maxima no longer correspond to single points but to two regions of the graph.

The dependence of $P_4(\mathbf{q},R_c)$ on the cutoff value is characterized in more detail in Fig. 7(c), analogous to Fig. 6(c) for $N=3$ (see explanations above). These curves display the same qualitative features as for $N=3$, although the numerical precision is considerably lower (especially for low R_c values) due to the more limited grid resolution. The fractional coverage function $f_4^{\max}$ reaches 1 at a cutoff value $R_4^*=0.712b$ for a specific (symmetry duplicated, i.e., two enantiomers) shape (barycentric shape $\mathbf{q}_4^*$) characterized by $\theta_1^*=\theta_2^*=36.0°$ and $\omega_1^*=\pm 162.4°$, and $P_4(\mathbf{q}_4^*,R_4^*)=1.11$. As could be anticipated from Fig. 7(b), the θ_1 and θ_2 values associated with the most likely shape (which remain identical to each other) decrease upon increasing R_c from 0 to R_4^*, while the corresponding single ω_1 value of 0° splits into two opposite (and increasingly larger) values for $R_c>0.4b$. Over a sizable range of R_c values ($0\leq R_c\leq 0.4b$), this most likely shape is consistently more probable by at least 40% ($P_4^{\max}$) than any shape taken at random. The function $f_4^{\min}$ reaches 1 at a cutoff value $R_4^{**}=1.236$. Above this R_c value, all shapes encompass the entire ensemble. As expected, the ranges of θ_1, θ_2, and ω_1 values satisfying $f_4(\mathbf{q},R_c)=1$ widen upon increasing R_c from R_4^* to R_4^{**}.

The results for $N=5$ beads are displayed in Fig. 8, the corresponding key parameters being reported in Table II. Here, the internal coordinate vector $\mathbf{q}$ consists of the cosines of the three angles θ_1, θ_2, and θ_3 along with the two dihedral angles ω_1 and ω_2 defined by the five beads.

The local shape probability density $P_5(\mathbf{q},0)$ is displayed in Fig. 8(a) in the form of a maximal value (over all possible θ_1, θ_2, and θ_3 combinations) as a function of the ω_1 and ω_2 dihedral angles. It was verified that the full (five-dimensional) distribution (not shown) satisfies the expected symmetry properties (Sec. II B). These translate at the level of the two-dimensional maximal-value projection to invariances with respect to the changes $\omega_1\leftrightarrow\omega_2$ and $\omega_1,\omega_2\leftrightarrow-\omega_1,-\omega_2$. The distribution displays a single maximum and is significantly biased toward flat *cis-cis* structures. The maximum (densest shape $\mathbf{q}_5^{\#}$) is located at $\theta_1^{\#}=\theta_3^{\#}=154.5°$, $\theta_2^{\#}=143.7°$, and $\omega_1^{\#}=\omega_2^{\#}=0.0°$ and is associated with a local shape probability density $P_5(\mathbf{q}_5^{\#},0)=2.94$. Here too, the presence of a single maximum is not a consequence of the above-mentioned symmetry properties (these merely imply that if the maximum is unique, it must satisfy $\theta_1^{\#}=\theta_3^{\#}$ and $\omega_1^{\#}=\omega_2^{\#}=0°$). This specific shape is about three times more likely (in a local sense) than any shape taken at random.

The finite-cutoff shape probability density $P_5(\mathbf{q},R_c)$ is displayed in Fig. 8(b), also in the form of a maximal-value projection, for four different cutoff values. All graphs preserve the symmetry features described above for $P_5(\mathbf{q},0)$. The curve corresponding to the lowest R_c value is again the closest to the limiting case $R_c\to 0$ [Fig. 8(a)]. Upon increasing R_c, the bias in the distribution progressively shifts from flat *cis-cis* structures in the direction of *gauche*$^{\pm}$-*gauche*$^{\mp}$ structures, while the single maximum becomes split into four symmetry-related (enantiomeric forms and reverse bead order) maxima with $\omega_1,\omega_2\neq 0$ values.

The dependence of $P_5(\mathbf{q},R_c)$ on the cutoff value is characterized in more detail in Fig. 8(c), analogous to Figs. 6(c) and 7(c) for $N=3,4$ (see explanations above). These curves display the same qualitative features as for $N=3$ and 4, although the numerical precision is again considerably lower (especially for low R_c values). The fractional coverage function $f_5^{\max}$ reaches 1 at a cutoff value $R_5^*=0.873b$ for a specific shape (barycentric shape $\mathbf{q}_5^*$) that is fourfold replicated by symmetry and is associated with $P_5(\mathbf{q}_5^*,R_5^*)=1.07$. One of these structures corresponds to $\theta_1^*=\theta_3^*=81°$, $\theta_2^*=67°$, $\omega_1^*=166°$, and $\omega_2^*=-83°$. The other three are obtained by the changes $\theta_1^*,\omega_1^*\leftrightarrow\theta_3^*,\omega_2^*$ and/or $\omega_1^*,\omega_2^*\leftrightarrow-\omega_1^*,-\omega_2^*$. As could be anticipated from Fig. 8(b), the dihedral angles associated with the most likely shape tend to shift away from flat *cis-cis* upon increasing R_c from 0 to R_5^*. Simultaneously, the angles tend to shift toward lower values. The function $f_5^{\min}$ reaches 1 at a cutoff value of $R_5^{**}=1.600$. Above this R_c value, all shapes encompass the entire ensemble.

The results for $N=6$ beads are displayed in Fig. 9, the corresponding key parameters being reported in Table II. Here, the internal coordinate vector $\mathbf{q}$ consists of the cosines of the four angles θ_1, θ_2, θ_3, and θ_4 along with the three dihedral angles ω_1, ω_2, and ω_3 defined by the six beads.

The local shape probability density $P_6(\mathbf{q},0)$ is displayed in Fig. 9(a) in the form of the maximal value (over all possible θ_1, θ_2, θ_3, and θ_4 combinations) as a function of the ω_1, ω_2, and ω_3 dihedral angles. It was verified that the full (seven-dimensional) distribution (not shown) satisfies the expected symmetry properties (Sec. II B). These translate at the level of the three-dimensional maximal-value projection to invariances with respect to the changes $\omega_1\leftrightarrow\omega_3$ and $\omega_1,\omega_2,\omega_3\leftrightarrow-\omega_1,-\omega_2,-\omega_3$. The distribution displays a single maximum and is significantly biased toward flat *cis-cis-cis* structures. The maximum (densest shape $\mathbf{q}_6^{\#}$) is located at $\theta_1^{\#}=\theta_4^{\#}=163.5°$, $\theta_2^{\#}=\theta_3^{\#}=151.8°$, and $\omega_1^{\#}=\omega_2^{\#}=\omega_3^{\#}=0.0°$ and is associated with a local shape probability density

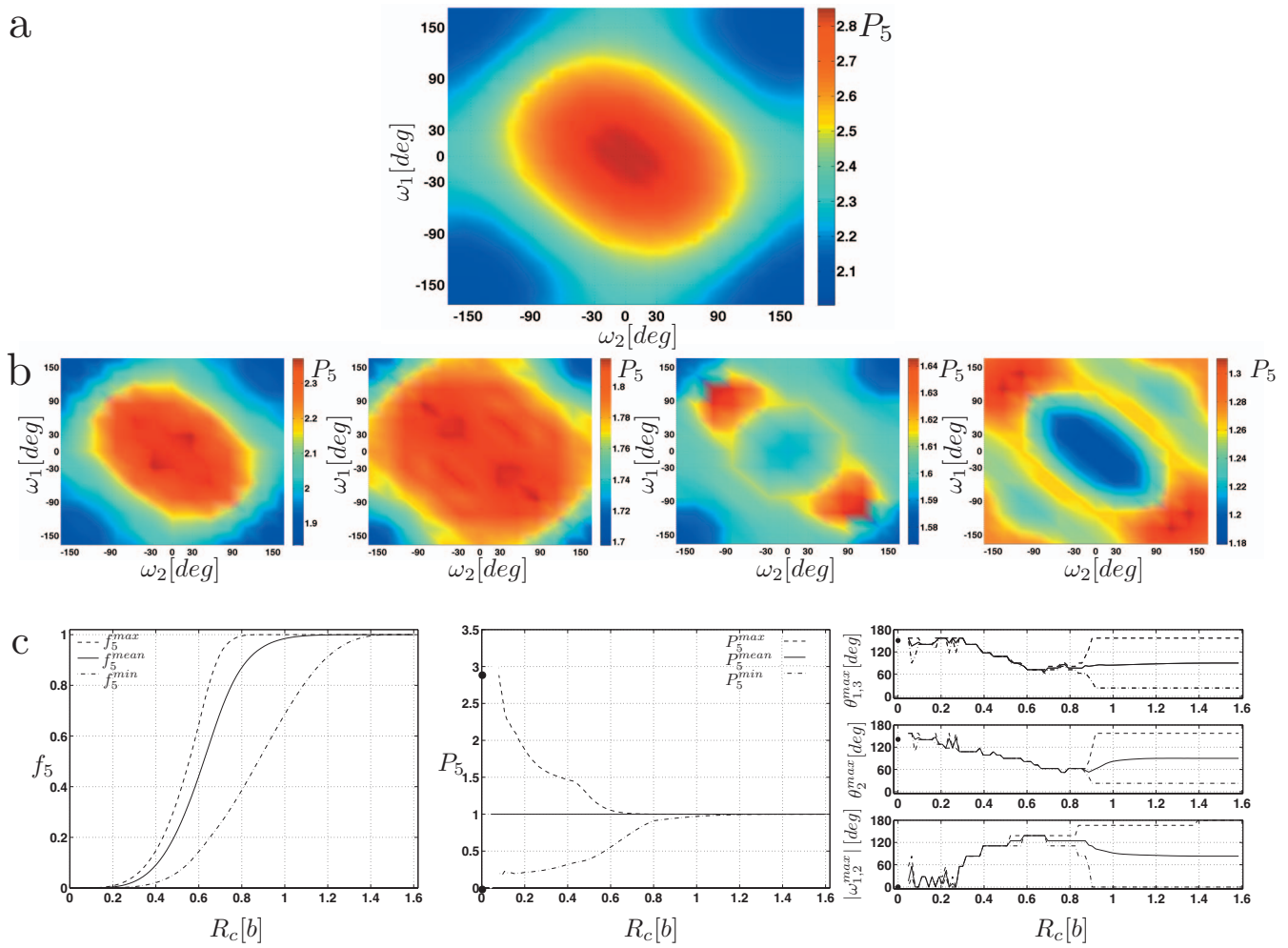

FIG. 8. (Color) (a) Normalized local shape probability distribution $P_5(\mathbf{q},0)$ [Eq. (19)], where $\mathbf{q}=\{\cos\theta_1,\cos\theta_2,\cos\theta_3,\omega_1,\omega_2\}$, displayed as a maximum projection onto the two dihedral angles ω_1 and ω_2 of the walk. (b) Corresponding normalized finite-cutoff shape probability distribution $P_5(\mathbf{q},R_c)$ [Eq. (16)] displayed as a maximum projection onto the two dihedral angles ω_1 and ω_2 of the walk; left to right: $R_c=0.16b$, $0.32b$, $0.40b$, and $0.72b$, with b being the bond length. (c) Left: maximum ($f_5^{\max}$), mean (f_5^{mean}), and minimum ($f_5^{\min}$) values of the fractional coverage function $f_5(\mathbf{q},R_c)$ [Eq. (17)] over $\mathcal{Q}_5$, displayed as a function of the cutoff R_c. Middle: maximum ($P_5^{\min}$), mean (P_5^{mean}), and minimum ($P_5^{\max}$) values of $P_5(\mathbf{q},R_c)$ over $\mathcal{Q}_5$ displayed as a function of the cutoff R_c. Right: maximum, mean, and minimum values of the outer angles θ_1 and θ_3, of the central angle θ_2, and of the two dihedral angles ω_1 and ω_2 (absolute values) over the set of structure maximizing $P_5(\mathbf{q},R_c)$, displayed as a function of the cutoff R_c; points indicated at $R_c=0$ in the middle and right panels correspond to expected values based on the local probability analysis. The data in (a) were evaluated using $g=24$ grid points per dimension, and the data in (b) and (c) using $g=13$ grid points per dimension.

$P_6(\mathbf{q}_6^{\#},0)=10.05$. Here too, the presence of a single maximum is not a consequence of the above-mentioned symmetry properties (these merely imply that if the maximum is unique, it must satisfy $\theta_1^{\#}=\theta_4^{\#}$, $\theta_2^{\#}=\theta_3^{\#}$ and $\omega_1^{\#}=\omega_2^{\#}=\omega_3^{\#}=0°$). This specific shape is about ten times more likely (in a local sense) than any shape taken at random.

The finite-cutoff shape probability density $P_6(\mathbf{q},R_c)$ is displayed in Fig. 9(b), also in the form of a maximal-value projection, for three different cutoff values R_c. All graphs preserve the symmetry features described above for $P_6(\mathbf{q},0)$. The curve corresponding to the lowest R_c value is again the closest to the limiting case $R_c\to 0$ [Fig. 9(a)]. Upon increasing R_c, the bias in the distribution progressively shifts from flat *cis-cis-cis* structures in the direction of *gauche*$^{\pm}$-*gauche*$^{+}$-*gauche*$^{\mp}$ or *gauche*$^{\pm}$-*gauche*$^{-}$-*gauche*$^{\mp}$ structures, while the single maximum becomes split into a pair (two alternative ω_2 values) of two symmetry-related (enantiomeric) maxima with $\omega_1,\omega_2,\omega_3\neq 0$ values.

The dependence of $P_6(\mathbf{q},R_c)$ on the cutoff value is characterized in more detail in Fig. 9(c), analogous to Figs. 6(c), 7(c), and 8(c) for $N=3$, 4, 5 (see explanations above). These curves display the same qualitative features as for $N=3$, 4, and 5, although the numerical precision is again considerably lower (especially for low R_c values). The function $f_6^{\max}$ reaches 1 at a cutoff value $R_6^{*}=0.998b$ for a specific pair of shapes (barycentric shapes $\mathbf{q}_6^{*}$) that are fourfold replicated by symmetry and associated with $P_6(\mathbf{q}_6^{*},R_6^{*})=1.06$. One pair of these structures corresponds to $\theta_1^{*}=\theta_4^{*}=66°$, $\theta_2^{*}=\theta_3^{*}=78°$, $\omega_1^{*}=144°$, $\omega_2^{*}=\pm 72°$, and $\omega_3^{*}=-144°$. The other pairs are obtained by the changes $\omega_1^{*}\leftrightarrow\omega_3^{*}$ or/and $\omega_1^{*},\omega_2^{*},\omega_3^{*}\leftrightarrow-\omega_1^{*}$, $-\omega_2^{*},-\omega_3^{*}$. The evolution of the angles and dihedral angles associated with the most likely shapes upon increasing R_c from 0 to R_6^{*} is difficult to assess in detail (numerical noise for small R_c) but agrees with the trends observed in Fig. 9(b). The function $f_6^{\min}$ reaches 1 at a cutoff value of $R_6^{**}=1.899$. Above this R_c value, all shapes encompass the entire ensemble.

The approximate relationships of Eqs. (22) and (23) con-

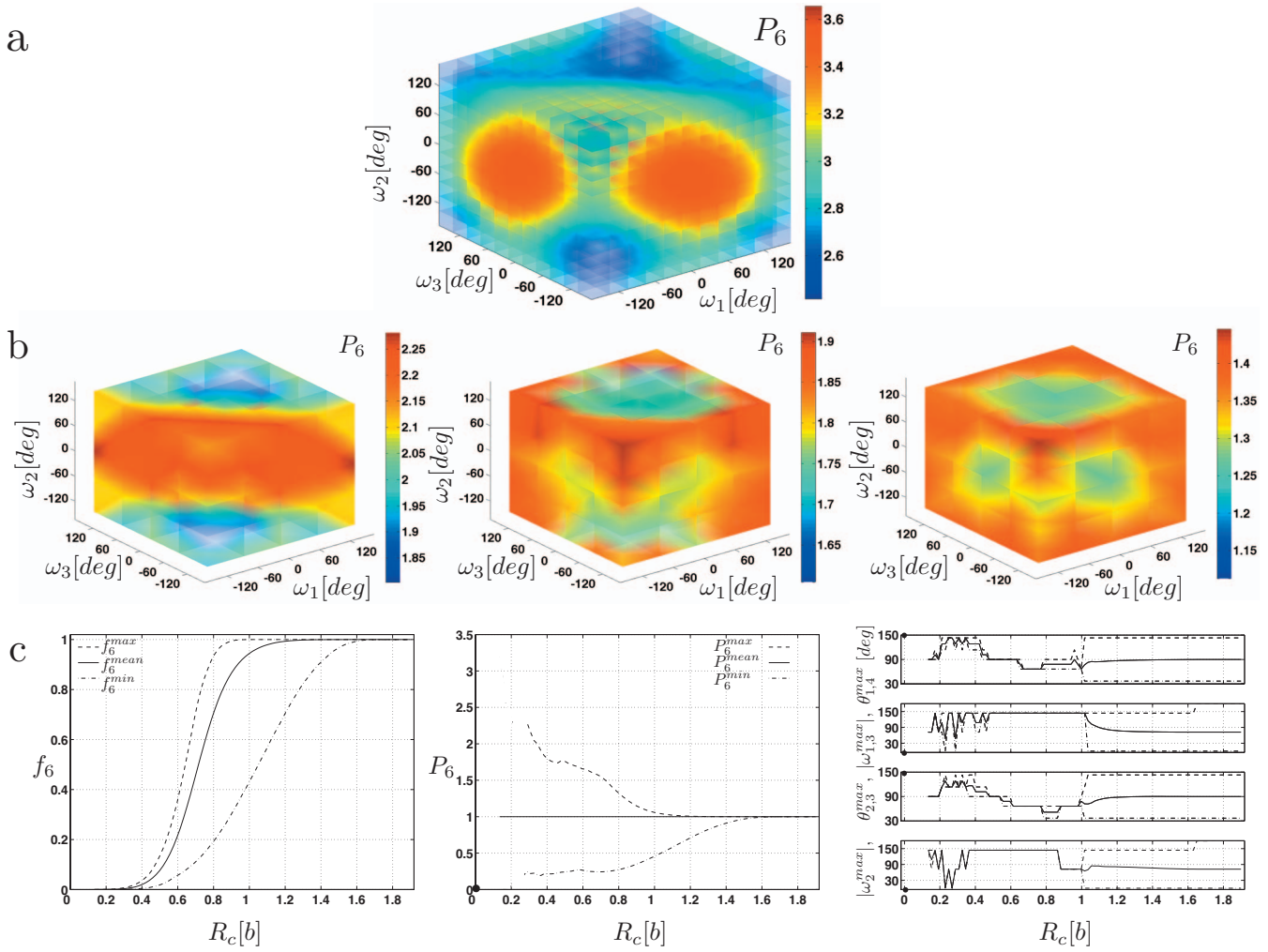

FIG. 9. (Color) (a) Normalized local shape probability distribution $P_6(\mathbf{q},0)$ [Eq. (19)], where $\mathbf{q}=\{\cos\theta_1,\cos\theta_2,\cos\theta_3,\cos\theta_4,\omega_1,\omega_2,\omega_3\}$, displayed as a maximum projection onto the three dihedral angles ω_1, ω_2, and ω_3 of the walk. (b) Corresponding normalized finite-cutoff shape probability distribution $P_6(\mathbf{q},R_c)$ [Eq. (16)] displayed as a maximum projection onto the three dihedral angles ω_1, ω_2, and ω_3 of the walk; left to right: $R_c=0.29b$, $0.38b$, and $0.76b$, with b being the bond length. (c) Left: maximum ($f_6^{\max}$), mean (f_6^{mean}), and minimum ($f_6^{\min}$) values of the fractional coverage function $f_6(\mathbf{q},R_c)$ [Eq. (17)] over $\mathcal{Q}_6$ displayed as a function of the cutoff R_c. Middle: maximum ($P_6^{\max}$), mean (P_6^{mean}), and minimum ($P_6^{\min}$) values of $P_6(\mathbf{q},R_c)$ over $\mathcal{Q}_6$ displayed as a function of the cutoff R_c. Right: maximum, mean, and minimum values of the of the outer angles θ_1 and θ_4, of the central angles θ_2 and θ_3, of the two outer dihedral angles ω_1 and ω_3 (absolute value) and of the central dihedral angle ω_2 (absolute value) over the set of structure maximizing $P_6(\mathbf{q},R_c)$, displayed as a function of the cutoff R_c; points indicated at $R_c=0$ in the middle and right panels correspond to expected values based on the local probability analysis. The data in (a) were evaluated using $g=12$ grid points per dimension, and the data in (b) and (c) using $g=5$ grid points per dimension.

necting the quantities $\Omega_N(\mathbf{q},R_c)$ or $f_N(\mathbf{q},R_c)$ to $\Gamma_N(\mathbf{q})$ for small R_c values (Sec. II E) are illustrated in Fig. 10 for $N=3,\ldots,6$, taking the corresponding densest structures $\mathbf{q}_N^{\#}$ (Table II) as examples. The approximate expression of Eq. (23) is seen to capture well the overall trend in the successive $f_N(\mathbf{q},R_c)$ curves at low R_c. At large R_c, negative deviations from this approximate expression result from the finiteness of the $\tilde{\mathcal{R}}_N$ hypersurfaces (f_N must converge to one for any shape, latest when $R_c=R_N^{**}$). Negative deviations may in the present case also result from the assumption that the local random-walk density is constant within the considered hyperdisk (while it actually decreases upon going away from the densest structures). On the other hand, positive deviations (clearly visible for $N=4$ when $R_c<0.29b$, and also present for $N\geq 5$ at small R_c values) arise from the neglect in the hyperdisk approximation of the curvature of $\mathcal{R}_N(\mathbf{r})$ in $\mathbb{R}^{3N}$ around $\mathbf{r}(\mathbf{q}_N^{\#})$ (an error that subsists even in the limit $R_c\to 0$).

Considering the key parameters of the distributions for $N=3,\ldots,6$ (Table II up to $N=9$ for some parameters) and the nature of the corresponding densest and barycentric structures, illustrated in Fig. 11, the following observations can be made.

The central structures associated with the densest shapes are all planar up to $N=6$ with bond angles progressively opening with increasing N [Fig. 11(a)]. In contrast, the central structures associated with the barycentric shapes are not planar (except for $N=3$) and do not present obvious systematic trends upon increasing N [Fig. 11(b)]. This may, however, also be due to the relatively low precision in the determination of these structures for $N=5$ and 6. Based on the present results, it is not possible to determine whether the above properties of the densest and barycentric structures also hold for larger N values.

The threshold cutoff radii R_N^* and R_N^{**} systematically increase upon increasing N. This is expected since the size of

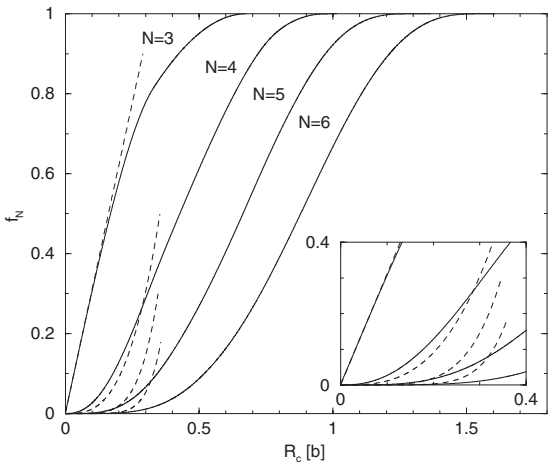

FIG. 10. Fractional coverage functions $f_N(\mathbf{q},R_c)$ [Eq. (17)] associated with the densest shapes $q_N^{\#}$ (Table II) for the random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,6$ beads, displayed as a function of the cutoff distance R_c. The numerically determined functions (solid lines) are compared to the corresponding approximate estimates from Eq. (23) (dashed lines). The inset represents a close-up of the low R_c region. The number of grid points per dimension was set to $g=10^4$, 200, 24, 12 for $N=3,\ldots,6$. The values of $\Gamma_N(\mathbf{q}^{\#})$ used in Eq. (23) can be deduced from the data in Table II via Eqs. (2) and (19).

the accessible conformational space also increases. As will be presented in a following article,[67] R_N^* values can be optimally fitted (for $N=3,\ldots,6$) by the logarithmic expression $R_N^* \approx (0.7393\ \log(N)-0.3207)b$, while R_N^{**} rapidly approaches the linear (analytically derived) expression $R_N^{**} \approx (5/48)^{1/2}Nb$. In contrast to the cutoff radii, the area $A_{\tilde{\mathcal{R}}_N}$ of the hypersurface $\tilde{\mathcal{R}}_N$ decreases upon increasing N. This is due to the fact that this quantity is associated with a M-dimensional space that has been scaled by $N^{1/2}$ in all its dimensions compared to a Cartesian space (Sec. II B). Thus, the present decrease in $A_{\tilde{\mathcal{R}}_N}$ rather accounts for a relative decrease in flexibility based on a size-consistent reference. When $A_{\tilde{\mathcal{R}}_N}$ is amplified by the factor $N^{M/2}$, the spanned area increases very rapidly with N.

Finally, the local shape probability density $P_N(\mathbf{q}_N^{\#},0)$ associated with the densest shape is seen to increase upon increasing N. This dependence can be approximately represented by the equation

$$P_N(\mathbf{q}_N^{\#},0) \approx \exp[(2/3)(N-3)]. \tag{27}$$

This equation is neither very accurate, nor strongly supported by the data (only $4N$ values), but nevertheless captures the apparent exponential increase in $P_N(\mathbf{q}_N^{\#},0)$ with N. Tentatively extrapolating the above sequence according to Eq. (27) suggests that $P_N(\mathbf{q}_N^{\#},0)$ could become as large as about 10^{28} for $N=100$. However, this extremely high number should be interpreted correctly. It represents the ratio of the local probability of the most likely shape to the corresponding probability averaged over all possible shapes, i.e., a relative probability defined by the ratio of two infinitesimal probabilities. When approximated using a finite discretization scheme, both probabilities will be seen to decrease very rapidly upon increasing N because the size of the $\mathcal{Q}_N$ space increases exponentially with M [Eq. (2)]. As a result, the densest shape becomes (in an absolute sense and just like all other shapes) increasingly improbable when N increases. For this reason, the properties of this single shape will never dominate those of the ensemble (i.e., dominate over those of all other possible shapes taken together). However, in a relative sense, this specific shape becomes overwhelmingly more probable than all other possible shapes taken individually.

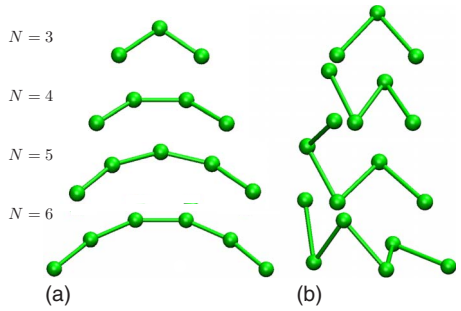

FIG. 11. (Color online) (a) Central structures associated with the densest shape. (b) Central structures associated with the barycentric shape. The different structures are associated to the random-walk ensembles $\mathcal{W}_N$ with different number of beads $N=3,\ldots,6$. All structures are drawn according to the internal coordinates $\mathbf{q}^{\#}$ and $\mathbf{q}^{*}$ reported in Table II. For $N=4-6$, only one of the 2 or 4 alternative barycentric shapes is represented ($N=4$: $\omega_1=162.4°$; $N=5$: $\omega_1=166°$, $\omega_2=-83°$; $N=6$: $\omega_1=144°$, $\omega_2=72°$, and $\omega_1=-144°$). The parameters of $\mathbf{q}^{\#}$ for $N=3,\ldots,6$ were refined to a precision of at least 1°. The parameters of $\mathbf{q}^{*}$ are of more limited precisions, especially for $N=5,6$.

To better appreciate the extent of inhomogeneity in the distribution of random walks into shapes, the integral $P_{\text{in},N}$ of the local shape probability distribution over the most likely shapes up to a fractional volume v_N of $\mathcal{Q}_N$, scaled by its value $V_{\mathcal{Q}_N}$ over the entire $\mathcal{Q}_N$ [Eqs. (24) and (26)], and the corresponding average $P_{\text{av},N}$ over the most likely shapes up to a fractional volume v_N of $\mathcal{Q}_N$ [Eqs. (24) and (26)] are displayed in Figs. 12(a) and 12(c), respectively, for the non-self-avoiding random-walk ensemble. For readability, $P_{\text{in},N}$ is actually shown in the form of $P_{\text{in},N}-v_N$, i.e., in excess to a diagonal line. As expected (Sec. II E), $P_{\text{in},N}-v_N$ [Fig. 12(a)] evaluates to 0 for $v_N=0$ (no shape selected, $P_{\text{in},N}=v_N=0$) or $v_N=1$ (all shapes selected, $P_{\text{in},N}=v_N=1$). At intermediate values, $P_{\text{in},N}-v_N$ is systematically positive indicating that the fraction of $\mathcal{Q}_N$ encompassed by the selected set of most likely shapes is larger than the corresponding fraction of the $\tilde{\mathcal{R}}_N$ hypersurface they cover. For $N=3,\ldots,9$, the maximum of the curve is close to 0.5 (shifting to the left upon increasing N), and the corresponding value ranges between 0.12 and 0.25 (increasing upon increasing N). For example, the maximal value $P_{\text{in},9}-v_9=0.25$ at $v_9=0.40$ indicates that the subset of most likely shapes covering 40% of $\mathcal{Q}_9$ actually encompass 50% the random walks of $\mathcal{W}_9$ (25% in excess of 40%). As also expected (Sec. II E), $P_{\text{av},N}$ [Fig. 12(c)] evaluates to $P_N(\mathbf{q}_N^{\#},0)$ for $v_N=0$ (densest shape) and to 1 for $v_N=1$ (all shapes selected). The functions monotonotically decrease upon increasing v_N and become consistently higher upon increasing N. Thus, the average of the local probability $P(\mathbf{q};0)$ over the selected set of most likely shapes is consistently higher than the corresponding average over all shapes for $v_N<1$ and the effect becomes more pronounced upon increasing N. For example, the value $P_{\text{av},9}=1.60$ at $v_9=0.40$ indicates that the subset of most likely shapes covering 40%

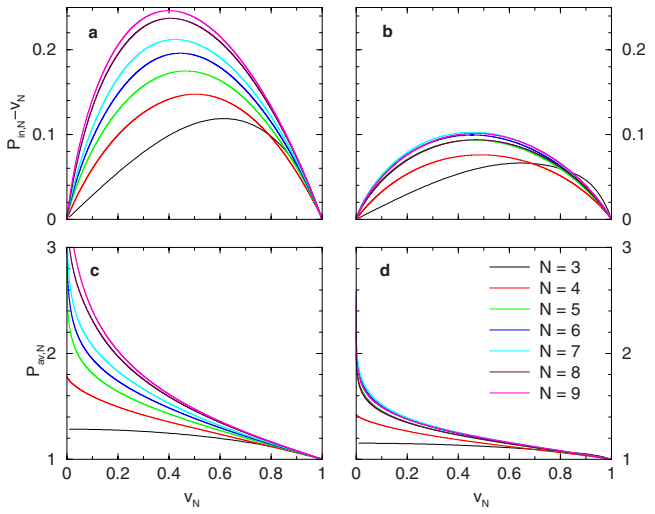

FIG. 12. (Color online) [(a) and (b)] Integral $P_{\mathrm{in},N}$ of the local shape probability distribution over the most likely shapes up to a fractional volume v_N of $\mathcal{Q}_N$, scaled by its value $V_{\mathcal{Q}_N}$ over the entire $\mathcal{Q}_N$ [Eqs. (24) and (25)], for the non-self-avoiding (a) and self-avoiding (b) random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,9$. [(c) and (d)] Average $P_{\mathrm{av},N}$ of the local shape probability distribution over the most likely shapes up to a fractional volume v_N of $\mathcal{Q}_N$ [Eqs. (24) and (26)], for the non-self-avoiding (c) and self-avoiding (d) random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,9$. For readability, $P_{\mathrm{in},N}$ is shown in the form of $P_{\mathrm{in},N}-v_N$, i.e., in excess to a diagonal line. The data were evaluated using grid-based sampling in the internal-coordinate space. The number of grid points per dimension was set to $g=10^4$, 200, 24, 12, 8, 4, and 4 for $N=3,\ldots,9$.

of $\mathcal{Q}_9$ have an average probability that is 60% higher than the corresponding average over all possible shapes of $\mathcal{Q}_9$.

The preceding discussion concerned exclusively non-self-avoiding walks. Key parameters of the shape probability distributions over the self-avoiding walk ensemble are also reported in Table II, while the corresponding curves for $P_{\mathrm{in},N}$ and $P_{\mathrm{av},N}$ are displayed in Figs. 12(b) and 12(d), respectively. For $N=3,\ldots,6$, the densest structures $\mathbf{q}_N^{\#}$ of the self-avoiding walk ensembles are identical to those of the non-self-avoiding walk ensembles (because these are self-avoiding). However, the corresponding local probabilities $P_N(\mathbf{q}^{\#};0)$ as well as the areas $A_{\tilde{\mathcal{R}}_N}$ are systematically lower (for the same N). The barycentric structures $\mathbf{q}_N^{*}$ are also self-avoiding. It is therefore probable that $\mathbf{q}_N^{*}$ and R_N^{*} are also identical for the two types of ensembles (but not guaranteed because the boundary of $\mathcal{Q}_N$ contains both self-avoiding and non-self-avoiding structures). On the other hand, it can be shown[67] that the R_N^{**} values are smaller for the self-avoiding walk ensemble. Finally the $P_{\mathrm{in},N}$ and $P_{\mathrm{av},N}$ curves (along with the above observations) suggest that the inhomogeneity in the shape probability distributions is less significant in the self-avoiding compared to the non-self-avoiding ensemble (at identical N).

V. DISCUSSION

The ideal random-walk model is a ubiquitous model in science, providing a very reasonable approximate description of many phenomena involving weakly correlated successive displacements. In the structural context relevant to polymer science, structural biology, and biopolymer dynamics this model accounts for the statistical (ensemble) properties of a polymer chain in three dimensions in the idealized situation where interatomic interactions are restricted to a generic connectivity constraint fixing the distance between successive sites (e.g., atom or reference point of a monomer) along the chain. The random-walk ensemble is then defined as the ensemble of chain conformations in which the directions of the (pseudo)bonds connecting these successive sites are isotropically distributed and uncorrelated among each other. Equivalently, this ensemble is also the one in which the internal degrees of freedom of the chain (bond-angle cosines and dihedral angles) are homogeneously distributed.

Because the random-walk ensemble appears to be the paramount of homogeneity, isotropy, and randomness, it may seem counterintuitive that its analysis in terms of shape distribution (at a low-resolution level, as done in previous works[7–11] or at a high-resolution level, as done in the present study) reveals inhomogeneity, anisotropy, and ordering features. However, the paradox is only apparent if one realizes that these features are actually introduced by the arbitrary definition selected for the concept of shape, i.e., by the specific properties of a corresponding structure-to-shape coordinate transformation.

At the low-resolution level, shapes may be characterized by single-structure observables such as the end-to-end distance, radius of gyration, radius of gyration tensor anisotropy, or number of contacts. As shown in previous studies[7–11] and in the present work, the distributions of these observables over the random-walk ensemble are already highly inhomogeneous. The distribution of the simplest observable, the end-to-end distance, can be evaluated analytically for an arbitrary number N of beads in the chain (see the Appendix). The root-mean-square value of the radius of gyration can also be given an analytical expression for an arbitrary N (see the Appendix). For the radius of gyration tensor anisotropy recent (analytical, computational, and experimental) work revealed that the average shape over the random-walk ensemble is not spherical but a prolate ellipsoid with a ratio of the mean square principal radii of gyration of approximately 9:2.3:1 (Refs. 11, 13, and 18) (high N limit; in good agreement with the present ratio of 9.4:2.6:1 for $N=9$). This ratio corresponds to a ratio of about 3.1:1.6:1 between the principal axes of the ellipsoid, in good qualitative agreement with the result of 4.1:2.3:1 obtained experimentally for DNA.[18] Finally, the present results for the reduced number of contacts suggest a relatively low compactness (only about 25% of the theoretically possible numbers of contacts being realized on average). When the above low-resolution observables are expressed in an appropriate reduced form (Sec. II D), the corresponding distributions and averages converge to unique (i.e., N-independent) expressions in the limit $N\rightarrow\infty$. These limiting expressions already represent very good approximations for relatively low values of N (Figs. 2–5) and are therefore typically the most relevant for practical purposes.

At the high-resolution level, the definition of shapes requires the specification of (i) a pairwise metric for the determination of the difference (distance) between two structures and (ii) an algorithm for classifying an arbitrary ensemble of structures into a corresponding ensemble of shapes based on

the knowledge on all pairwise distances (along with specific assignment parameters). The present study has focused exclusively on the most commonly used comparison metric, namely, the RMSD,[39,51–55] and on one possible classification scheme involving a single assignment parameter (cutoff distance) and a rather simple shape definition scheme (collection of all structures within the given cutoff distance of a central structure). Alternative metrics (e.g., unit-vector RMSD,[41] universal RMSD,[43] distance-matrix root-mean-square difference,[36,56] ρ measure,[40] or TM-align parameter[46]) and more complex shape assignment algorithms (e.g., clustering[27–34]) are also commonly used.

However, even if derived in the context of one specific metric and assignment scheme, the main conclusions of the present work certainly also pertain to any other metrics or assignment schemes in a qualitative sense. Given the above choices of metric and shape definition, the main observations concerning the high-resolution shape analysis of the random-walk ensemble may be summarized as follows.

The local shape probability distribution within the random-walk ensemble (i.e., the local density of structures in the immediate neighborhood of a given central structure, relative to its average over all possible central structures) is by no means homogeneous across all possible shapes. Even in the absence of interatomic interactions (beyond the mere connectivity constraint), some shapes are intrinsically more probable, while others (e.g., those defined by a central structure with one or more bond angles equal to 0 or π) have a vanishing probability. Over the limited range of sizes ($N=3,\ldots,6$) that could be probed in the present study, the bias in favor of the most probable (densest) shape increases in the sequence 1.28, 1.79, 2.94, and 10.05, as measured by the probability of this shape relative to the corresponding average over all possible shapes. In other words, given a random structure with $N=6$ beads and prompted to make a guess for a shape to which this structure belongs, one would have about a tenfold higher chance of success by proposing the densest shape than by proposing any shape at random. Tentatively extrapolating the above sequence according to Eq. (27) suggests that for $N=100$, the densest shape could become as much as 10^{28} times more probable compared to the average. This number represents a relative probability (one shape relative to the average over all shapes). Of course, in an absolute sense, the probability of this shape, as well as that of any shape, decreases even faster. In other words, this number should not be interpreted as meaning that one shape will dominate the properties of the entire ensemble for large N but simply that one shape will become overwhelmingly likely compared to any other one (although it becomes increasingly less likely compared to the union of all other ones). Over the range $N=3,\ldots,6$, the central structures associated with the densest shape were all found to be planar with bond angles progressively opening upon increasing N [Fig. 11(a)]. Based on the present results, it is not possible to determine whether these features also hold for larger N values.

The finite-cutoff shape probability distribution (i.e., the integrated density of structures within a specified cutoff distance from a given central structure, relative to its average over all possible central structures) as well as the fractional coverage function (i.e., the integrated density of structures within a specified cutoff distance from a given central structure, relative to its value at infinite cutoff) evidence similar qualitative features for all values of N considered. Three regimes are observed upon increasing the cutoff R_c: (i) for R_c below some threshold value R_N^*, all shapes only encompass part of the ensemble; (ii) for R_c above some threshold value $R_N^{**}>R_N^*$, all shapes encompass the entire ensemble; (iii) for intermediate values $R_N^*\leq R_c\leq R_N^{**}$, a single shape ($R_c=R_N^*$; barycentric shape), and then an increasingly larger set of shapes ($R_c>R_N^*$), encompasses the entire ensemble. A more detailed analysis of the properties of R_N^* and R_N^{**} will be presented in a following article.[67] Over the range $N=3,\ldots,6$, the central structures associated with the barycentric shapes are not planar (except for $N=3$) and do not appear to present obvious systematic trends [Fig. 11(b)]. This may, however, also be due to the relatively low precision in the determination of these structures for $N=5$ and 6.

In view of the limited number of beads considered in the present work ($N=3,\ldots,6$), these results can only be extrapolated to longer polymer chains (e.g., proteins) in a qualitative manner. It is nevertheless quite revealing to attempt such an extrapolation because the determination of polymer shapes and the evaluation of their relative probabilities is a core problem in structural biology and biopolymer dynamics. Two prototypical examples are given below.

As a first example, schemes have been designed to classify experimentally determined three-dimensional protein structures (e.g., via x-ray crystallography or NMR spectroscopy) into a limited number of protein shapes[19–22] (usually termed folds in this case). The assignment of individual structures to common folds is generally performed on the basis of pairwise distance (typically RMSD) comparisons between the different structures in a database. In this specific context, the present results suggest that the same analysis applied to the random-walk ensemble (i.e., in the total absence of the interatomic interactions characteristic of a specific protein) would also lead to the identification of random-walk "folds," thereby suggesting that part of what is called a protein fold is contained in the nature of the assignment scheme used to define it rather in the physics of protein systems.

As a second example, peptide conformational ensembles generated through computer simulation techniques [e.g., molecular dynamics (MD)] are often analyzed by clustering so as to identify a limited number of peptide shapes[27–34] (usually termed states in this case). This clustering of individual structures into common states is also generally performed on the basis of pairwise distance (typically RMSD) comparisons between the configurations sampled along the simulation. In addition, the populations of the various states are then commonly interpreted in terms of corresponding relative free energies. In this specific context, the present results suggest that the same analysis applied to the random-walk ensemble (i.e., in the total absence of the interatomic interactions characteristic of a specific peptide) would also lead to the identification random-walk "states" and associated relative free energies, thereby suggesting that part of what is considered

to represent the peptide conformational landscape (and associated free-energy surface) is contained in the nature of the clustering scheme used to defined it rather in the physics of peptide systems. Note that the above bias is entirely distinct in nature from the so-called metric-tensor effects in MD simulations.[68–78] The latter effects arise from the dependence of the kinetic energy on the atomic coordinates in MD simulations involving constraints. Owing to this bias, MD simulations of a chain in the absence of interatomic interactions (besides the bond-length constraints) will not exactly sample the random-walk ensemble. Monte Carlo (MC) simulations,[79–81] in contrast, are exempt of such effects.

As illustrated by the two above examples, the inhomogeneity of the shape probability distribution within the random-walk ensemble will depend on the chosen pairwise metric and cutoff value (or on the corresponding resolution parameters in more complex assignment schemes), i.e., on choices that characterize the way one has decided to look at structures rather than the physics of the system itself. This bias, most easily identified in the shape analysis of the random-walk ensemble, will remain present as an underlying artifact in the analysis of real polymer chain ensembles. As a consequence, part of what is called a polymer (e.g., protein) shape (e.g., fold or state) may actually reside "in the eye of the beholder" rather than in the nature of the interactions between the constituting atoms. By analogy, one could say that the "ideal-gas" model of shape analysis (i.e., the non-self-avoiding random-walk ensemble seen as the baseline model in the absence of interatomic interactions specific of a given system, beyond the generic connectivity constraint) does not present a homogeneous distribution in the different shapes (as defined by the specific metric and cutoff), and that this inhomogeneity should be taken into account when drawing conclusions from such an analysis as applied to real structural ensembles. In other words, shape analysis should be concerned with the excess probability of a shape relative to its corresponding probability in the random-walk ensemble rather than with the corresponding absolute probability. The corresponding baseline correction could, in principle, be evaluated by running the specific shape analysis over the random-walk ensemble.

From a Bayesian perspective,[82–85] one could interpret the covalent connectivity as the "prior knowledge" on the model. In the absence of "additional information" to complement this prior knowledge, the most likely shape probability distribution is just the inhomogeneous distribution derived in the present work for the random-walk ensemble (Figs. 6–9), the most likely shapes being the corresponding densest shapes [Fig. 11(a)]. For example, the optimal answer to the question "what is the shape (in a local sense, in three dimensions and based on a RMSD metric) of a triatomic molecule," with the prior knowledge that it involves two bonds of identical lengths and no additional experimental information, is the shape corresponding to an angle of 105.5° (which coincidentally turns out to be identical to the corresponding angle of 105.5° obtained by *ab initio* calculations[86] and very close to that of 106° obtained by neutron scattering[87] in the water molecule). However, the shape analysis of real polymer ensembles should ideally characterize solely the additional information (shape distribution induced by the effect of interatomic interactions in a specific polymer), without being biased by the prior knowledge (heterogeneous shape distribution associated with the random-walk ensemble).

From another perspective, one might interpret the heterogeneity in the shape probability density $P_N(\mathbf{q},R_c)$ within the random-walk ensemble as the cause of a corresponding entropic bias $S_N(\mathbf{q},R_c)=k_B \ln P_N(\mathbf{q},R_c)$, where k_B is Boltzmann's constant. [Note that due to the chosen normalization for P_N in Eq. (14), S_N may be positive or negative.] However, it should be kept in mind that this entropy is an "entropy of observation." It is not related to the physics the system but to the properties of the shape-assignment procedure.[59,60]

Due to the limited number of beads considered in the present work ($N=3,\ldots,6$), the nature and magnitude of the bias in the context of real (bio)polymers is difficult to assess. It is nevertheless interesting to speculate about the possibility that structural motives commonly found e.g., in proteins (such as helices and sheets) might be to some extent favored by this bias in the shape distribution (in addition to the enthalpic and entropic effects related to interatomic interactions and solvation). This does not necessarily need to imply that these motives should represent the densest shapes for large N but merely that their probabilities as determined by physical (enthalpic and entropic) effects may be artificially enhanced by the present entropy of observation. One might also consider the possibility that a similar bias exists in the interpretation of experimental data. For example, reflection intensities from crystal of fiber diffraction experiments are commonly used to refine models for the three-dimensional structure of biomolecules. In a similar way as described in the present work, it is likely that there exists a specific (nonuniform) diffraction pattern associated with the random-walk ensemble, thereby suggesting that part of what is considered to be experimental information stems again from the nature of the diffraction experiment. This possibility has already been partly explored previously.[88,89]

Finally, it is interesting to draw a parallel between the present results and the work of Chan and Dill concerning random walks on cubic lattices.[90–92] Their work considered self-avoiding random walks (excluded-volume constraint) on cubic lattices (constraints on the allowed pseudoangle and pseudodihedral values) and with a specified number t of topological contacts (nonconnected beads of the walk that are adjacent on the lattice; compactness constraint), along with a shape definition based on patterns of topological contacts (secondary structure). The main conclusions (restated in the wording of the present article) were that (i) excluded-volume constraints (along with the connectivity constraints) promote structuring, i.e., higher shape probabilities for regular shapes (e.g., helical or sheet patterns of topological contacts) compared to the average shape probability over all possible (regular and nonregular) shapes and (ii) increasing compactness constraints (along with the connectivity and excluded-volume constraints) promote an increasing amount of structuring (i.e., increasing t amplifies the bias toward regular shapes; high compactness being most relevant in the context of globular macromolecules, e.g., proteins). The first conclu-

sion should probably be reassessed in view of the results of the present article, suggesting that more probable shapes already exist in the absence of any excluded-volume constraints (this could be done, e.g., by comparing the properties of non-self-avoiding and self-avoiding walks on cubic lattices). The second conclusion is not affected by the results of the present article, since it relies on a comparison between properties of different walk ensembles (depending parametrically on t).

VI. CONCLUSION

The present work represents a first attempt to investigate and characterize the heterogeneous probability distribution of high-resolution shapes within the random-walk ensemble. The results are probably more qualitatively than quantitatively useful due to the use of a simplified assignment procedure and the restriction to short chains. This work could be extended along three main lines: (i) the consideration of more complex shape assignment schemes, e.g., various forms of structural clustering[27–29,31–34] (including intuitively more appealing and practically more relevant schemes leading to nonoverlapping shape definitions), (ii) the consideration of alternative metrics,[36,40,41,43,46,56] and (iii) the consideration of longer chains (beyond $N=6$), more relevant to biopolymers (e.g., peptides, proteins, oligosaccharides, and polysaccharides). The latter extension using the systematic (grid-based) approach employed in the present work appears clearly impossible beyond $N \approx 10$, considering present-day computational resources. This scheme could, however, be somewhat extended by reducing the problem dimensionality (e.g., considering chains with rigid bond angles, i.e., reducing the problem to dihedral-angle degrees of freedom sampled in terms of three or six relevant conformations). Alternative approaches could involve heuristic (nonsystematic) schemes, specifically designed for the preferential sampling of the densest shapes, such as, e.g., MC (random) sampling,[80] quasi-MC or low-discrepancy sampling,[93,94] or evolutionary optimization algorithms.[95] The application of common fold classification[19–22] or clustering algorithms[27–29,31–34] to the random-walk ensemble might also be useful in characterizing this bias.

Clearly, this study only represents a preliminary step toward a precise characterization of the intrinsic (observation) bias involved in any high-resolution shape analysis as those commonly used in structural biology, biopolymer dynamics, and possibly also, in the interpretation of polymer diffraction data in terms of structure. However, it is to our knowledge the first investigation to date (with the possible exception of the above-mentioned work of Chan and Dill[90–92]) pointing toward the possible existence of such a bias.

ACKNOWLEDGMENTS

This work was financed by the National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia (EMBO Installation Grant, B.Z.) and by the Ministry of Science, Education and Sports of the Republic of Croatia (Unity through Knowledge Fund Grant 1A, B.Z.). The authors would like to thank Halvor Hansen for useful algorithmic suggestions and help in generating the figures of this article, Michel Cuendet for his careful reading of the manuscript (including many insightful comments), as well as the members of the IGC group (ETH Zürich, Switzerland) and MedILS institute (Split, Croatia) for useful discussions.

APPENDIX: ANALYTICAL RESULTS

In this appendix, expressions are derived for the probability distribution $P_N^E(r_E)$ of the reduced end-to-end distance r_E [Eq. (9)] and for the root-mean-square value $\langle r_G^2 \rangle_N^{1/2}$ of the reduced radius of gyration r_G [Eq. (10), along with an expression for the corresponding probability distribution $P_3^G(r_G)$ for $N=3$] over the (non-self-avoiding) random-walk ensemble $\mathcal{W}_N$ (Sec. II C). Although these derivations rely on well-known analytical approaches,[8–10] some of the final expressions reported here have, to our knowledge, not been explicitly formulated.

For these derivations, it is convenient to introduce the $3(N-1)$ dimensional vector $\mathbf{u} \doteq \{\mathbf{u}_n | n=1,\ldots,N-1\}$, where $\mathbf{u}_n \doteq \mathbf{r}_{n+1}-\mathbf{r}_n$ is the Cartesian displacement between beads n and $n+1$. The transformations $\mathbf{r} \rightarrow \mathbf{u}(\mathbf{r})$ and $\mathbf{u} \rightarrow \mathbf{r}(\mathbf{u})$ are trivial, keeping in mind the bond-length constraints as well as the six constraints imposed to the $\mathbf{r}$ components of the first three beads (mapping to three constraints on the $\mathbf{u}$ components corresponding to the first two displacements). However, in order to evaluate the distribution of an observable α that is translationally invariant, i.e., can be written $\alpha(\mathbf{u})$ and rotationally invariant i.e., independent of the specific values of the three constraints imposed on $\mathbf{u}$, it is actually easier to also integrate over the three constrained $\mathbf{u}$ components.

Using this approach, the distribution $P_N^\alpha(\alpha)$ of the observable α over $\mathcal{W}_N$ can be written as

$$P_N^\alpha(\alpha^o) = \int d^{3(N-1)}\mathbf{u} \Delta_N(\mathbf{u}) \delta[\alpha(\mathbf{u}) - \alpha^o] \tag{A1}$$

with

$$\Delta_N(\mathbf{u}) \doteq \prod_{n=1}^{N-1} (4\pi b^2)^{-1} \delta(u_n - b). \tag{A2}$$

Consider first the reduced end-to-end distance r_E as an observable α, defined by Eq. (9). Introducing the (three-dimensional) end-to-end vector $\mathbf{R}$ of a walk, defined as

$$\mathbf{R}(\mathbf{u}) \doteq \mathbf{r}_N - \mathbf{r}_1 = \sum_{n=1}^{N-1} \mathbf{u}_n, \tag{A3}$$

and the function

$$g(\mathbf{R}, r_E) \doteq \delta[(N-1)^{-1/2} b^{-1} |\mathbf{R}| - r_E], \tag{A4}$$

Equation (A1) can be rewritten in the specific case as

$$P_N^E(r_E) = \int d^{3(N-1)}\mathbf{u} \Delta_N(\mathbf{u}) g(\mathbf{R}(\mathbf{u}), r_E). \tag{A5}$$

The function $g(\mathbf{R}, r_E)$ can be expanded in a Fourier series, as

$$g(\mathbf{R},r_E) = (2\pi)^{-3}\int d^3\mathbf{k}_R \hat{g}(\mathbf{k}_R,r_E)\exp[i\mathbf{k}_R\cdot\mathbf{R}] \tag{A6}$$

with the coefficients

$$\hat{g}(\mathbf{k}_R,r_E) \doteq \int d^3\mathbf{R} g(\mathbf{R},r_E)\exp[-i\mathbf{k}_R\cdot\mathbf{R}]. \tag{A7}$$

Inserting Eq. (A4) into the latter expression and expanding the integral in spherical coordinates with the z-axis along $\mathbf{k}_R$ gives

$$\begin{aligned}\hat{g}(\mathbf{k}_R,r_E) &= 2\pi\int_0^\infty dRR^2\delta[(N-1)^{-1/2}b^{-1}R - r_E] \\ &\quad\times\int_0^\pi d\theta \sin\theta \exp[-ik_R R\cos\theta] \\ &= 4\pi\int_0^\infty dRR^2\delta[(N-1)^{-1/2}b^{-1}R - r_E] \\ &\quad\times(k_R R)^{-1}\sin(k_R R).\end{aligned} \tag{A8}$$

Making the change in variable $R\rightarrow\tilde{R}\doteq(N-1)^{-1/2}b^{-1}R$ and contracting the resulting $\tilde{R}$ integral to $\tilde{R}=r_E$ using the definition of the δ-function further leads to

$$\hat{g}(\mathbf{k}_R,r_E) = 4\pi(N-1)b^2 r_E k_R^{-1}\sin[k_R(N-1)^{1/2}br_E]. \tag{A9}$$

Using this result within Eq. (A6), Eq. (A5) can be rewritten as

$$\begin{aligned}P_N^E(r_E) &= (2\pi^2)^{-1}(N-1)b^2 r_E \\ &\quad\times\int d^3\mathbf{k}_R k_R^{-1}\sin[k_R(N-1)^{1/2}br_E]\hat{h}(\mathbf{k}_R)\end{aligned} \tag{A10}$$

with

$$\hat{h}(\mathbf{k}_R) \doteq \int d^{3(N-1)}\mathbf{u}\Delta_N(\mathbf{u})\exp[i\mathbf{k}_R\cdot\mathbf{R}(\mathbf{u})]. \tag{A11}$$

Using Eqs. (A2) and (A3), noting that the resulting $3(N-1)$-dimensional integral can be factored into a product of $N-1$ identical three-dimensional integrals and evaluating this integral by expansion in spherical coordinates, one finds

$$\hat{h}(\mathbf{k}_R) = [(k_R b)^{-1}\sin(k_R b)]^{N-1}. \tag{A12}$$

Inserting this expression into Eq. (A10), expanding the resulting integral in spherical coordinates (the integrand actually only depends on the norm of $\mathbf{k}_R$), one gets

$$\begin{aligned}P_N^E(r_E) &= 2\pi^{-1}(N-1)b^2 r_E\int_0^\infty dk_R k_R \sin[k_R(N-1)^{1/2}br_E] \\ &\quad\times[(k_R b)^{-1}\sin(k_R b)]^{N-1}.\end{aligned} \tag{A13}$$

This result can be further simplified by making the change of variable $k_R\rightarrow\tilde{k}_R\doteq k_R b$, leading to

$$\begin{aligned}P_N^E(r_E) &= 2\pi^{-1}(N-1)r_E\int_0^\infty d\tilde{k}_R\tilde{k}_R \sin(\tilde{k}_R(N-1)^{1/2}r_E) \\ &\quad\times[\tilde{k}_R^{-1}\sin\tilde{k}_R]^{N-1}.\end{aligned} \tag{A14}$$

Equation (A14) can be evaluated in an iterative way for successive values of N (starting from $N=3$). This leads [after introduction of the constraint $P_N^E(r_E)=0$ for $r_E<0$] to a series of piecewise-defined polynomials, i.e., functions defined as the union of polynomials over $[N-\mathrm{O}_N]/2+2$ successive intervals in $\mathbb{R}$, where O_N evaluates to one for N odd and zero otherwise. These functions can be written as

$$P_N^E(r_E) = (N-1)^{-1/2}\tilde{P}_N^E((N-1)^{1/2}r_E) \tag{A15}$$

with

$$\tilde{P}_N^E(x) \doteq \begin{cases} 0 & \text{if } x<0 \\ \cup_{n=1}^{[N-\mathrm{O}_N]/2}\sum_{m=1}^{N-2} c_{N,n,m}x^m & \text{if } \max\{0,2n-3+\mathrm{O}_N\}<x<2n-1+\mathrm{O}_N \\ 0 & \text{if } x>N-1, \end{cases} \tag{A16}$$

where the coefficients $\{c_{N,n,m}|N=1,\ldots,10,n=1,\ldots,[N-\mathrm{O}_N]/2,m=1,\ldots,N-2\}$ are listed in Table III. The $P_N^E(r_E)$ functions are either discontinuous ($N=3$), continuous to the first derivative ($N=4$), or continuous to the second derivative ($N\geq 5$). The corresponding average values of r_E are given by

$$\langle r_E\rangle_N \doteq \int_0^\infty dr_E r_E P_N^E(r_E) = \tilde{c}_N(N-1)^{-1/2}, \tag{A17}$$

where the coefficients $\{\tilde{c}_N|N=1,\ldots,10\}$ are also listed in Table III. These distributions are displayed in Fig. 2 and the corresponding averages listed in Table I for $N=3,\ldots,10$. Finally, the limiting distribution $P_\infty^E(r_E)$ for $N\rightarrow\infty$ is obtained by inserting the approximation (valid for large N)

$$\begin{aligned}(\tilde{k}_R^{-1}\sin\tilde{k}_R)^{N-1} &\approx [1-(1/6)\tilde{k}_R^2]^{N-1} \\ &\approx \exp[-(1/6)(N-1)\tilde{k}_R^2]\end{aligned} \tag{A18}$$

into Eq. (A14), both approximations resulting from Taylor expansions up to second order around $\tilde{k}_R=0$. Using this limiting expression, one obtains

TABLE III. Coefficients $\{c_{N,n,m}|N=1,\ldots,10, n=1,\ldots,[N-O_N]/2, m=1,\ldots,N-2\}$ of the piecewise-defined polynomials [Eqs. (A15) and (A16)] representing the probability distribution $P_N^E(r_E)$ of the reduced end-to-end distance [Eq. (9)] over the (non-self-avoiding) random-walk ensembles $\mathcal{W}_N$ with $N=3,\ldots,10$ The coefficients $\tilde{c}_N$ involved in the corresponding average $\langle r_E\rangle_N$ [Eq. (A17)] are also listed.

N	$\tilde{c}_N$	n	$c_{N,n,1}$	$c_{N,n,2}$	$c_{N,n,3}$	$c_{N,n,4}$	$c_{N,n,5}$	$c_{N,n,6}$	$c_{N,n,7}$	$c_{N,n,8}$
3	4/3	1	1	⋯	⋯	⋯	⋯	⋯	⋯	⋯
4	13/8	1	0	3/2	⋯	⋯	⋯	⋯	⋯	⋯
4	⋯	2	9/4	−3/4	⋯	⋯	⋯	⋯	⋯	⋯
5	28/15	1	0	2	−3/4	⋯	⋯	⋯	⋯	⋯
5	⋯	2	4	−2	1/4	⋯	⋯	⋯	⋯	⋯
6	1199/576	1	0	25/16	0	−5/16	⋯	⋯	⋯	⋯
6	⋯	2	−25/48	25/8	−25/16	5/24	⋯	⋯	⋯	⋯
6	⋯	3	625/96	−125/32	25/32	−5/96	⋯	⋯	⋯	⋯
7	239/105	1	0	3/2	0	−3/8	5/64	⋯	⋯	⋯
7	⋯	2	−15/8	21/4	−45/16	9/16	−5/128	⋯	⋯	⋯
7	⋯	3	81/8	−27/4	27/16	−3/16	1/128	⋯	⋯	⋯
8	113 149/46 080	1	0	539/384	0	−49/192	0	7/384	,,,	⋯
8	⋯	2	49/1536	637/512	245/768	−147/256	245/1536	−7/512	⋯	⋯
8	⋯	3	−8869/1920	2303/256	−931/192	147/128	−49/384	7/1280	⋯	⋯
8	⋯	4	117 649/7680	−16 807/1536	2401/768	−343/768	49/1536	−7/7680	⋯	⋯
9	1487/567	1	0	4/3	0	−2/9	0	1/48	−7/2304	⋯
9	⋯	2	14/45	2/5	7/6	−1	7/24	−3/80	7/3840	⋯
9	⋯	3	−434/45	46/3	−49/6	19/9	−7/24	1/48	−7/11 520	⋯
9	⋯	4	1024/45	−256/15	16/3	−8/9	1/12	−1/240	1/11 520	⋯
10	14 345 663/5 160 960	1	0	2601/2048	0	−387/2048	0	27/2048	0	−1/2048
10	⋯	2	−9/10 240	3267/2560	−189/10 240	−81/512	−63/2048	81/2560	−63/10 240	1/2560
10	⋯	3	13 113/10 240	−1755/1024	30 429/10 240	−1863/1024	1071/2048	−81/1024	63/10 240	−1/5120
10	⋯	4	−1 314 459/71 680	13 185/512	−138 321/10 240	1881/512	−1179/2048	27/512	−27/10 240	1/17 920
10	⋯	5	4 782 969/143 360	−531 441/20 480	177 147/20 480	−6561/4096	729/4096	−243/20 480	9/20 480	−1/143 360

$$P_\infty^E(r_E) = 2\pi^{-1}(N-1)r_E\Theta(r_E)\int d\tilde{k}_R\tilde{k}_R \sin[\tilde{k}_R(N-1)^{1/2}r_E] \times\exp[-(1/6)(N-1)\tilde{k}_R^2] = (54/\pi)^{1/2}r_E^2\Theta(r_E)\exp[-(3/2)r_E^2], \quad \text{(A19)}$$

where Θ is the Heaviside function. The corresponding average is

$$\langle r_E\rangle_\infty \doteq \int_0^\infty dr_E r_E P_\infty^E(r_E) = (3\pi/8)^{-1/2} \approx 0.9213. \quad \text{(A20)}$$

This distribution is displayed in the panel of Fig. 2 corresponding to $N=10$.

Consider next the reduced radius of gyration r_G as an observable α, defined by Eq. (10). By straightforward sum manipulations, this definition may be rewritten as

$$r_G(\mathbf{r}) = (N-1)^{-1/2}N^{-1}b^{-1}\left[\sum_{n=1}^{N}\sum_{m>n}^{N}(\mathbf{r}_m-\mathbf{r}_n)^2\right]^{1/2}. \quad \text{(A21)}$$

Reformulating in terms of the $\mathbf{u}$ vectors, one obtains

$$r_G(\mathbf{u}) = (N-1)^{-1/2}N^{-1}b^{-1}\left[\sum_{n=1}^{N}\sum_{m>n}^{N}\left(\sum_{l=n}^{m-1}\mathbf{u}_l\right)^2\right]^{1/2} = (N-1)^{-1/2}N^{-1}b^{-1}\left[\sum_{n=1}^{N-1}n(N-n)\mathbf{u}_n^2 + 2\sum_{n=1}^{N-1}\sum_{m>n}^{N-1}n(N-m)\mathbf{u}_n\cdot\mathbf{u}_m\right]^{1/2}. \quad \text{(A22)}$$

Recalling that all bond lengths are equal to b and that distinct bond vectors are uncorrelated in their directions, i.e., $\langle\mathbf{u}_n\cdot\mathbf{u}_m\rangle = b^2\delta_{n,m}$, one obtains

$$\langle r_G^2\rangle_N^{1/2} = (N-1)^{-1/2}N^{-1}\left[\sum_{n=1}^{N-1}n(N-n)\right]^{1/2} = [(1/6)(1+1/N)]^{1/2}. \quad \text{(A23)}$$

The limiting average $\langle r_G\rangle_\infty^{1/2}$ for $N\to\infty$ is then

$$\langle r_G^2\rangle_\infty^{1/2} = 6^{-1/2} \approx 0.4082. \quad \text{(A24)}$$

The minimum possible value $r_{G,N}^{\min}$ of r_G for a chain of length N is reached when all bond angles are equal to 0. In this case, straightforward application of Eq. (10) leads, after simplification, to the expression

$$r_{G,N}^{\min} = \frac{1}{2}\left[\frac{N + O_N}{N(N - 1 + O_N)}\right]^{1/2}. \tag{A25}$$

The corresponding maximum possible value $r_{G,N}^{\max}$ of r_G is reached when all bond angles are equal to π. In this case, straightforward application of Eq. (10) leads, after simplification, to the expression

$$r_{G,N}^{\max} = (12)^{-1/2}(N+1)^{1/2}. \tag{A26}$$

The distribution $P_N^G(r_G)$ is easily derived in the special case $N=3$. In this case, Eq. (A1) becomes

$$P_3^G(r_G^o) = \int d^6\mathbf{u}\Delta_3(\mathbf{u})\delta[r_G(\mathbf{u}) - r_G^o] \tag{A27}$$

with [Eq. (A22)]

$$r_G(\mathbf{u}) = (N-1)^{-1/2}N^{-1}b^{-1}2(1/2)(\mathbf{u}_1^2 + \mathbf{u}_2^2 + \mathbf{u}_1 \cdot \mathbf{u}_2)^{1/2}. \tag{A28}$$

Chosing $\mathbf{u}_1$ along the z-axis and $\mathbf{u}_2$ in the xz-plane, this can be rewritten in spherical coordinates

$$P_3^G(r_G^o) = (1/2)\int_0^\pi d\theta \sin\theta\delta[(1/3)(2+\cos\theta)^{1/2} - r_G^o]. \tag{A29}$$

Integrating this expression results in

$$P_3^G(r_G^o)\begin{cases} 0 & \text{if } r_G < 1/3 \quad \text{or} \quad r_G > (1/3)^{1/2} \\ 9r_G & \text{otherwise.} \end{cases} \tag{A30}$$

Analytical forms for $P_N^G(r_G)$ could not be derived for $N \geq 4$.

[1] R. M. Mazo, *Brownian Motion: Fluctuations, Dynamics and Applications* (Clarendon, Oxford, 2002).
[2] H. C. Berg, *Random Walks in Biology* (Princeton University Press, Princeton, 1983).
[3] T. Royama, *Analytical Population Dynamics* (Chapman and Hall, London, 1992).
[4] R. Engbert and R. Kliegl, Biol. Cybern. **85**, 77 (2001).
[5] B. Malkiel, *A Random Walk Down Wall Street* (W. W. Norton, New York, 1973).
[6] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. (Academic, New York, 1975).
[7] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Hanser, New York, 1969).
[8] M. Doi and S. J. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, Oxford, 1986).
[9] F. W. Wiegel, *Introduction to Path-Integral Methods in Physics and Polymer Science* (World Scientific, Singapore, 1986).
[10] A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).
[11] J. Rudnick and G. Gaspari, *Elements of the Random Walk: An Introduction for Advanced Students and Researchers* (Cambridge University Press, Cambridge, 2004).
[12] K Solc and WH Stockmayer, J. Chem. Phys. **54**, 2756 (1971).
[13] J. Rudnick and G. Gaspari, Science **237**, 384 (1987).
[14] F. Fougere and J. Desbois, J. Phys. A **26**, 7253 (1993).
[15] G. Y. Wei, Physica A **222**, 155 (1995).
[16] P. Biswas, A. Paramekanti, and B. J. Cherayil, J. Chem. Phys. **104**, 3360 (1996).
[17] G. Y. Wei, Physica A **237**, 413 (1997).
[18] C. Haber, S. A. Ruiz, and D. Wirtz, Proc. Natl. Acad. Sci. U.S.A. **97**, 10792 (2000).
[19] D. L. Bostick, M. Shen, and I. I. Vaisman, Proteins: Struct., Funct., Bioinf. **56**, 487 (2004).
[20] P. D. Sun, C. E. Foster, and J. C. Boyington, *Current Protocols in Protein Science* (Wiley and Sons, New York, 2004), Chap. 17.
[21] R. Kolodny, D. Petrey, and B. Honig, Curr. Opin. Struct. Biol. **16**, 393 (2006).
[22] Y. J. Kim and J. M. Patel, BMC Bioinf. **7**, 456 (2006).
[23] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker, Science **310**, 638 (2005).
[24] R. Das, Q. Bin, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. Tyka, D. Bhat, D. Chivian, D. E. Kim, W. H. Sheffler, L. Malmstrom, A. M. Wollacott, C. Wang, I. Andre, and D. Baker, Proteins: Struct., Funct., Bioinf. **69**, 118 (2007).
[25] H. Rangwala and G. Karypis, Proteins **72**, 1005 (2008).
[26] Y. Zhang, Curr. Opin. Struct. Biol. **18**, 342 (2008).
[27] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[28] X. Daura, W. F. van Gunsteren, and A. E. Mark, Proteins: Struct., Funct., Genet. **34**, 269 (1999).
[29] F. A. Hamprecht, C. Peter, X. Daura, W. Thiel, and W. F. van Gunsteren, J. Chem. Phys. **114**, 2079 (2001).
[30] G. M. Downs and J. M. Barnard, *Clustering Methods and Their Uses in Computational Chemistry* (Wiley, New York, 2002), Vol. 18, pp. 1–40.
[31] J. Y. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, J. Chem. Theory Comput. **3**, 2312 (2007).
[32] C. H. Jensen, D. Nerukh, and R. C. Glen, J. Chem. Phys. **128**, 115107 (2008).
[33] B. Zagrovic, Z. Gattin, J. K. C. Lau, M. Huber, and W. F. van Gunsteren, Eur. Biophys. J. Biophys. Lett. **37**, 903 (2008).
[34] X. Daura, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, J. Mol. Biol. **280**, 925 (1998).
[35] W. Huisinga, C. Best, R. Roitzsch, C. Schutte, and F. Cordes, J. Comput. Chem. **20**, 1760 (1999).
[36] B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande, J. Mol. Biol. **323**, 927 (2002).
[37] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, Biopolymers **68**, 91 (2003).
[38] R. Boned, W. F. van Gunsteren, and X. Daura, Chem.-Eur. J. **14**, 5039 (2008).
[39] A. D. McLachlan, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **28**, 656 (1972).
[40] V. N. Maiorov and G. M. Crippen, Proteins: Struct., Funct., Genet. **22**, 273 (1995).
[41] K. Kedem, L. P. Chew, and R. Elber, Proteins: Struct., Funct., Genet. **37**, 554 (1999).
[42] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Nature (London) **406**, 287 (2000).
[43] M. R. Betancourt and J. Skolnick, Biopolymers **59**, 305 (2001).
[44] P. Koehl, Curr. Opin. Struct. Biol. **11**, 348 (2001).
[45] P. Rogen and B. Fain, Proc. Natl. Acad. Sci. U.S.A. **100**, 119 (2003).
[46] Y. Zhang and J. Skolnick, Nucleic Acids Res. **33**, 2302 (2005).
[47] M. Kouza, C. K. Hu, and M. S. Lia, J. Chem. Phys. **128**, (2008).
[48] B. Zagrovic, E. J. Sorin, and V. Pande, J. Mol. Biol. **313**, 151 (2001).
[49] M. A. Kastenholz, T. U. Schwartz, and P. H. Hünenberger, Biophys. J. **91**, 2976 (2006).
[50] J. N. Onuchic, N. D. Socci, Z. Luthey Schulten, and P. G. Wolynes, Folding Des. **1**, 441 (1996).
[51] W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 922 (1976).
[52] W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **34**, 827 (1978).
[53] G. R. Kneller, Mol. Simul. **7**, 113 (1991).
[54] G. R. Kneller, J. Comput. Chem. **26**, 1660 (2005).
[55] T. Shibuya, J. Comput. Biol. **14**, 1201 (2007).
[56] A. F. Pereira de Araújo, A. L. C. Gomes, A. A. Bursztyn, and E. I. Shakhnovich, Proteins: Struct., Funct., Bioinf. **70**, 971 (2008).
[57] M. Andrec, D. A. Snyder, Z. Zhou, J. Young, G. T. Montelione, and R. M. Levy, Proteins: Struct., Funct., Bioinf. **69**, 449 (2007).
[58] E. Saccenti and A. Rosato, J. Biomol. NMR **40**, 251 (2008).
[59] D. C. Sullivan and I. D. Kuntz, Proteins: Struct., Funct., Genet. **42**, 495 (2001).
[60] D. C. Sullivan and I. D. Kuntz, Biophys. J. **87**, 113 (2004).
[61] T. E. Creighton, *Proteins*, 2nd ed. (Freeman, New York, 1993).
[62] K. Kaindl and B. Steipe, Acta Crystallogr., Sect. A: Found. Crystallogr. **53**, 809 (1997).
[63] B. Steipe, Acta Crystallogr., Sect. A: Found. Crystallogr. **58**, 507 (2002).
[64] B. Steipe, Acta Crystallogr., Sect. A: Found. Crystallogr. **58**, 506 (2002).
[65] B. K. P. Horn, J. Opt. Soc. Am. A **4**, 629 (1987).

[66] B. K. P. Horn, M. H. Hilden, and S. Negahdaripour, J. Opt. Soc. Am. A **5**, 1127 (1988).
[67] C. L. Müller, P. H. Hünenberger, B. Zagrovic, and I. F. Sbalzarini (unpublished).
[68] M. Fixman, Proc. Natl. Acad. Sci. U.S.A. **71**, 3050 (1974).
[69] N. Go and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[70] M. Gottlieb and R. B. Bird, J. Chem. Phys. **65**, 2467 (1976).
[71] E. Helfand, J. Chem. Phys. **71**, 5000 (1979).
[72] E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, Chem. Phys. Lett. **156**, 472 (1989).
[73] S. Boresch and M. Karplus, J. Chem. Phys. **105**, 5145 (1996).
[74] M. Sprik and G. Ciccotti, J. Chem. Phys. **109**, 7737 (1998).
[75] W. K. den Otter and W. J. Briels, J. Chem. Phys. **109**, 4139 (1998).
[76] W. K. den Otter, J. Chem. Phys. **112**, 7283 (2000).
[77] W. K. den Otter and W. J. Briels, Mol. Phys. **98**, 773 (2000).
[78] D. Trzesniak, A. P. E. Kunz, and W. F. van Gunsteren, ChemPhysChem **8**, 162 (2007).
[79] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[80] D. Frenkel and B. Smit, *Understanding Molecular Simulation* (Academic, Orlando, 2001).
[81] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2002).
[82] J. Drenth, *Principles of Protein X-Ray Crystallography* (Springer, New York, 1994).
[83] E. T. Jaynes, *Probability Theory : The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[84] W. Rieping, M. Habeck, and M. Nilges, Science **309**, 303 (2005).
[85] W. Rieping, M. Nilges, and M. Habeck, Bioinformatics **24**, 1104 (2008).
[86] P. L. Silvestrelli and M. Parrinello, J. Chem. Phys. **111**, 3572 (1999).
[87] K. Ichikawa, Y. Kameda, T. Yamaguchi, H. Wakita, and M. Misawa, Mol. Phys. **73**, 79 (1991).
[88] B. Zagrovic and V. S. Pande, J. Am. Chem. Soc. **128**, 11742 (2006).
[89] B. Zagrovic, Mol. Phys. **105**, 1299 (2007).
[90] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[91] H. S. Chan and K. A. Dill, J. Chem. Phys. **107**, 10353 (1997).
[92] H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990).
[93] W. J. Morokoff and R. E. Caflisch, J. Comput. Phys. **122**, 218 (1995).
[94] C. Schlier, Comput. Phys. Commun. **159**, 93 (2004).
[95] K. A. De Jong, *Evolutionary Computation. A unified Approach* (The MIT Press, Cambridge, 2006).