06.48

Trends in Physical Chemistry
Vol. 11, 2006

Review

Estimating the configurational entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the quasi-harmonic approximation

Riccardo Baron, Wilfred F. van Gunsteren and Philippe H. Hünenberger*
Laboratorium für Physikalische Chemie, ETH, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

ABSTRACT

During the past decades, the calculation of accurate free-energy differences from molecular simulations has become feasible in practice. In contrast, the reliable estimation of absolute entropies and entropy differences from these simulations is a notoriously difficult problem. This article investigates critically the method to estimate configurational entropies from molecular dynamics simulations based on the quasi-harmonic approximation. The theory, assumptions, and approximations underlying this method are presented, as well as its connection with essential-mode and normal-mode analyses. In particular, the following points are considered: (i) the relationship between quasi-harmonic and essential modes; (ii) the requirement of mass-weighting (or metric-tensor-weighting) in quasi-harmonic analysis; (iii) the effect of anharmonicities in the individual modes on the estimated entropy; (iv) the effect of pairwise (supralinear) correlations among the different modes on the estimated entropy. The analyses are carried out in the context of long (hundreds of nanoseconds) molecular dynamics simulations involving the reversible folding of β-peptides, considering individually the specific properties of the folded and unfolded ensembles. The anharmonicity correction to the quasi-harmonic entropy is small. In contrast, the pairwise (supralinear) correlation correction is large and affects to a larger extent the entropy of the folded state than that of the unfolded state. The proposed procedure to evaluate corrections for anharmonicity and correlation effects allows for an improved calculation of absolute entropies, as well as of entropy differences for molecular systems which undergo conformational transitions.

KEYWORDS: computer simulation; entropy; quasi-harmonic analysis; anharmonicity; correlation; peptide folding

1. INTRODUCTION

Entropy is a key property to the understanding of a wide variety of physical, chemical and biochemical phenomena [1-8]. In particular, entropic effects play a fundamental role in protein folding [9-15] and stability [16-19], molecular association [20-22], ligand binding [18,23-29], and enzyme catalysis [30]. Often, chemical and biochemical processes are characterized by a delicate balance between a large change in enthalpy and a large opposite change in entropy (enthalpy-entropy compensation), resulting in comparatively small free energy differences [4-10,12,16,20,24,29,31,32].

During the past decades, the calculation of accurate free-energy differences from molecular simulations has become feasible in practice [33-43]. In contrast, the reliable estimation of absolute entropies and entropy differences from these

*Corresponding author
phil@igc.phys.chem.ethz.ch

simulations is a notoriously difficult problem [41,44-53] and one of the key current challenges in computational chemistry.

In principle, the calculation of exact (within the force field and simulation methodology employed) absolute entropies from molecular simulations is an impossible task, because this quantity measures the overall extent of accessible phase space (*i.e.* would require infinitely long simulations to converge) [41,52]. There are two ways of circumventing this problem. The first approach is to restrict the task to the calculation of the entropy difference between two states of the system. In this case, appropriate generalizations of the coupling-parameter approach commonly used to evaluate free-energy differences can be used to estimate entropy differences via *e.g.* free-energy perturbation [52,54-56], thermodynamic integration [38,52,55], particle insertion [57,58], umbrella sampling [59], or finite-temperature differences [60]. In this case, appropriate sampling is only required for the regions of phase space where the Hamiltonians corresponding to the two states differ significantly. Unfortunately, this sampling still generally requires (series of) prohibitively long simulations and the method only provides converged results in favorable cases [41,52]. The second approach is to estimate the absolute entropy based on an analytical approximation to the configurational probability distribution of the system, expressed in a specified set of generalized coordinates. This analytical function involves parameters (*e.g.* moments of the distribution) that can in principle be estimated from a single (sufficiently long) simulation of the system.

This second approach was introduced by Karplus and Kushick [44], who applied it to bound (non-diffusive) systems (*e.g.* a single covalently linked macromolecule) on the basis of internal coordinates (subset of the distances, angles and torsional angles defining a configuration of the molecule, excluding overall translation and rotation) and with a multivariate Gaussian approximation for the configurational probability distribution (the parameters of which are the averages and covariances of the selected internal coordinates during the simulation). Because the normalized Gaussian is the probability distribution associated with a classical one-dimensional harmonic oscillator, this method was called quasi-harmonic analysis. In practice, however, the quasi-harmonic entropy is evaluated by application of the quantum-mechanical equation for the entropy of a harmonic oscillator, rather than the classical one. This point is essential in the presence of stiff degrees of freedom (*e.g.* bond-stretching or bond-angle bending vibrations, geometric constraints), because the quantum-mechanical entropy converges to zero in the limit of high oscillation frequencies, while the corresponding classical entropy incorrectly diverges to minus infinity. The quasi-harmonic approach was applied to biomolecular systems [16,44,45,61,62], and further extended by critical analyses of the approximations involved and of the resulting errors [45-47,53,61]. The application of the method as initially suggested by Karplus and Kushick [44] has a number of drawbacks: (i) the use of internal coordinates complicates the computational implementation of the method; (ii) an approximation must be made for the configurational dependence of the Jacobian associated with the internal-to-Cartesian coordinate transformation (metric-tensor effects); (iii) the choice of a non-redundant set of internal coordinates is not unique and affects the resulting entropy estimate; (iv) some internal coordinates (*e.g.* torsional angles) are periodic (*i.e.* their representation by a single-well harmonic oscillator is questionable and the associated averages and covariances cannot be unambiguously defined).

About ten years after the introduction of the quasi-harmonic approach, Schlitter suggested that the above problems could be alleviated by performing the analysis (still restricted to bound systems) in terms of Cartesian coordinates instead of internal ones [48]. He also realized that the method was capable of estimating absolute (rather than relative) entropies and suggested that, in the limit of infinite sampling, the estimated absolute entropy (*i.e.* that of a multi-dimensional system harmonic in a specified set of generalized coordinates and with metric-tensor-weighted coordinate averages and covariances identical to those monitored during the simulation) always provides an upper-bound to the true entropy of the simulated system. As shown in Appendix A, this latter statement is only rigorously exact at the classical level for a generalized coordinate system involving a configuration-independent Jacobian for the generalized-to-Cartesian coordinate

transformation (*e.g.* for a Cartesian coordinate system). For such coordinate systems, the statement remains valid at the quantum-mechanical level, although the proof given by Schlitter [48] is incomplete [63]. In other cases, this principle may be violated. However, the maximum entropy distribution will generally remain very close to that associated with a corresponding harmonic system and violations are only likely to be encountered for systems that are inherently very close to harmonicity, which is not the case of complex (macro)molecular systems. In the present article, the upper-bound property of the quasi-harmonic entropy estimate will be considered as being of general validity. Since the actual quasi-harmonic entropy estimate depends on the choice of a coordinate system (*i.e.* on the nature of the underlying harmonic model), the 'optimal' coordinate system will be the one leading to the lowest estimate [46]. Unfortunately, there is currently no systematic procedure for determining this optimal coordinate system.

In view of the previously mentioned drawbacks of internal coordinates (and of the difficulty in applying metric-tensor-weighting to these coordinates, see below), the use of Cartesian coordinates probably represents a good compromise. Yet, this choice also has three main drawbacks: (i) ambiguity may arise when removing the overall translational and rotational motions from the configurations sampled during the simulation of a (macro) molecule; (ii) the average molecular structure (equilibrium configuration in the harmonic model) may be less realistic when defined in terms of Cartesian coordinates (after removal of overall translational and rotational motions) compared to internal ones; (iii) the different potential energy terms of a force field are generally associated with specific internal coordinates, so that these coordinates appear to be a more 'natural' choice for the quasi-harmonic analysis. Indeed, calculations on model systems have shown that internal coordinates tend to result in lower (*i.e.* better) entropy estimates [53].

Irrespective of the chosen coordinate system, the difference between the true entropy of the simulated system and its quasi-harmonic estimate arises from: (i) anharmonicities (*i.e.* non-Gaussian behavior) in the probability distributions along individual coordinates; (ii) correlations among the probability distributions associated with different coordinates (beyond the pairwise linear correlations accounted for in the harmonic model). These effects are neglected in the quasi-harmonic analysis and (nearly; see Appendix A) always correspond to a negative entropy contribution. From a different perspective, the (positive) error associated with a quasi-harmonic entropy estimate may be viewed as arising from the approximation of the typically frustrated (many local minima and barriers) potential energy landscape characteristic of most molecular systems by a single (effective) harmonic basin [53]. In favourable cases, better entropy estimates can be obtained through methods that explicitly take into account the multiple-minima structure of this landscape (*e.g.* mining-minima approach [53,64]). However, in the context of biomolecular systems with explicit solvation, the enumeration of all minima (or even of only the most relevant ones) is not possible. In these cases, the quasi-harmonic approach (possibly including corrections such as the ones described in the present article) is probably the only option currently available.

The application of quasi-harmonic analysis in terms of Cartesian coordinates requires the removal of the overall (center of mass) translational motion from the sampled configurations (because the analysis assumes a bound system). If required, the translational entropy contribution can be later reintroduced using the appropriate quantum-mechanical expression (Sackur-Tetrode equation; based on a specified standard state for the system volume or pressure). Although the removal of the overall rotational motion is in principle not required, it is recommended in practice [49], because: (i) overall rotation is a highly correlated motion (associated with a relatively small entropy contribution), while the quasi-harmonic analysis interprets it as a superposition of uncorrelated harmonic motions along the individual Cartesian coordinates (associated with a much larger entropy contribution); (ii) the tumbling time of macro-molecules is generally long [65], so that rotational motions cannot be sampled accurately on the timescales reachable nowadays in explicit-solvent simulations of these systems. However, the overall rotation of a flexible molecule cannot be separated

from internal motions in a unique fashion [66-68], and this separation is always somewhat arbitrary [28,69,70]. In practice, the removal of overall translation and rotation from the sampled configurations is commonly performed by atom-positional least-squares fitting of successive structures along a trajectory onto a common reference structure [49,71,72]. Alternatively, simulations can be performed with constrained roto-translational motion [73]. If necessary, the resulting quasi-harmonic entropy estimate can be (approximately) corrected using the quantum-mechanical expression for the entropy of a rigid rotor (based on the average inertia tensor of the molecule).

The procedure outlined by Schlitter to evaluate quasi-harmonic entropies based on Cartesian coordinates [48] was implemented [49] and tested on simulations of reversible peptide folding in methanol at different temperatures [74]. The results showed that the solute quasi-harmonic entropy does not give sufficient information on the thermodynamics of the folding process, because: (i) the quasi-harmonic entropy estimate is an upper bound to the true entropy of the simulated peptide, both in the folded and unfolded states, but the magnitude of the associated error is unknown and may differ between the two states; (ii) solute-solvent and solvent-solvent correlations provide an unknown (but probably large) contribution to the entropy, which may also significantly differ between the folded and unfolded states. The same procedure was applied to investigate the influence of different solvents, folds and peptide chain lengths on the configurational entropy of carbopeptoids [75]. The results evidenced significant dependence of the estimated solute entropy on the properties of the solvent (in contrast to an earlier suggestion [16]) and on the characteristics of the folded configuration. In another study, the configurational entropy was calculated for trialanine in water and compared with vibrational spectroscopy experiments to interpret the stability of different peptide conformations [76]. The application of the Schlitter approach to fairly rigid organic compounds was also undertaken in order to estimate the translational and rotational contributions to the absolute entropy [28]. Its application to liquid hydrocarbons was undertaken to compare the configurational spaces accessible using simulation models of different spatial resolutions [77]. Applications to larger biomolecules involved the estimation of side-chain entropies in a model for a protein molten-globule state [78], as well as of the entropy changes upon binding a fatty acid to a fatty-acid binding protein [79], upon complexation of a protein with its receptor [80], upon binding of the CAP and λ-repressor to cognate DNA sequences [81], and upon binding of netropsin and distamycin ligands to the minor groove of a model DNA duplex [82].

The practical implementation of quasi-harmonic analysis using a specific set of generalized coordinates requires the diagonalization of the corresponding metric-tensor-weighted covariance matrix as evaluated from the simulation (in a Cartesian coordinate system, metric-tensor-weighting is equivalent to mass-weighting). The resulting eigenvectors represent motional modes (quasi-harmonic modes) around the average system configuration, associated with probability distributions that have zero averages and are pairwise linearly uncorrelated. Furthermore, if the configurational probability distribution of the system is approximated by a multivariate Gaussian with the same (metric-tensor-weighted) covariance matrix as the real distribution, the probability distributions associated with the quasi-harmonic modes are Gaussian (of widths determined by the corresponding eigenvalues) and fully uncorrelated. In this case, the approximate system entropy (quasi-harmonic estimate) is a sum of harmonic entropy contributions associated with each of the quasi-harmonic modes. These contributions can be calculated analytically based on the corresponding eigenvalues, using the appropriate classical or (preferably) quantum-mechanical expression for the entropy of a one-dimensional harmonic oscillator. As suggested by Schlitter [48], the diagonalization process can be substituted by a determinant calculation if the correct quantum-mechanical formula is replaced by an approximate heuristic expression (that slightly overestimates the quasi-harmonic entropy). However, since the computational gain is only moderate and the quasi-harmonic modes can no longer be determined, the use of this approximate formula is probably not recommended in practice [50].

As mentioned above (and more explicitly formulated in Section 2) the diagonalization process in a quasi-harmonic analysis must be applied to the metric-tensor-weighted covariance matrix. However, previous applications of this approach in terms of internal coordinates [16,44-47,61,62] incorrectly omitted this weighting, leading to an error in the estimated entropy (the nature and magnitude of which is difficult to characterize) and to an inconsistency in the dimensionality of the estimated entropy (dependence on the arbitrary choice for the unit of mass).

The above procedure for quasi-harmonic analysis bears strong analogies with the essential-mode analysis developed by the Berendsen group [83-86]. The aim of this approach is to decompose the atom-positional fluctuations of a (bound) molecular system as obtained from a simulation into pairwise linearly uncorrelated motional modes (essential modes), associated with additive contributions to the total mean-square atom-positional fluctuations. In fact, essential-mode analysis is equivalent to a quasi-harmonic analysis applied in Cartesian coordinates, but without mass-weighting of the covariance matrix prior to diagonalization. However, to our knowledge, the relationship between quasi-harmonic and essential modes (and between the probability distributions along these modes) has never been investigated so far.

Finally, both quasi-harmonic and essential-mode analyses should be distinguished from normal-mode analysis [87-97]. This analysis, typically performed in terms of Cartesian coordinates (see, however, Section 2.1), relies on evaluating and diagonalizing the mass-weighted Hessian matrix (second derivative of the potential energy with respect to the atomic coordinates) associated with a single structure (usually corresponding to a stationary point on the potential-energy surface). Normal-mode analysis probes the vibrational modes associated with this single structure, *i.e.* a very local region of the configurational space, while quasi-harmonic and essential-mode analyses aim at characterizing the global extent of the configurational space accessible to the system at a given temperature (figure 1). Therefore, an entropy estimate based on normal-mode analysis

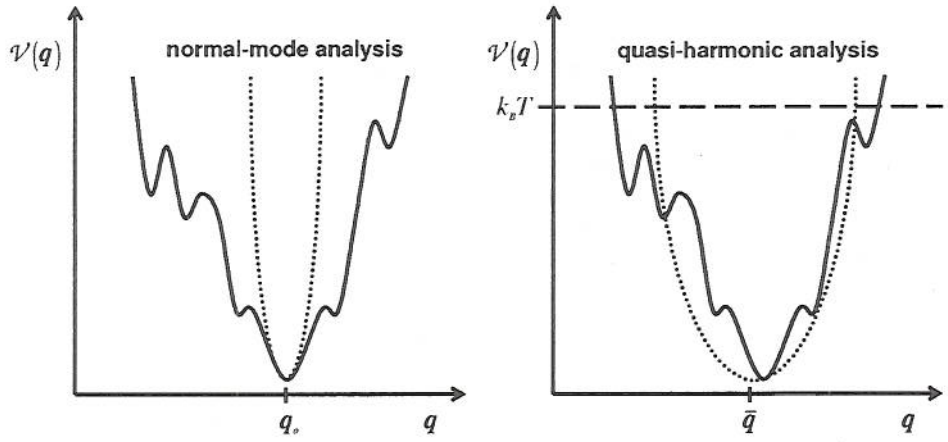

Figure 1. Schematic comparison of the principles underlying normal-mode (left) and quasi-harmonic (right) analyses. Normal-mode analysis probes the local configurational space (q) around a single (usually stationary) point (q_o), based on the local curvature of the potential energy surface ($\mathcal{V}(q)$) at this point. Quasi-harmonic analysis probes the global extent of configurational space (q) accessible to a molecular system at a given temperature (k_BT).

is only meaningful for systems that are intrinsically highly harmonic (*e.g.* stiff molecules in the gas phase), but certainly not for solvated (macro)molecules.

The estimation of absolute entropies through quasi-harmonic analysis is very appealing for five main reasons: (i) absolute entropies rather than entropy differences can be estimated; (ii) the entropy estimate is based on a single simulation of the molecular system; (iii) the entropy of a bound system of N atoms (exempt of overall translation and rotation) can be decomposed into an ideal gas contribution plus a sum of $3N - 6$ configurational contributions arising from independent internal motional modes, thereby providing insight into its nature; (iv) the estimated entropy (nearly; see Appendix A) always represents an upper bound to the true entropy of the simulated system; (v) the entropy estimate, even if originally based on a classical simulation, can be performed using an exact quantum-mechanical formula rather than a classical approximation. However, this approach also has two major shortcomings: (i) it is currently restricted to bound (non-diffusive) systems, excluding the determination of entropy contributions associated with intermolecular (*e.g.* solute-solvent or solvent-solvent) correlations; (ii) it provides upper-bound entropies with an error that depends on the (arbitrary) choice of a generalized coordinate system and is rather difficult to quantify. The second point is particularly problematic when trying to estimate entropy differences between two states of a molecular system (*e.g.* B and Z forms of DNA [62] or folded and unfolded states of peptides [74]), because error cancellation is by no means guaranteed.

The present article addresses a number of issues related to the quasi-harmonic approach. In particular, it aims at: (i) analyzing the relationship between quasi-harmonic and essential modes; (ii) evaluating critically the requirement of mass-weighting (or metric-tensor-weighting) in quasi-harmonic analysis; (iii) quantifying the effect of anharmonicities in the individual modes on the estimated entropy; (iv) quantifying the effect of pairwise supralinear correlations among the different modes on the estimated entropy. This analysis is carried out in the context of long (hundreds of nanoseconds) molecular dynamics simulations involving the reversible folding of β-peptides in methanol [98-103], and considers individually the specific properties of the ensembles corresponding to folded and unfolded configurations.

2. Theory

2.1 Normal-mode analysis

Consider a classical bound (non-diffusive) system involving M' degrees of freedom, of which M are unconstrained and specified by a M - dimensional generalized-coordinate vector $\boldsymbol{q} = \{q_m \mid m = 1..M\}$. The $M' \times M$ - dimensional Jacobian matrix $\underline{\mathbf{J}}_q$ relating infinitesimal generalized-coordinate displacements $d\boldsymbol{q}$ to the corresponding Cartesian displacements $d\boldsymbol{r}$ is defined as

$$\underline{\mathbf{J}}_q(\boldsymbol{q}) = \frac{d\boldsymbol{r}(\boldsymbol{q}, d\boldsymbol{q})}{d\boldsymbol{q}} \tag{1}$$

where the $\boldsymbol{q}$ subscript indicates a dependence on the specific choice of generalized coordinate system.

If the potential energy solely depends on the system configuration, the M-dimensional conjugate-momentum vector $\boldsymbol{p}_q$ associated with $\boldsymbol{q}$ (derivative of the Lagrangian with respect to $\dot{\boldsymbol{q}}$) is given by

$$\boldsymbol{p}_q = \underline{\mathbf{A}}_q(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{2}$$

with the $M \times M$ - dimensional (symmetric) mass-metric tensor $\underline{\mathbf{A}}_q$ (further simply referred to as the metric tensor) defined as

$$\underline{\mathbf{A}}_q(\boldsymbol{q}) = \underline{\mathbf{J}}_q^T(\boldsymbol{q})\,\underline{\mathbf{M}}\,\underline{\mathbf{J}}_q(\boldsymbol{q}) \tag{3}$$

where $\underline{\mathbf{M}}$ is the diagonal $M' \times M'$ - dimensional matrix containing the masses associated with the Cartesian degrees of freedom and the T superscript indicates the transpose of a vector or matrix.

The corresponding kinetic energy is given by

$$\mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_q) = \frac{1}{2}\underline{\boldsymbol{p}}_q^T\,\underline{\mathbf{A}}_q^{-1}(\boldsymbol{q})\,\underline{\boldsymbol{p}}_q \tag{4}$$

Normal-mode analysis assumes that the neighborhood of the potential-energy surface around

a stationary point q_o can be approximated by the multivariate harmonic function (h subscript)

$$\mathcal{V}_h(q) = \frac{1}{2}(q-q_o)^T \underline{\mathbf{H}}_q (q-q_o) \tag{5}$$

where $\underline{\mathbf{H}}_q$ is the $M \times M$ - dimensional (symmetric) Hessian matrix containing the second derivatives of the potential energy with respect to any pair of generalized coordinates at q_o. From eqs. (4) and (5), the Hamiltonian of the system in the neighbourhood of q_o is therefore approximated as

$$\mathcal{H}_h(q,p_q) = \frac{1}{2}(q-q_o)^T \underline{\mathbf{H}}_q (q-q_o) + \frac{1}{2} p_q^T \underline{\mathbf{A}}_q^{-1}(q)\, p_q \tag{6}$$

The corresponding (normalized) configurational probability distribution in the canonical ensemble is given by (see Appendix B; eq. (B8))

$$p_h(q) = \frac{\int dp_q\, e^{-\beta \mathcal{H}_h(q,p_q)}}{\int dq\, dp_q\, e^{-\beta \mathcal{H}_h(q,p_q)}} = \frac{\left|\underline{\mathbf{A}}_q(q)\right|^{1/2} e^{-\frac{1}{2}\beta (q-q_o)^T \underline{\mathbf{H}}_q (q-q_o)}}{\int dq \left|\underline{\mathbf{A}}_q(q)\right|^{1/2} e^{-\frac{1}{2}\beta (q-q_o)^T \underline{\mathbf{H}}_q (q-q_o)}} \tag{7}$$

where $\beta = (k_B T)^{-1}$, k_B being the Boltzmann's constant and T the absolute temperature. This distribution is generally complicated by the presence of the configuration-dependent quantity $\left|\underline{\mathbf{A}}_q\right|^{1/2}$ related to metric-tensor effects. This difficulty can be overcome by assuming that, in the close neighborhood of the potential energy surface surrounding q_o, the matrix $\underline{\mathbf{A}}_q$ is nearly configuration-independent, *i.e.*

$$\underline{\mathbf{A}}_q(q) \approx \underline{\mathbf{A}}_q(q_o) = \underline{\mathbf{A}}_q \tag{8}$$

which is reasonable for small oscillations around an equilibrium configuration. In this case, the Hamiltonian $\mathcal{H}_h$ of eq. (6) may be approximated by a Hamiltonian $\mathcal{H}_o$ defined as

$$\mathcal{H}_o(q,p_q) = \frac{1}{2}(q-q_o)^T \underline{\mathbf{H}}_q (q-q_o) + \frac{1}{2} p_q^T \underline{\mathbf{A}}_q^{-1} p_q \tag{9}$$

The corresponding configurational probability distribution in the canonical ensemble (see eq. (7)) can be written

$$p_o(q) = \frac{e^{-\frac{1}{2}\beta (q-q_o)^T \underline{\mathbf{H}}_q (q-q_o)}}{\int dq\, e^{-\frac{1}{2}\beta (q-q_o)^T \underline{\mathbf{H}}_q (q-q_o)}} \tag{10}$$

Now, consider the coordinate transformation

$$a_q = \underline{\mathbf{U}}_q^T \underline{\mathbf{A}}_q^{1/2} (q-q_o) \tag{11}$$

where $\underline{\mathbf{U}}_q$ is the $M \times M$ - dimensional (orthogonal, *i.e.* with $\underline{\mathbf{U}}^T = \underline{\mathbf{U}}^{-1}$) matrix diagonalizing the (symmetric) metric-tensor-weighted inverse Hessian matrix, *i.e.*

$$\underline{\mathbf{U}}_q^T \underline{\mathbf{A}}_q^{1/2} \underline{\mathbf{H}}_q^{-1} \underline{\mathbf{A}}_q^{1/2} \underline{\mathbf{U}}_q = \beta \underline{\mathbf{E}}_q \tag{12}$$

$\underline{\mathbf{E}}_q$ being the corresponding $M \times M$ - dimensional (diagonal) eigenvalue matrix amplified by β^{-1}. To make the definitions of $\underline{\mathbf{U}}_q$ and $\underline{\mathbf{E}}_q$ unique (except for a possible degeneracy in the eigenvalues), it is further specified that: (i) the entry of largest absolute value in any column of $\underline{\mathbf{U}}_q$ is a positive number; (ii) the eigenvalues occur along the diagonal of $\underline{\mathbf{E}}_q$ in decreasing order. This convention will be applied for all subsequent diagonalization processes in the present article (see eqs. (25) and (32)).

The generalized momenta associated with the new coordinates (derivative of the Lagrangian with respect to $\dot{a}_q$) are given by

$$p_a = \dot{a}_q = \underline{\mathbf{U}}_q^T \underline{\mathbf{A}}_q^{-1/2} p_q \tag{13}$$

where the notation p_{a_q} has been simplified to p_a.

Thus, the Hamiltonian of eq. (9) can be rewritten in terms of the transformed coordinate system as

$$\mathcal{H}'_o(a_q, p_a) = \frac{1}{2}\beta^{-1} a_q^T \underline{\mathbf{E}}_q^{-1} a_q + \frac{1}{2} p_a^T p_a \tag{14}$$

By writing the corresponding Hamiltonian equations of motion it is easily seen that this system corresponds to a collection of independent harmonic oscillators with angular frequencies

$$\omega_m = \left(\beta E_{q,m}\right)^{-1/2}, \quad m = 1,2,...,M \tag{15}$$

The associated configurational probability distribution in the canonical ensemble can be written as (see eq. (10))

$$p'_o(a_q) = \frac{e^{-\frac{1}{2}a_q^T \underline{\mathbf{E}}_q^{-1} a_q}}{\int da_q\, e^{-\frac{1}{2}a_q^T \underline{\mathbf{E}}_q^{-1} a_q}} = \prod_{m=1}^{3f} p'_{o,m}(a_{q,m}) \tag{16}$$

with

$$p'_{o,m}(a_{q,m}) = \left(2\pi E_{q,m}\right)^{-1/2} e^{-\frac{1}{2}E_{q,m}^{-1} a_{q,m}^2} \tag{17}$$

i.e. it factorizes into a product of normalized origin-centered Gaussians along each of the coordinates.

The eigenvectors contained in the columns of $\underline{\mathbf{U}}_q$ can be referred to as the generalized normal modes associated with the selected generalized coordinate system q and the corresponding model Hamiltonian $\mathcal{H}_o$. The corresponding transformed coordinates a_q will be referred to as generalized normal coordinates, and satisfy the properties

$$\langle a_q \rangle = \int da_q\, a_q\, p'_o(a_q) = 0 \tag{18}$$

and

$$\langle a_q \otimes a_q \rangle = \int da_q \left(a_q \otimes a_q\right) p'_o(a_q) = \underline{\mathbf{E}}_q \tag{19}$$

where the angular brackets indicate canonical ensemble averaging and the notation $a \otimes b$ is used for the matrix with components μ,ν equal to $a_\mu b_\nu$. Because $\underline{\mathbf{E}}_q$ is diagonal, the second equation implies in particular that the motions along distinct generalized normal modes are pairwise linearly uncorrelated. These results can be rewritten in terms of the original coordinate system as

$$\overline{q} = \langle q \rangle = q_o + \underline{\mathbf{A}}_q^{-1/2}\, \underline{\mathbf{U}}_q \langle a_q \rangle = q_o \tag{20}$$

and

$$\begin{aligned} \underline{\mathbf{C}}_q &= \langle (q - q_o) \otimes (q - q_o) \rangle \\ &= \left\langle \left(\underline{\mathbf{A}}_q^{-1/2}\, \underline{\mathbf{U}}_q\, a_q\right) \otimes \left(\underline{\mathbf{A}}_q^{-1/2}\, \underline{\mathbf{U}}_q\, a_q\right) \right\rangle \\ &= \underline{\mathbf{A}}_q^{-1/2}\, \underline{\mathbf{U}}_q \langle a_q \otimes a_q \rangle \underline{\mathbf{U}}_q^T\, \underline{\mathbf{A}}_q^{-1/2} \\ &= \underline{\mathbf{A}}_q^{-1/2}\, \underline{\mathbf{U}}_q\, \underline{\mathbf{E}}_q\, \underline{\mathbf{U}}_q^T\, \underline{\mathbf{A}}_q^{-1/2} \\ &= \beta^{-1}\, \underline{\mathbf{H}}_q^{-1} \end{aligned} \tag{21}$$

where $\overline{q}$ and $\underline{\mathbf{C}}_q$ are the average and the covariance matrix associated with the original coordinate system, respectively, the third equality following from $(\underline{\mathbf{A}}\, a) \otimes (\underline{\mathbf{B}}\, b) = \underline{\mathbf{A}}\,(a \otimes b)\,\underline{\mathbf{B}}^T$. The above results (eqs. (9)-(21)) hold within the harmonic approximation of eq. (5) and are limited by the validity of eq. (8).

Although the above formulation is quite general, normal-mode analysis is typically only performed in terms of the Cartesian displacements r relative to a reference (stationary) point at $r = 0$ (as a particular case of the generalized coordinates q). In this specific case, $\underline{\mathbf{J}}_r(r) = \underline{\mathbf{J}}_r = \underline{\mathbf{1}}$ in eq. (1) and $\underline{\mathbf{A}}_r(r) = \underline{\mathbf{A}}_r = \underline{\mathbf{M}}$ in eq. (3), so that eq. (8) is exactly satisfied and the Hamiltonians $\mathcal{H}_h$ and $\mathcal{H}_o$ (eqs. (6) and (9)) become indistinguishable. The corresponding generalized normal modes and generalized normal coordinates (the latter with units of mass$^{1/2}$ × length) are then simply referred to as normal modes and normal coordinates, respectively.

2.2 Quasi-harmonic analysis

Although the mathematical formalism employed in quasi-harmonic analysis is very similar to that employed in normal-mode analysis, the underlying principle is different (figure 1). While normal-mode analysis attempts to describe the local area of the potential energy surface in the neighborhood of a stationary point, quasi-harmonic analysis tries to account for motions in the overall extent of configurational space accessible to a molecular system (on the simulation timescale and at a given temperature). For a given choice of generalized coordinate system q, the input quantity of a

normal-mode analysis is the Hessian matrix $\underline{\mathbf{H}}_q$ characterizing the curvature of the potential energy surface at a stationary point q_o, as evaluated from the force field employed. In contrast, the input quantity of a quasi-harmonic analysis is the covariance matrix $\underline{\mathbf{C}}_q$ characterizing the coordinate fluctuations around an average configuration $\overline{q}$, as obtained from a simulation of the molecular system (of a given duration and at a given constant temperature). Note that a meaningful determination of $\overline{q}$ and $\underline{\mathbf{C}}_q$ requires that the simulated trajectory contains at least $M+1$ configurations (which will be assumed from here on).

The covariance matrix describes the motions within a (macro)molecule in the form of atom-positional fluctuations and their correlations [74,83,88,93,104]. Assuming that these fluctuations result from an underlying harmonic potential of the form

$$\tilde{\mathcal{V}}_h(q) = \frac{1}{2}(q-\tilde{q}_o)^T\,\underline{\tilde{\mathbf{H}}}_q\,(q-\tilde{q}_o) \qquad (22)$$

where $\underline{\tilde{\mathbf{H}}}_q$ is an effective Hessian matrix and $\tilde{q}_o$ an effective equilibrium configuration, eqs. (20) and (21) show immediately that

$$\tilde{q}_o = \overline{q} \quad \text{and} \quad \underline{\tilde{\mathbf{H}}}_q = \beta^{-1}\underline{\mathbf{C}}_q^{-1} \qquad (23)$$

Note that the corresponding harmonic model only strictly produces the correct average configuration $\overline{q}$ and covariance matrix $\underline{\mathbf{C}}_q$ when eq. (8) is satisfied, *i.e.* for generalized coordinate systems where the metric tensor $\underline{\mathbf{A}}_q$ is configuration-independent. This is the case when employing Cartesian coordinates, but in general not for internal ones. Furthermore, in the latter case, the configurational dependence of the metric tensor is likely to be more pronounced in quasi-harmonic analysis compared to normal-mode analysis due to the larger extent of configurational space covered by the model.

In practice, quasi-harmonic analysis involves the following steps.

First, a generalized coordinate system q is selected.

Second, the average configuration $\overline{q}$ and the covariance matrix $\underline{\mathbf{C}}_q$ in this coordinate system are evaluated from a (sufficiently long) molecular simulation, as

$$\overline{q} = \langle q \rangle \text{ and } \underline{\mathbf{C}}_q = \langle (q-\overline{q}) \otimes (q-\overline{q}) \rangle \qquad (24)$$

The equilibrium configuration $\tilde{q}_o$ and Hessian matrix $\underline{\tilde{\mathbf{H}}}_q$ of the effective underlying harmonic model are then defined according to eq. (23), assuming the (exact or approximate) validity of eq. (8).

Third, the (symmetric) metric-tensor-weighted covariance matrix is diagonalized, *i.e.* (compare with eq. (12))

$$\underline{\mathbf{V}}_q^T\,\underline{\mathbf{A}}_q^{1/2}\,\underline{\mathbf{C}}_q\,\underline{\mathbf{A}}_q^{1/2}\,\underline{\mathbf{V}}_q = \underline{\mathbf{F}}_q \qquad (25)$$

where $\underline{\mathbf{V}}_q$ is a $M \times M$ - dimensional (orthogonal) matrix the columns of which represent the M components of the eigenvectors $\{v_{q,m} \mid m = 1..M\}$ (called quasi-harmonic modes) of the metric-tensor-weighted covariance matrix in the original coordinate system, and $\underline{\mathbf{F}}_q$ is a diagonal matrix containing the corresponding eigenvalues. These eigenvalues are related to the associated angular frequencies in the effective underlying harmonic model as (see eqs. (12), (15), (23) and (25))

$$\omega_m = \left(\beta F_{q,m}\right)^{-1/2}, \quad m = 1,2,...,M \qquad (26)$$

Fourth, the simulated trajectory is projected onto the quasi-harmonic modes, *i.e.* one considers the transformed coordinates b_q defined as (compare with eq. (11))

$$b_q = \underline{\mathbf{V}}_q^T\,\underline{\mathbf{A}}_q^{1/2}\,(q-\overline{q}) \qquad (27)$$

These transformed coordinates will be referred to as quasi-harmonic coordinates, and satisfy the properties (compare with eqs. (18) and (19))

$$\langle b_q \rangle = \left\langle \underline{\mathbf{V}}_q^T\,\underline{\mathbf{A}}_q^{1/2}\,(q-\overline{q}) \right\rangle = \mathbf{0} \qquad (28)$$

and

$$\begin{aligned} &\langle b_q \otimes b_q \rangle \\ &= \left\langle \left[\underline{\mathbf{V}}_q^T\,\underline{\mathbf{A}}_q^{1/2}\,(q-\overline{q})\right] \otimes \left[\underline{\mathbf{V}}_q^T\,\underline{\mathbf{A}}_q^{1/2}\,(q-\overline{q})\right] \right\rangle \\ &= \underline{\mathbf{V}}_q^T\,\underline{\mathbf{A}}_q^{1/2}\,\underline{\mathbf{C}}_q\,\underline{\mathbf{A}}_q^{1/2}\,\underline{\mathbf{V}}_q = \underline{\mathbf{F}}_q \end{aligned} \qquad (29)$$

Because $\underline{\mathbf{F}}_q$ is diagonal, the second equation implies that the projected coordinates b_q are pairwise linearly uncorrelated which, however, does not imply the absence of higher-order (*i.e.* pairwise supralinear or beyond pairwise) correlations. The sum of the eigenvalues in $\underline{\mathbf{F}}_q$ is equal to the total mean-square metric-tensor-weighted fluctuation of the system, *i.e.*

$$Tr\left[\underline{\mathbf{F}}_q\right] = Tr\left[\underline{\mathbf{A}}_q^{\dagger/2}\,\underline{\mathbf{C}}_q\,\underline{\mathbf{A}}_q^{\dagger/2}\right] = \left\langle\left[\underline{\mathbf{A}}_q^{\dagger/2}\left(q-\overline{q}\right)\right]\cdot\left[\underline{\mathbf{A}}_q^{\dagger/2}\left(q-\overline{q}\right)\right]\right\rangle \tag{30}$$

so that the eigenvalues can be interpreted as contributions of the individual quasi-harmonic modes to this quantity.

In the particular case where the analysis is performed in a Cartesian coordinate system r (after removal of the overall translational and rotational motion from the sampled trajectory), one has $\underline{\mathbf{A}}_r(r) = \underline{\mathbf{A}}_r = \underline{\mathbf{M}}$ in eq. (3). Thus, eq. (8) and, consequently, eq. (23) are exactly satisfied. In this case, the analysis relies on the diagonalization of the mass-weighted Cartesian covariance matrix (see eq. (25))

$$\underline{\mathbf{D}}_r = \underline{\mathbf{M}}^{\dagger/2}\,\underline{\mathbf{C}}_r\,\underline{\mathbf{M}}^{\dagger/2} \tag{31}$$

In the absence of geometric constraints, the corresponding eigenvalue matrix $\underline{\mathbf{F}}_r$ contains $3N-6$ non-zero and 6 vanishing (or quasi-vanishing) elements. If geometrical constraints are present in the system (*e.g.* bond-length constraints), these will map to a corresponding number of zero eigenvalues (see Appendix A in ref. 83). Note that in this case, the quasi-harmonic coordinates have units of mass$^{1/2}$ × length.

Previous applications of the quasi-harmonic analysis in terms of internal coordinates [16, 44-47,61,62] omitted the metric-tensor weighting in eqs. (25) and (27). This modification (which essentially leads to an essential-mode rather than a quasi-harmonic analysis, see Section 2.3) is incorrect, and leads in particular to results that are inconsistent in their dimensionality. This is easily seen *e.g.* from eqs. (25) and (26). If the units of q are noted 'coord', $\underline{\mathbf{C}}_q$ has units of coord^2, $\underline{\mathbf{A}}_q$ has units of mass × length^2 × coord^{-2} (eqs. (1) and (3)), β has units of mass^{-1} × length^{-2} × time^2 and $\underline{\mathbf{V}}_q$ is unitless. In this case, ω cannot represent an angular frequency with units of time^{-1} if the metric-tensor weighting is omitted in eq. (25). However, even if metric-tensor-weighting was properly applied (which may be cumbersome to implement in practice), the validity of the analysis would still be restricted by the approximate nature of eq. (8). On the other hand, when internal coordinates are used, center of mass (translational and rotational) motion as well as possible geometrical constraints are automatically mapped out of the analysis.

As discussed in Section 2.4, the main interest of quasi-harmonic analysis resides in its connection with the estimation of the absolute entropy of the simulated system.

2.3 Essential-mode analysis

The quasi-harmonic analysis as described above (Section 2.2) bears strong similarities with essential-mode (or principal-component) analysis [83-87,105-109]. However, while the former method aims at analyzing a simulated trajectory in terms of an underlying effective harmonic model, the latter method attempts to decompose the system fluctuations into independent (pairwise linearly uncorrelated) motional modes with additive contributions to the system total mean-square fluctuations. To this purpose, it is the covariance matrix rather than the metric-tensor-weighted covariance matrix that is diagonalized. In the context of biomolecules, the result of this analysis is usually interpreted under the assumptions that: (i) the modes with the largest contributions to the total system fluctuations are also the most functionally relevant; (ii) the modes associated with comparatively lower eigenvalues (*i.e.* providing a negligible contribution to the overall motion) are characterized by simple unimodal probability distributions (Gaussian-like) and largely uncorrelated among each other and with the more relevant modes. Under these assumptions, the outcome of an essential-mode analysis can be used to perform more efficient simulations in a reduced space where the least relevant modes are treated as constraints, thereby

eliminating high-frequency motions and enabling the use of a longer time step (*e.g.* essential dynamics [83]; it is debatable whether quasi-harmonic analysis would not be also better suited for this purpose, see *e.g.* ref. 109). However, the application of this approach is limited in practice by two problems: (i) the questionable validity of the above assumptions (in particular, some low amplitude motions in biomolecules may be functionally very relevant); (ii) the introduction of constraints, leading to thermodynamical (metric-tensor-related) as well as dynamical (enhanced torsional barriers) artifacts [111-113]. For details about the latter point, the reader is referred to Appendix B in ref. 83.

In analogy with eq. (25), essential-mode analysis relies on diagonalizing the (symmetric) covariance matrix $\underline{\mathbf{C}}_q$ as

$$\underline{\mathbf{W}}_q^T \, \underline{\mathbf{C}}_q \, \underline{\mathbf{W}}_q = \underline{\mathbf{G}}_q \tag{32}$$

where $\underline{\mathbf{W}}_q$ is a $M \times M$ - dimensional (orthogonal) matrix the columns of which represent the M components of the eigenvectors $\{\boldsymbol{w}_{q,m} \mid m = 1..M\}$ (referred to here as essential modes) of the covariance matrix in the original coordinate system, and $\underline{\mathbf{G}}_q$ is a diagonal matrix containing the corresponding eigenvalues. These eigenvalues are related to the associated contributions to the total mean-square fluctuation (see below). A set of transformed coordinates $\boldsymbol{c}_q$ is then defined as the projection of the original coordinate deviations in the basis of the essential modes [83], *i.e.* as

$$\boldsymbol{c}_q = \underline{\mathbf{W}}^T \left(\boldsymbol{q} - \overline{\boldsymbol{q}} \right) \tag{33}$$

These projected coordinates will be referred to as essential coordinates, and satisfy the properties (compare with eqs. (28) and (29))

$$\left\langle \boldsymbol{c}_q \right\rangle = \left\langle \underline{\mathbf{W}}_q^T \left(\boldsymbol{q} - \overline{\boldsymbol{q}} \right) \right\rangle = \mathbf{0} \tag{34}$$

and

$$\begin{aligned} &\left\langle \boldsymbol{c}_q \otimes \boldsymbol{c}_q \right\rangle \\ &= \left\langle \left[\underline{\mathbf{W}}_q^T \left(\boldsymbol{q} - \overline{\boldsymbol{q}} \right) \right] \otimes \left[\underline{\mathbf{W}}_q^T \left(\boldsymbol{q} - \overline{\boldsymbol{q}} \right) \right] \right\rangle \\ &= \underline{\mathbf{W}}_q^T \, \underline{\mathbf{C}}_q \, \underline{\mathbf{W}}_q = \underline{\mathbf{G}}_q \end{aligned} \tag{35}$$

Because $\underline{\mathbf{G}}_q$ is diagonal, the second equation implies that the projected coordinates $\boldsymbol{c}_q$ are pairwise linearly uncorrelated. Finally, the sum of the eigenvalues in $\underline{\mathbf{G}}_q$ is equal to the total mean-square fluctuation (MSF) of the system, *i.e.* (compare with eq. (30))

$$\begin{aligned} Tr\left[\underline{\mathbf{G}}_q\right] &= Tr\left[\underline{\mathbf{C}}_q\right] = \left\langle \left(\boldsymbol{q} - \overline{\boldsymbol{q}} \right)^2 \right\rangle \\ &= \left\langle \boldsymbol{q}^2 \right\rangle - \overline{\boldsymbol{q}}^2 \end{aligned} \tag{36}$$

so that the eigenvalues themselves can be interpreted as contributions of the individual essential modes to the total mean-square fluctuation. In the literature (*e.g.* refs. 83-87,105-109), essential-mode analysis has always been applied in terms of Cartesian coordinates (after removal of overall translation and rotation). In this specific case, the elements of $\underline{\mathbf{G}}_q$ represent contributions to the total Cartesian mean-square atom-positional fluctuation, and the essential coordinates have units of length.

In spite of the striking similarities in the definitions of quasi-harmonic and essential modes, the two sets of eigenvectors are not related in any simple way. Although eqs. (25) and (32) suggest that the matrix $\underline{\mathbf{V}}'_q = \mathbf{A}_q^{-1/2} \underline{\mathbf{W}}_q$ diagonalizes $\underline{\mathbf{A}}_q^{1/2} \, \underline{\mathbf{C}}_q \, \underline{\mathbf{A}}_q^{1/2}$, this matrix is generally not orthogonal. However, even if there is no simple relation between $\underline{\mathbf{V}}_q$ and $\underline{\mathbf{W}}_q$, the projections of the trajectory onto either quasi-harmonic ($\boldsymbol{b}_q$) or essential ($\boldsymbol{c}_q$) modes appear to be heavily correlated (see Section 4.1).

2.4 Estimation of the entropy

Entropy estimates may be performed based on either the normal-mode or the quasi-harmonic mode analyses described above (Sections 2.1 and 2.2). The application to normal-mode analysis is detailed below as an example. The adaptation of the formalism to quasi-harmonic analysis is then briefly described.

Based on an arbitrary generalized coordinate system $\boldsymbol{q}$ (and the associated conjugate momenta $\boldsymbol{p}_q$), the classical entropy of a (bound) system of $M'/3$ particles with M unconstrained degrees of freedom may be evaluated as [45,46].

$$S_{cl} = -k_B \int dq\, dp_q\ \tilde{p}(q, p_q) \ln\left[\xi h^M\ \tilde{p}(q, p_q)\right] \quad (37)$$

where h is Planck's constant, $\tilde{p}$ the (normalized) phase-space probability distribution, and ξ a factor related to the indistinguishability of the particles. If the particles are distinguishable (*e.g.* individual atoms in a covalently bound molecule) one has $\xi = 1$. If they are indistinguishable (*e.g.* atoms in a monoatomic fluid) one has $\xi = (M'/3)!$. The $M' - M$ constrained degrees of freedom (*e.g.* fixed center of mass position, molecular orientation, and geometrical constraints) are excluded from eq. (37) because they would cause the divergence of the calculated entropy to minus infinity.

In the canonical ensemble, the integration of eq. (37) over the momenta can be carried out analytically (Appendix B), leading to

$$\begin{aligned} S_{cl} = k_B \Bigg[& \frac{M}{2}\left(1 - \ln\frac{\beta h^2}{2\pi}\right) - \ln\xi \\ & + \frac{1}{2}\int dq\ p(q)\ \ln\left|\underline{\mathbf{A}}_q(q)\right| \\ & - \int dq\ p(q) \ln p(q) \Bigg] \end{aligned} \quad (38)$$

where p is the (normalized) configurational probability distribution and $\underline{\mathbf{A}}_q(q)$ the metric tensor introduced in eq. (3).

Within the harmonic approximation (eq. (5)) and assuming that the metric tensor is configuration-independent (eq. (8)), the latter expression can be applied to the generalized normal modes a_q (eq. (11)) as a particular case of q (Appendix B), leading to

$$\begin{aligned} S_{cl,o} = k_B \Bigg\{ & M\left(1 - \ln\beta\hbar\right) - \ln\xi \\ & - \sum_{m=1}^{M} \ln\left[\left(\beta E_{q,m}\right)^{-1/2}\right] \Bigg\} \end{aligned} \quad (39)$$

where $\hbar = (2\pi)^{-1} h$ and $\underline{\mathbf{E}}_q$ is the diagonal eigenvalue matrix (amplified by β^{-1}) introduced in eq. (12). This equation can be rewritten as

$$S_{cl,o} = -k_B \ln\xi + \sum_{m=1}^{M} s_{cl,o}\left(\left(\beta E_{q,m}\right)^{-1/2}\right) \quad (40)$$

where $s_{cl,o}(\omega)$ is the classical entropy of a single harmonic degree of freedom oscillating at an angular frequency ω, *i.e.*

$$s_{cl,o}(\omega) = k_B\left(1 - \ln\beta\hbar\omega\right) \quad (41)$$

To this classical expression, one may substitute the corresponding (more accurate) quantum-mechanical expression

$$s_{qm,o}(\omega) = k_B\left[\frac{\beta\hbar\omega}{e^{\beta\hbar\omega} - 1} - \ln\left(1 - e^{-\beta\hbar\omega}\right)\right] \quad (42)$$

leading to a total quantum-mechanical entropy

$$S_{qm,o} = -k_B \ln\xi + \sum_{m=1}^{M} s_{qm,o}\left(\left(\beta E_{q,m}\right)^{-1/2}\right) \quad (43)$$

The procedure described above is easily adapted to obtain an entropy estimate from a quasi-harmonic analysis. In this case, the angular frequencies ω_m are obtained from $\mathbf{F}_q$ (eq. (25)) according to eq. (26). As suggested by Schlitter [48], the diagonalization process (eq. (25)) can be substituted by a determinant calculation if the correct quantum-mechanical formula is replaced by the approximate heuristic expression

$$S'_{qm,o} = \frac{k_B}{2} \ln\det\left(\mathbf{1} + \frac{e^2}{\beta\hbar^2}\underline{\mathbf{D}}_r\right) \quad (44)$$

where $\mathbf{1}$ is the unit matrix and e Euler's number. In practice $S'_{qm,o}$ is always larger than $S_{qm,o}$, and usually very close. However, the computational gain achieved by the latter substitution is typically moderate [50] and the diagonalization is still required if one wishes to obtain the quasi-harmonic modes (necessary to correct for anharmonicity and mode correlation, see following). Therefore, unless otherwise specified, eq. (44) will not be further considered in this article.

As mentioned in Section 2.2, previous applications of the quasi-harmonic analysis in terms of internal coordinates incorrectly omitted the metric-tensor-weighting in eq. (25). The consequences of this 'approximation', which effectively amounts to setting $\underline{\mathbf{A}}_q = \mathbf{1}$ (in unspecified units of mass), in terms of the estimated entropy are, however, difficult to quantify.

In practice, a quasi-harmonic entropy estimate is affected by three types of errors: (i) neglect of anharmonicities in the individual modes; (ii) neglect of correlations (beyond pairwise-linear ones) among the modes; (iii) neglect of metric-tensor effects. If the estimate could be corrected for these effects, the calculated entropies would converge to a common value (the entropy of the simulated system) irrespective of the original coordinate system and methodology employed. At the classical level, it is possible to investigate errors due to anharmonicities and pairwise supralinear correlations.

The effect of anharmonicities in the individual modes can be assessed in the following way. Qualitatively, one may evaluate how close the actual distributions $p'_m(a_{q,m})$ associated with the individual quasi-harmonic coordinates, as obtained from a simulation, are to the model Gaussians $p'_{o,m}$ in eq. (17). To this purpose, a linear regression is performed, the correlation coefficient of which may serve as a measure of the degree of harmonicity of the mode considered. More precisely, the function

$$f(a_{q,m}) = \left[-F_{q,m} \ln \left(\frac{p'_m(a_{q,m})}{p'_m(0)} \right) \right]^{\frac{1}{2}} \tag{45}$$

should be a straight line of unit slope going through the origin if the mode is perfectly harmonic.

Quantitatively, given the actual distribution $p'_m(a_{q,m})$ from the simulations, it is possible to calculate the exact (anharmonic) contribution of a single eigenmode to the classical entropy as (compare with eqs. (39), (40) and (41))

$$s^{ah}_{cl,m} = k_B \left[\frac{1}{2} \left(1 - \ln \frac{\beta h^2}{2\pi} \right) - \int da_{q,m}\ p'_m(a_{q,m}) \ln p'_m(a_{q,m}) \right] \tag{46}$$

The overall entropy correction for the non-harmonic behavior of the actual probability distributions along the individual modes then reads

$$\begin{aligned} \Delta S^{ah}_{cl} &= S^{ah}_{cl} - S_{cl,o} \\ &= \sum_{m=1}^{M} \left[s^{ah}_{cl,m} - s_{cl,o}\left(\left(\beta F_{q,m} \right)^{-1/2} \right) \right] \end{aligned} \tag{47}$$

The quantity ΔS^{ah}_{cl} is always negative or zero (see Appendix A).

The effect of pairwise mode correlations can be assessed quantitatively in the following way. As indicated by eq. (29) distinct quasi-harmonic modes are (by construction) linearly independent from each other. However, the absence of linear correlation between pairs of modes does not rule out the possibility of higher-order correlations. Given the actual two-dimensional distribution $p'_{m,n}(a_{q,m}, a_{q,n})$, it is possible to calculate the exact (correlated) contribution of the two modes to the classical entropy as (compare with eqs. (39) and (46))

$$\begin{aligned} s^{pc}_{cl,mn} &= k_B \left[\left(1 - \ln \frac{\beta h^2}{2\pi} \right) \right. \\ &\left. - \int da_{q,m}\, da_{q,n}\ p'_{m,n}(a_{q,m}, a_{q,n}) \ln p'_{m,n}(a_{q,m}, a_{q,n}) \right] \end{aligned} \tag{48}$$

The overall entropy correction for the pairwise (supralinear) correlation between modes in the actual probability distribution then reads

$$\begin{aligned} \Delta S^{pc}_{cl} &= \sum_{m=1}^{M} \sum_{n=m+1}^{M} \Delta s^{pc}_{cl,mn} \\ &= \sum_{m=1}^{M} \sum_{n=m+1}^{M} \left(s^{pc}_{cl,mn} - s^{ah}_{cl,m} - s^{ah}_{cl,n} \right) \end{aligned} \tag{49}$$

This correction is relative to an entropy already including the effect of mode anharmonicities, *i.e.* the corrections ΔS^{ah}_{cl} and ΔS^{pc}_{cl} (eqs. (47) and (49)) are cumulative. Note that although the quantity $\Delta S^{ah}_{cl} + \Delta S^{pc}_{cl}$ is always negative or zero (see Appendix A), it cannot be guaranteed that the same result holds for ΔS^{pc}_{cl} taken alone. In principle, the same approach could be used to quantify the effect of higher-order correlations (*i.e.* involving more than two modes). However, extending the numerical integration in eq. (48) to more than two dimensions is currently not feasible (except possibly for very small systems) due

computational and memory costs, as well as to significantly poorer statistics based on simulations over a limited timescale. Again, although the total correction (including anharmonicities and correlations up to an arbitrary order) must be negative, it cannot be guaranteed that the successive corrections are individually negative. One may assume that pairwise correlation represents the dominating term. Yet, higher-order correlations, even if intrinsically smaller, may still be significant because they are more numerous (*e.g.* $O\left[N^3\right]$ for three-mode correlations compared to $O\left[N^2\right]$ for pairwise ones).

In practice, the entire analysis relies on the representation of the probability distributions $p'_m\left(a_{q,m}\right)$ and $p'_{mn}\left(a_{q,m}, a_{q,n}\right)$ through one and two-dimensional histograms evaluated based on the simulation trajectory. The choice of the corresponding histogram bin width is of importance, as discussed in Appendix C. This approach is conceptually similar to a previously reported analysis of the effect of anharmonicity on the estimated entropy [45], but considers the anharmonicity in the linearly uncorrelated motional modes, rather than in the original set of coordinates.

An alternative route to derive anharmonicity and mode-coupling corrections to the estimated entropies is to make use of higher-order moments (*i.e.* beyond pairwise covariances) of the configurational probability distributions as evaluated from the simulated trajectory [47]. However, the present method should in principle be more accurate because it relies on full one-dimensional and two-dimensional projections of this probability distribution, rather than on a limited set of moments.

3. METHODS

The molecular dynamics (MD) trajectories of two different peptides, the β-hexapeptide H-β^2-HVal-β^3-HAla-β^2-HLeu-β^3-HVal-β^2-HAla-β^3-HLeu-OH and the β-heptapeptide H-β^3-HVal-β^3-HAla-β^3-HLeu-(*S*,*S*)-β^3-HAla(αMe)-β^3-HVal-β^3-HAla-β^3-HLeu-OH, both simulated for 100 ns in explicit-solvent methanol at 298 K or 340 K were used for the analyses. The initial configurations were chosen as extended (*i.e.* all backbone torsional dihedral angles in *trans* conformation) for the β-hexapeptide and folded (based on a NMR model structure) for the β-heptapeptide, respectively. All simulations were carried out using the GROMOS96 program [114,115], the 43A1 united-atom force field [116,117], and a GROMOS-

Table 1. MD simulations and reference codes for the corresponding ensembles of configurations.[a]

Code	System	Solvent	T [K]	Nr. Solute Atoms	Ensemble	Nr. Config.
X_340_F					F	18197
X_340_U	β-hexapeptide	CH_3OH	340	56	U	181803
X_340_A					A	200000
X_298_A	β-hexapeptide	CH_3OH	298	56	A	200000
P_340_F					F	33895
P_340_U	β-heptapepeptide	CH_3OH	340	64	U	166105
P_340_A					A	200000
P_298_A	β-heptapepeptide	CH_3OH	298	64	A	200000

[a] For each system, the number of configurations belonging to the folded (F), unfolded (U), and overall (A) ensembles is given. All ensembles were generated based on 100 ns simulations with a sampling frequency of 0.5 ps.

compatible three-site methanol model [118]. The simulation setup and trajectory analyses for the two β-peptides are described in details elsewhere [98-101,103]. All simulations were characterized by multiple folding-unfolding events on the timescale considered.

Ensembles of folded and unfolded configurations were defined using backbone atom-positional root-mean-square deviations (RMSD) from ideal folded structures matching a (*P*)-12/10-helix (β-hexapeptide, ref. 101) and a (*M*)-3_1-helix (β-heptapeptide, ref. 98) as similarity criterion, with cutoffs of 0.08 nm and 0.10 nm, respectively. All backbone atoms (N, C^{α}, C^{β}, and C^{CO}) were used in the RMSD calculations. A summary of the ensembles used for the present analyses is given in table 1. The distinction between folded and unfolded ensembles was only made for the simulations at 340 K, due to the limited number of unfolded structures present in the ensembles generated at 298 K.

The quasi-harmonic and essential-mode analyses were performed by calculating the solute all-atom covariance matrix (eq. (24)) in Cartesian coordinates after least-squares fit superposition of all configurations onto a reference structure so as to eliminate overall translation and rotation [49], and diagonalization of the mass-weighted form of this matrix (eqs. (25) and (31)) or of the matrix itself (eq. (32)). The reference structure for least-squares fitting was taken to be the corresponding experimental model helical (folded) structure, unless otherwise specified. The six eigenvalues of lowest magnitude (corresponding to the suppressed overall translation and rotation) are typically very close to zero, and were left out from all subsequent analyses. After determination of the quasi-harmonic modes (columns of the matrix $\underline{\mathbf{V}}$ in eq. (25); to simplify the notation, the $\boldsymbol{r}$ subscript indicating the use of Cartesian coordinates will be omitted here and in section 4) and essential-modes (columns of the matrix $\underline{\mathbf{W}}$ in eq. (32)), the trajectory was projected in this basis to obtain the time series of the corresponding quasi-harmonic coordinates $\boldsymbol{b}$ (eq. (27)) and essential coordinates $\boldsymbol{c}$ (eq. (33)). The eigenvectors (and the corresponding projected coordinate components) were sorted in order of decreasing eigenvalues (*i.e.* decreasing variance or increasing frequency). The relationship between the coordinates $\boldsymbol{b}$ and $\boldsymbol{c}$, as well as between the eigenvectors contained in the columns of $\underline{\mathbf{V}}$ and $\underline{\mathbf{W}}$, was first analyzed in detail. The quasi-harmonic configurational entropy for increasingly long segments of the simulations (differing in length by 1 ns) was then calculated as described in Section 2.4 (eq. (43); the factor ξ was set to 1). Corrections for mode anharmonicity (eq. (47)) and pairwise supralinear mode correlation (eq. (49)) were also calculated based on ensembles of $2 \cdot 10^5$ structures. Corrections for pairwise supralinear mode correlation calculated using only $2 \cdot 10^4$ structures of each of the ensembles considered differ by less than 1% ($6\ \mathrm{J \cdot K^{-1} \cdot mol^{-1}}$) suggesting that the reported values are well converged.

4. RESULTS AND DISCUSSION

4.1 Quasi-harmonic and essential modes

The time series of selected components of the quasi-harmonic coordinates $\boldsymbol{b}$ and essential coordinates $\boldsymbol{c}$ are shown in figure 2, together with the corresponding time series of the backbone atom-positional root-mean-square deviation (RMSD) with respect to the helical (folded) structure, for the X_340_A simulation. The fluctuations of the coordinates associated with low-frequency modes (eigenvectors with low index) are clearly correlated with fluctuations in the RMSD of the system (figure 2(b)-2(d) vs. 2(a)). The fluctuations of the first components of $\boldsymbol{b}$ and $\boldsymbol{c}$ are also highly correlated among each other. For the components of $\boldsymbol{b}$, the amplitude of the fluctuations becomes smaller for higher eigenvector indices, corresponding to higher frequency modes (figure 2(c)-2(f)). A similar feature holds for the components of $\boldsymbol{c}$ (data not shown).

The probability distributions along selected components of the transformed coordinates $\boldsymbol{b}$ and $\boldsymbol{c}$ are shown in figures 3 and 4, respectively, for the same ensemble X_340_A. The corresponding model Gaussian functions with the same variances and vanishing averages (eq. (17) for the quasi-harmonic coordinates) are also represented. For both types of coordinates, the actual distributions

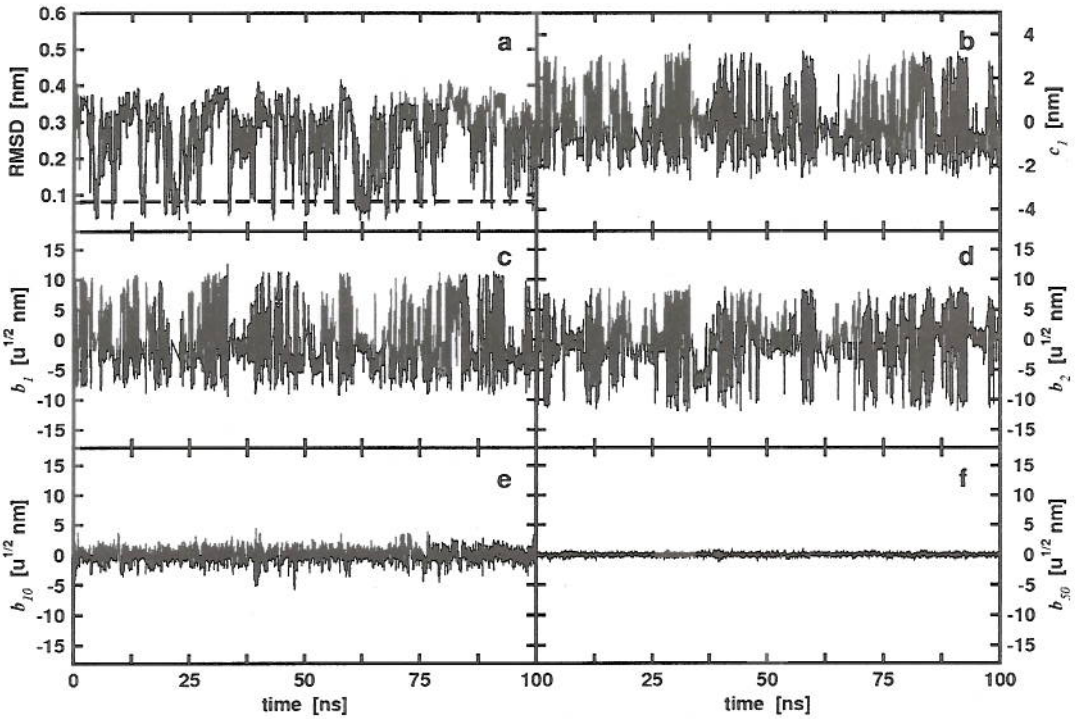

Figure 2. Time series of the backbone atom-positional root-mean-square deviation (RMSD) with respect to the helical (folded) structure and of selected components of the quasi-harmonic coordinates $\boldsymbol{b}$ and essential coordinates $\boldsymbol{c}$ for the β-hexapeptide simulation X_340_A (table 1). (a) RMSD as function of time (the dashed line indicates the RMSD similarity criterion chosen to define the folded ensemble). (b) First component of $\boldsymbol{c}$, *i.e.* projection of the simulated trajectory onto the first eigenvector of the solute all-atom covariance matrix. (c) First, (d) second, (e) tenth, and (f) fiftieth components of $\boldsymbol{b}$, *i.e.* projection of the simulated trajectory onto the corresponding eigenvectors of the solute all-atom mass-weighted covariance matrix. The least-square fitting was performed with the experimental model helical (folded) configuration as a reference structure. The letter 'u' stands for atomic mass unit.

become narrower and increasingly similar to the model Gaussians for higher mode indices, *i.e.* the quasi-harmonic modes become increasingly stiff and harmonic. Similarly, the essential mode distributions become less structured and provide contributions of decreasing magnitudes to the total system fluctuation. However, the distributions along the lowest-frequency modes (figures 3 and 4, upper six panels) are far from Gaussian, and evidently result from the superposition of two distributions. This feature was previously observed for Cartesian displacements of a single atom in a α-helical peptide (figure 4 in ref. 92). The contributions to the probability distributions provided by the subset of folded configurations during the simulations are also displayed in the figures. Clearly, the folded configurations provide the characteristic narrow peak (overlayed by a broader distribution for the unfolded state) in the overall distributions for low-frequency modes. Similar features are found for the β-heptapeptide simulation at 340 K (data not shown). The comparison of figures 3 and 4 also reveals a striking correspondence between the distributions along individual components of the quasi-harmonic and essential coordinates with identical indices.

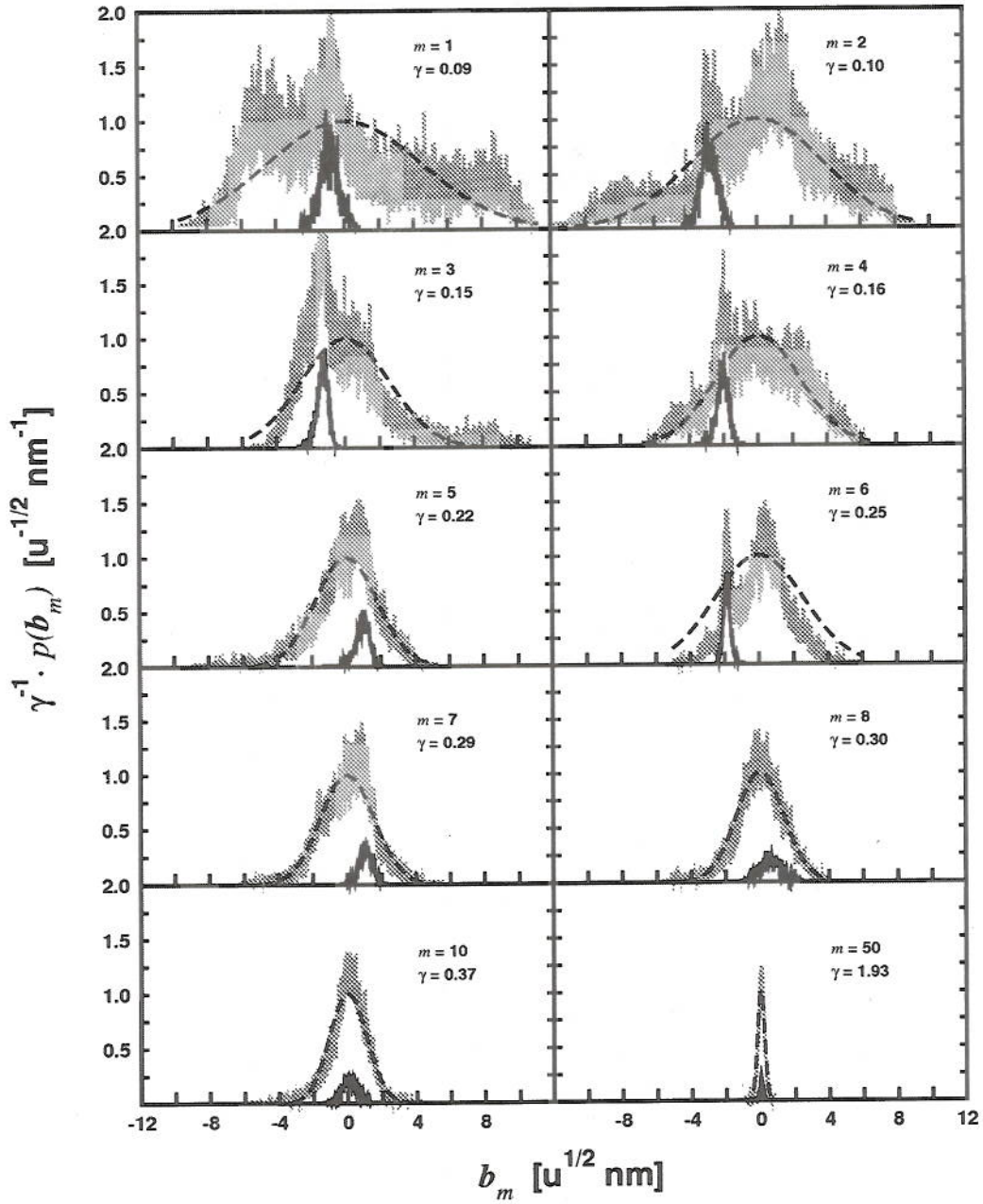

Figure 3. Probability distributions along selected components of the quasi-harmonic coordinates $\boldsymbol{b}$ for the β-hexapeptide simulation X_340_A (table 1). The actual distributions from the whole simulation (gray line) are displayed together with the corresponding approximate Gaussians (dashed line, eq. (17)) for increasing component indices m. The contribution to the distributions provided by the folded configurations only are also displayed (solid line). All probability distributions are normalized and shown after scaling by a factor γ for graphical purposes. The letter 'u' stands for atomic mass unit.

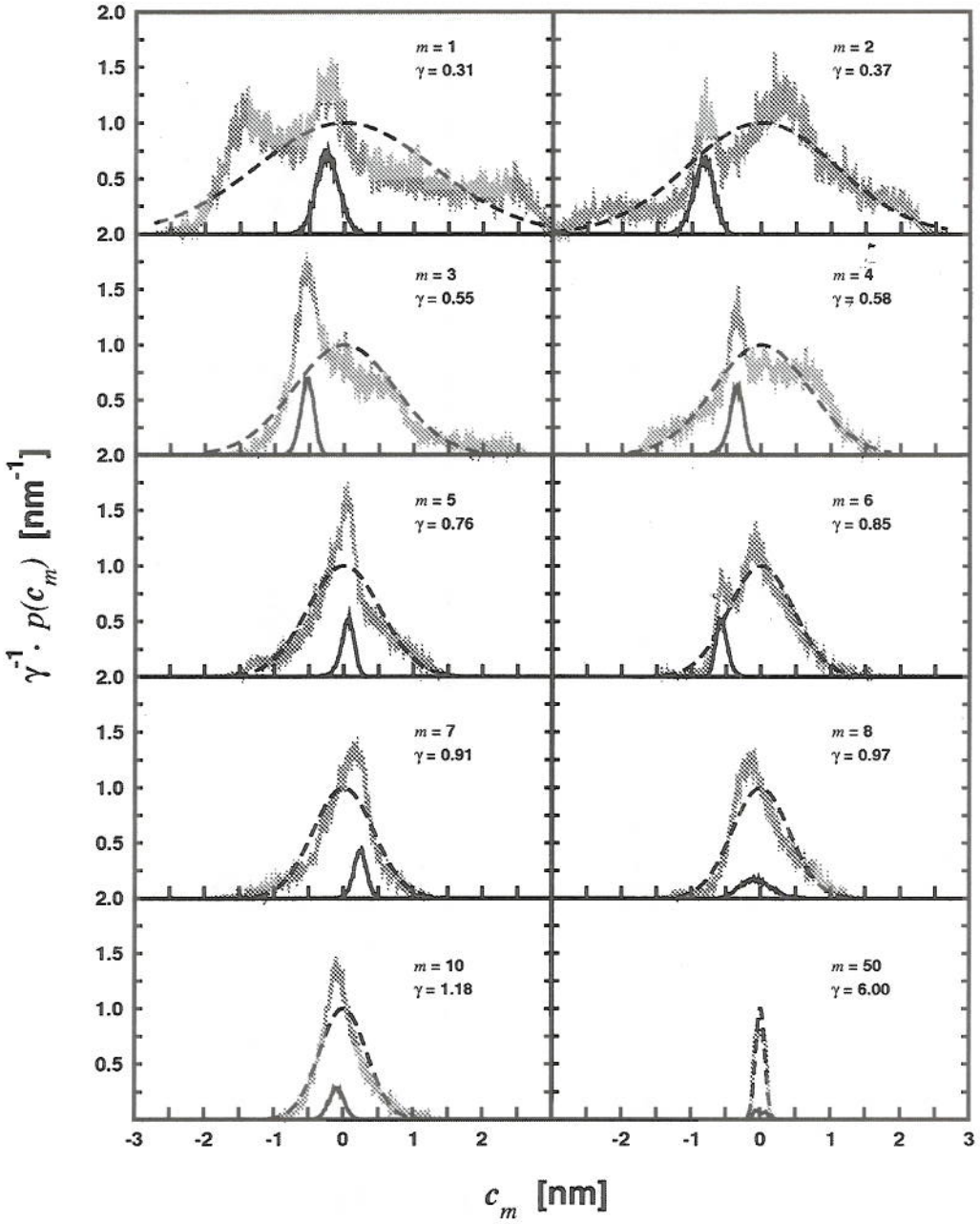

Figure 4. Probability distributions along selected components of the essential coordinates c for the β-hexapeptide simulation X_340_A (table 1). The actual distributions from the whole simulation (gray line) are displayed together with the corresponding approximate Gaussians with the same variance and a vanishing average (dashed line) for increasing component indices m. The contribution to the distributions provided by the folded configurations only are also displayed (solid line). All probability distributions are normalized and shown after scaling by a factor ν for graphical purposes.

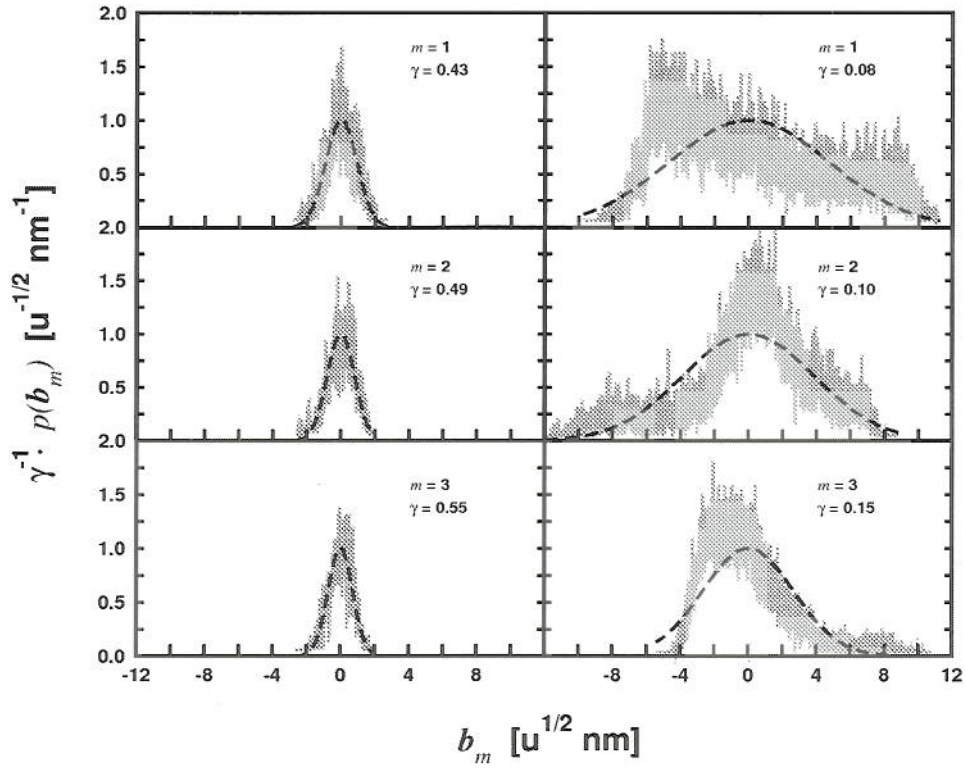

Figure 5. Probability distributions along the first three components of the quasi-harmonic coordinates $\boldsymbol{b}$ for the folded ensemble X_340_F (left panels) and the unfolded ensemble X_340_U (right panels) of the β-hexapeptide simulation at 340 K (table 1). The actual distributions (gray line) are displayed together with the corresponding approximate Gaussians (dashed line, eq. (17)) for increasing component indices m. All probability distributions are normalized and shown after scaling by a factor γ for graphical purposes. The letter 'u' stands for atomic mass unit.

The results of separate quasi-harmonic analyses carried out for the folded and unfolded ensembles of configurations of the β-hexapeptide simulation at 340 K are displayed in figure 5. The probability distributions along the three lowest-frequency components of the quasi-harmonic coordinates $\boldsymbol{b}$ differ significantly between the X_340_F and X_340_U ensembles. The compact helical fold is characterized by motions of relatively small amplitude, whose distributions are nearly Gaussian even for the first components. In contrast, for the unfolded ensemble, the first components are characterized by broad distributions, which deviate significantly from the ideal Gaussian shape. Thus, the folded ensemble appear to be intrinsically more harmonic in its low-frequency modes compared to the unfolded one. The corresponding distributions for the overall ensemble X_340_A (figure 3) are very close to a superposition of the probability distributions associated with the folded and unfolded ensembles (weighted by the relative populations of the two states and with the folded distribution shifted to a non-zero average value). Analogous observations are made for the folded, unfolded and overall ensembles in the β-heptapeptide simulation at 340 K (data not shown).

The correlation between the time-evolutions of the individual components b_m of the quasi-harmonic

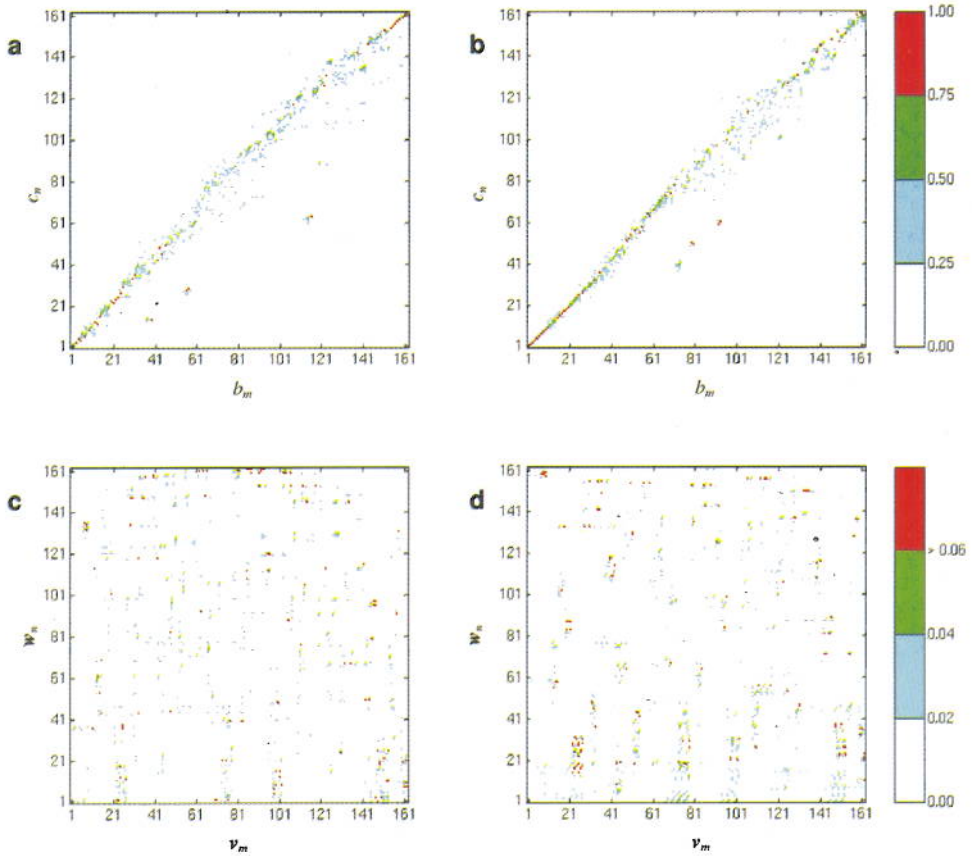

Figure 6. Correlation between the time evolutions of individual components b_m of the quasi-harmonic coordinates $\boldsymbol{b}$ and c_n of the essential coordinates $\boldsymbol{c}$ for the folded ensemble X_340_F (a) and the whole ensemble X_340_A (b) of the β-hexapeptide simulation at 340 K (table 1), the correlation being calculated as $\left|\langle b_m \rangle^{-1/2} \langle c_n \rangle^{-1/2} \langle b_m c_n \rangle\right|$, and corresponding scalar products (absolute values) between the eigenvectors $\boldsymbol{v_m}$ of the mass-weighted covariance matrix and the eigenvectors $\boldsymbol{w_n}$ of the covariance matrix for the X_340_F (c) and X_340_A (d) ensembles. The last six eigenvectors with (nearly) vanishing eigenvalues are omitted from the analysis.

coordinates $\boldsymbol{b}$ and c_n of the essential coordinates $\boldsymbol{c}$ are displayed in the form of correlation matrices in figures 6(a) and 6(b), for either the folded ensemble X_340_F or the whole ensemble X_340_A corresponding to the β-hexapeptide simulation at 340 K. A high correlation is found for pairs of components with close indices (*i.e.* along the diagonal), as previously pointed out for the first components of the two vectors (figure 2(b) vs. figure 2(c)). However, the correlations are typically negligible for all other combinations of indices (*i.e.* away from the diagonal), with only a few exceptions. The scalar products (absolute values) between eigenvectors $\boldsymbol{v_m}$ of the mass-weighted covariance matrix and corresponding eigenvectors $\boldsymbol{w_n}$ of the covariance matrix are displayed in the form of matrices in figures 6(c) and 6(d) for the same ensembles. These scalar products are fairly low (maximum observed magnitude of 0.36 over the two ensembles) and show no clear trends in terms of combination of indices. These observations are valid for both the folded and

overall ensembles of both peptides (data for the heptapeptide at 340 K not shown). Thus, although the motions along the quasi-harmonic and essential coordinate components with similar indices are highly correlated, the corresponding eigenvectors do not seem to be connected by any simple mathematical relationship. This absence of simple connection is illustrated for the first eigenvectors v_1 and w_1 of the X_340_A ensemble in figure 7(a), which displays the components $v_{1,l}$ and $w_{1,l}$ of the two vectors along the $l = 1...M$ ($M = 3N$, N being the number of solute atoms) Cartesian coordinates. The correlation between the corresponding time series of b_1 and c_1 (figure 6(b)) is 0.997, while the scalar product between the two eigenvectors is as low as 0.002 (figure 6(d)). Indeed, the distributions of the eigenvector components along atomic Cartesian coordinates depend largely on whether mass- weighting is applied or not prior to diagonalization, and the two types of modes involve significantly different parts of the peptide. Furthermore, these differences affect both hydrogen and heavy atoms, and are not correlated with the atomic masses in any obvious way. The convergence of these eigenvector components as a function of the sampling time is slow, as shown in figures 7(a)-7(d). The component distribution is significantly different when comparing the results from the first 40, 60 or 80 ns of the simulation to the results obtained from the entire 100 ns simulation. The corresponding scalar products (absolute values) between the first eigenvector of the covariance (or mass-weighted covariance) matrix calculated using different sampling times are reported in table 2. The values are generally higher for the first eigenvector of the mass-weighted covariance matrix (0.07-0.47) compared to the first eigenvector of the covariance matrix (0.05-0.27), suggesting a better convergence of the eigenvector components in the former case. The data indicate that proper convergence of the eigenvector components may require sampling times beyond 100 ns.

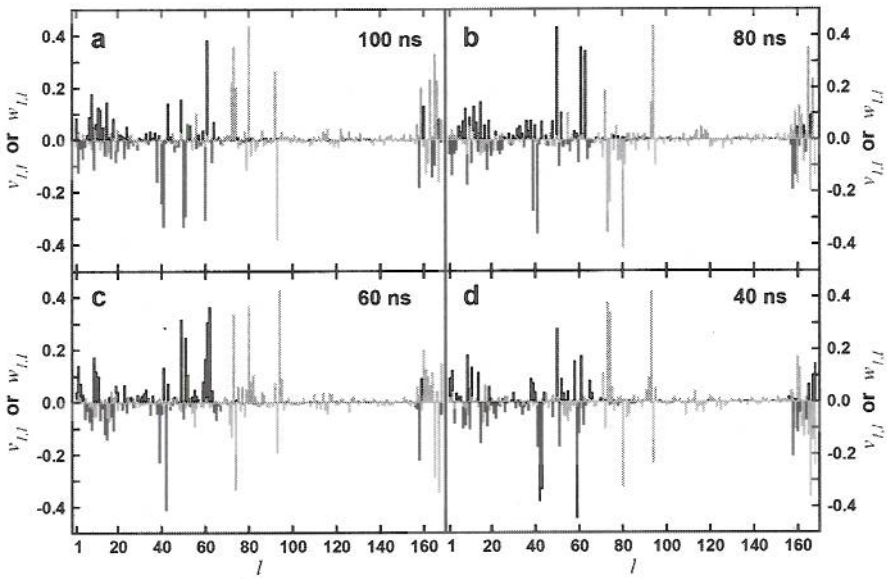

Figure 7. Components $l = 1..M$ ($M = 3N$, where N is the number of solute atoms) of the first eigenvector v_1 of the mass-weighted covariance matrix (grey) and the first eigenvector w_1 of the covariance matrix (black) for the β-hexapeptide simulation at 340 K (table 1). The different graphs correspond to (a) the whole ensemble X_340_A (100 ns). (b) First 80 ns. (c) First 60 ns. (d) First 40 ns.

Table 2. Scalar products (absolute values) of the first eigenvectors v_1 of the mass-weighted covariance matrix evaluated based on different simulation times (upper right triangle) and scalar products of the first eigenvectors w_1 of the covariance matrix evaluated based on different simulation times (lower left triangle) for the β-hexapeptide simulation at 340 K.[a]

Time [ns]	20	40	60	80	100
20	1	0.07	0.22	0.26	0.37
40	0.07	1	0.11	0.22	0.10
60	0.06	0.07	1	0.25	0.40
80	0.05	0.16	0.07	1	0.47
100	0.24	0.05	0.08	0.27	1

[a] The corresponding components of v_1 and w_1 (for 40 ns, 60 ns, 80 ns and 100 ns sampling periods) are reported in figure 6.

The eigenvalues F_m and G_m of the mass-weighted covariance matrix (eq. (25)) and of the covariance matrix (eq. (32)), respectively, are shown in figure 8 as a function of the eigenvector index m for the β-hexapeptide simulation at 340 K. The corresponding cumulative estimates of the quasi-harmonic entropy $S_{qm,o}$ (eq. (43)) and of the total system mean-square fluctuation (MSF; eq. (36)) are also displayed. In both cases, the eigenvalues decrease similarly rapidly in relative magnitude with increasing eigenvector index. However, although 90 % (99 %) of the overall MSF is already captured by the first 14 (53) lowest-index essential modes, the entropy is significantly influenced by higher-frequency modes. In the latter case, accounting for 90 % (99 %) of the total entropy requires the inclusion of the first 96 (138) lowest-frequency quasi-harmonic modes. The reason for this difference resides mainly in the different functional dependences for the contribution of an essential mode m to the MSF (*i.e.* the eigenvalue G_m itself) and that of a quasi-harmonic mode to the entropy (*i.e.* $s_{qm,o}$ in eq. (42)

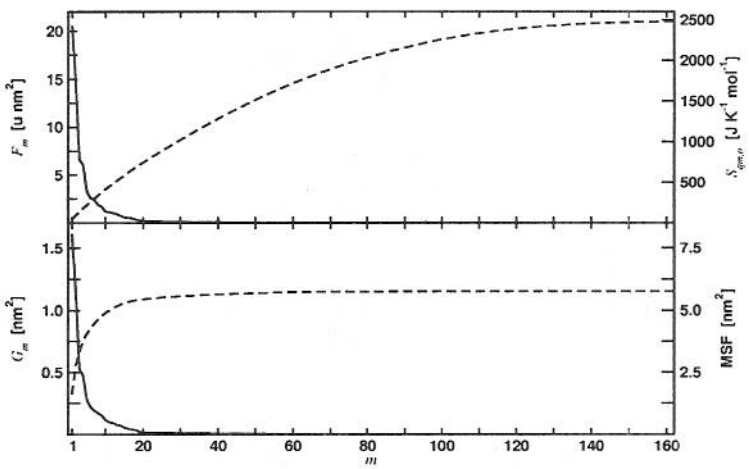

Figure 8. Eigenvalues F_m of the mass-weighted covariance matrix (top panel, solid line) and G_m of the covariance matrix (bottom panel, solid line) as a function of the eigenvector index m for the β-hexapeptide simulation at 340 K (table 1). The corresponding cumulative estimates for the entropy $S_{qm,o}$ (eq. (43)) and total system mean-square fluctuations (MSF; eq. (36)) are also shown (dashed lines). The last six eigenvectors with (nearly) vanishing eigenvalues are omitted from the analysis. The letter 'u' stands for atomic mass unit.

with $\omega_m = \left(\beta F_m\right)^{-1/2}$). Therefore, the estimation of quasi-harmonic entropies from a reduced number of low-frequency modes, considered to be 'conformational' as opposed to 'vibrational' modes [16,21,45,62], is questionable.

4.2 Convergence of the entropy estimate

The convergence properties of the quasi-harmonic entropy as a function of the sampling time are illustrated in figure 9 for the β-heptapeptide simulations at 298 and 340 K, along with the corresponding time series of the backbone atom-positional RMSD with respect to the (helical) folded structure. Quasi-harmonic entropy estimates based on Cartesian coordinates depend on the reference structure chosen to perform the least-squares fitting procedure aimed at removing the overall solute translation and rotation. Entropies calculated with fitting to either the folded structure or the extended structure are displayed. In addition, the results obtained by application of the exact quantum-mechanical formula ($S_{qm,o}$; eq. (43)) are compared to those obtained by application of the alternative (approximate) entropy formula proposed by Schlitter ($S'_{qm,o}$; eq. (44)). At 298 K, all entropy curves show a stepwise build-up with time [74]. Relatively long plateaus are interrupted by steep increases when the system visits a new set of unfolded configurations. This suggests that the unfolded state is not appropriately sampled at this temperature on the 100 ns time scale, so that the final entropy estimate for the peptide is probably not well converged. At 340 K, the sampling of the unfolded state is much more effective and the entropy estimate appears to converge after about 50 ns. The fitting based on a folded reference structure systematically leads to lower entropies compared to the fitting based on

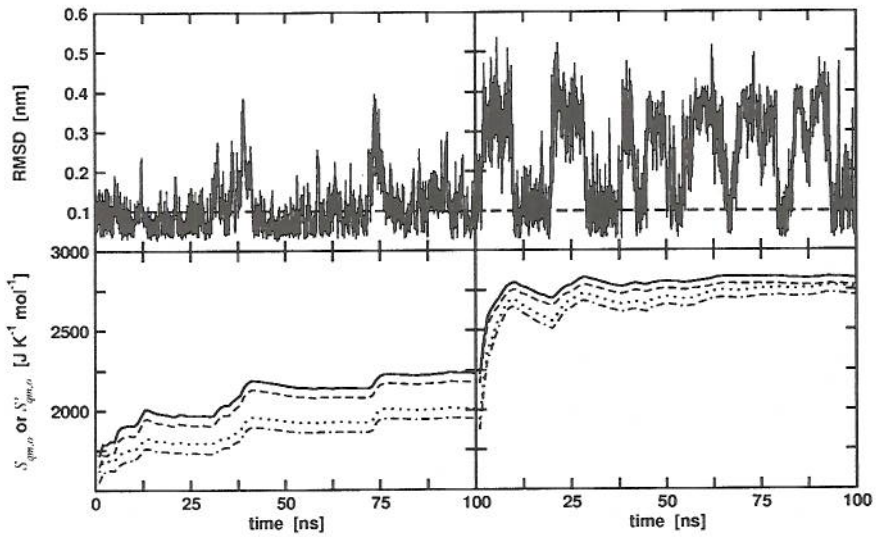

Figure 9. Time series of the backbone root-mean-square backbone atom-positional deviation (RMSD) with respect to the helical (folded) structure (top panels) and build-up curves of the quasi-harmonic entropy (bottom panels) for the β-heptapeptide simulations at 298 K (left panels) and 340 K (right panels). The horizontal dashed line (top panels) indicates the RMSD similarity criterion used to define the folded ensemble. Different entropy estimates are reported (lower panels) with fitting to extended (solid and dashed lines) and folded (dotted and dot-dashed lines) reference structures, and based on the exact quantum-mechanical formula for the entropy ($S_{qm,o}$, eq. (43); dashed and dot-dashed lines) and on the approximate (upper bound) formula proposed by Schlitter ($S'_{qm,o}$, eq. (44); solid and dotted lines).

an extended reference structure (the difference being larger in the simulation at 298 K). This probably results from a more effective removal of the overall peptide rotation in the former case. The corresponding relative differences in the final entropy estimates are 10 and 2 % (based on either entropy formula) at 298 and 340 K, respectively. As expected, the values provided by the approximate expression of Schlitter are systematically larger than the exact quasi-harmonic ones. The relative differences in the final entropy estimates are 2 and 3 % at 298 and 340 K, respectively.

4.3 Correction for mode anharmonicities

The influence of anharmonicity in the individual quasi-harmonic modes is illustrated in figure 10 for the β-hexapeptide simulation at 340 K. The correlation coefficient R involved in the fit of eq. (45) to a straight line is shown as a function of the eigenvector index. Deviations from unity indicate that the probability distribution associated with the corresponding quasi-harmonic coordinate component deviates from an ideal Gaussian. The corresponding contribution of the mode to the estimated entropy is also displayed, as calculated using three different expressions: the quantum-mechanical harmonic expression $s_{qm,o}$ $s_{qm,o}$ (eq. (42)), the classical harmonic expression $s_{cl,o}$ (eq. (41)) and the classical anharmonic expression $s^{ah}_{cl,m}$ (eq. (46)). Agreement

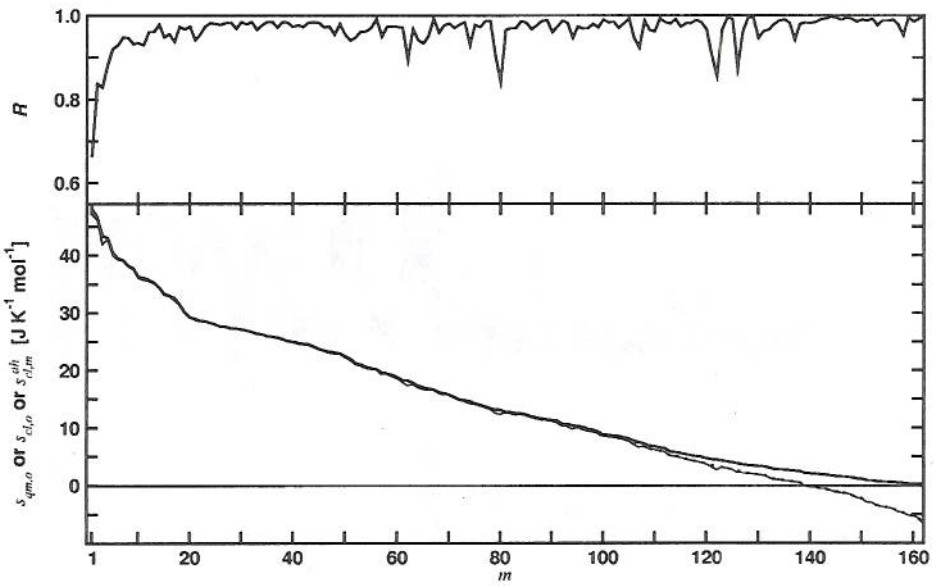

Figure 10. Correlation coefficient R for the fitting of the probability distribution associated with a quasi-harmonic coordinate to a Gaussian (top panel) and per-mode entropy contribution (bottom panel) for the β-heptapeptide simulation at 340 K (table 1) displayed as a function of the eigenvector index m. The coefficient R is the linear-regression coefficient corresponding to the fit of eq. (45) to a straight line. The per-mode contributions to the quasi-harmonic entropy are evaluated using either the quantum-mechanical harmonic expression $s_{qm,o}$ (eq. (42); thick solid line), the classical harmonic expression $s_{cl,o}$ (eq. (41); dotted line) or the classical anharmonic expression $s^{ah}_{cl,m}$ (eq. (46); thin solid line). The last six eigenvectors with (nearly) vanishing eigenvalues are omitted from the analysis.

between $s_{cl,o}$ and $s_{qm,o}$ is expected (and found) for low-frequency modes, where quantum effects become negligible [45,48]. In contrast, for the high-frequency modes, the quantum-mechanical entropy contributions correctly converge to zero, while the corresponding classical estimates become negative (they are expected to ultimately diverge to minus infinity in the limit of very high frequencies). Agreement between $s_{cl,o}$ and $s^{ah}_{cl,m}$ is expected (and found) for high-frequency modes, where anharmonicity becomes unimportant (see coefficient R). However, even for the distributions associated with low-frequency modes, which differ significantly from ideal Gaussians (figure 3), the effect of mode anharmonicity on the calculated entropy contributions remains surprisingly small.

The overall corrections ΔS^{ah}_{cl} for anharmonicity effects (eq. (47)), to be applied to the quasi-harmonic entropy estimates $S_{qm,o}$ (eq.(43)), are reported in table 3 for all ensembles considered. In all cases, these corrections are negative and rather small (at most 1.4 % of the quasi-harmonic estimate). The small magnitude of anharmonicity

Table 3. Quantum-mechanical quasi-harmonic configurational entropy $S_{qm,o}$, its (classically derived) corrections for mode anharmonicities (ΔS^{ah}_{cl}) and (supralinear) pairwise mode correlations (ΔS^{pc}_{cl}), together with the corrected value $S^{ctd} = S_{qm,o} + \Delta S^{ah}_{cl} + \Delta S^{pc}_{cl}$ for the different ensembles considered (table 1).[a]

Code	$S_{qm,o}$	ΔS^{ah}_{cl}	ΔS^{pc}_{cl}	S^{ctd}
	$\left[\mathrm{J \cdot K^{-1} \cdot mol^{-1}}\right]$	$\left[\mathrm{J \cdot K^{-1} \cdot mol^{-1}}\right]$	$\left[\mathrm{J \cdot K^{-1} \cdot mol^{-1}}\right]$	$\left[\mathrm{J \cdot K^{-1} \cdot mol^{-1}}\right]$
X_340_F	1530	-9 (0.6)	-830 (54)	691
X_340_U	2499	-8 (0.3)	-376 (15)	2115
X_340_A	2477	-9 (0.4)	-380 (15)	2088
X_298_A	2273	-10 (0.4)	-377 (16)	1886
P_340_F	1793	-8 (0.4)	-959 (53)	826
P_340_U	2766	-10 (0.4)	-533 (19)	2223
P_340_A	2716	-15 (0.5)	-546 (20)	2155
P_298_A	1945	-27 (1.4)	-688 (35)	1230

[a] The anharmonicity correction ΔS^{ah}_{cl} was calculated as in eq. (47) summing the per-mode values up to eigenvector 50 (*i.e.* in the domain of validity of the classical approximation, and where the anharmonicity effects are significant). The pairwise (supralinear) correlation correction ΔS^{pc}_{cl} was calculated as in eq. (49), excluding the last six eigenvectors with (nearly) vanishing eigenvalues. Relative values of the entropy corrections (with respect to $S_{qm,o}$ and reported in percent) are given between parentheses.

effects was previously suggested by comparison of molecular dynamics simulations and harmonic dynamics for the motions of an α-helical polypeptide [92] and of a small compact protein [119]. This observation need not necessarily apply to more complex biomolecules (*e.g.* larger proteins).

4.4 Correction for pairwise (supralinear) mode correlations

The pairwise (supralinear) correlation entropy contributions $\Delta s^{pc}_{cl,mn}$ (eq. (49)) associated with all unique combinations of quasi-harmonic modes m and n are displayed in figure 11 for the folded and complete ensembles of the β-hexapeptide simulation at 340 K. Pairwise correlation corrections for the whole ensemble tend to be small (between -0.07 and -0.02 $J \cdot K^{-1} \cdot mol^{-1}$) with larger contributions (< -0.07) predominating for combinations of low-frequency modes. In contrast, for the folded ensemble, the corrections are in general more significant (between -0.09 and -0.05 $J \cdot K^{-1} \cdot mol^{-1}$) with larger contributions (< -0.09) for combinations of high-frequency modes. Similar results are obtained for the β-heptapeptide simulation (data not shown). These observations are intuitively reasonable. For the whole ensemble, the leading type of correlations are expected to occur between large-scale backbone motions accompanying the folding-unfolding process *i.e.* between low-frequency modes. In contrast, for the folded ensemble, correlations probably dominate in the vibrations associated with the hydrogen-bonding network, *i.e.* between the high-frequency modes.

The overall corrections ΔS^{pc}_{cl} for pairwise (supralinear) correlation effects (eq. (49)), to be applied to the anharmonicity-corrected entropy estimates $S_{qm,a} + \Delta S^{ah}_{cl}$ (eqs. (43) and (47)), are reported in table 3 for all ensembles considered. In all cases, these corrections are negative and represent a significant fraction of the (uncorrected) quasi-harmonic entropies, namely about 15-35 % for both β-peptides (overall and unfolded ensembles), the effect being comparatively larger (53-54 %) for the ensembles of folded structures. Because pairwise correlation effects are more important in the folded state than

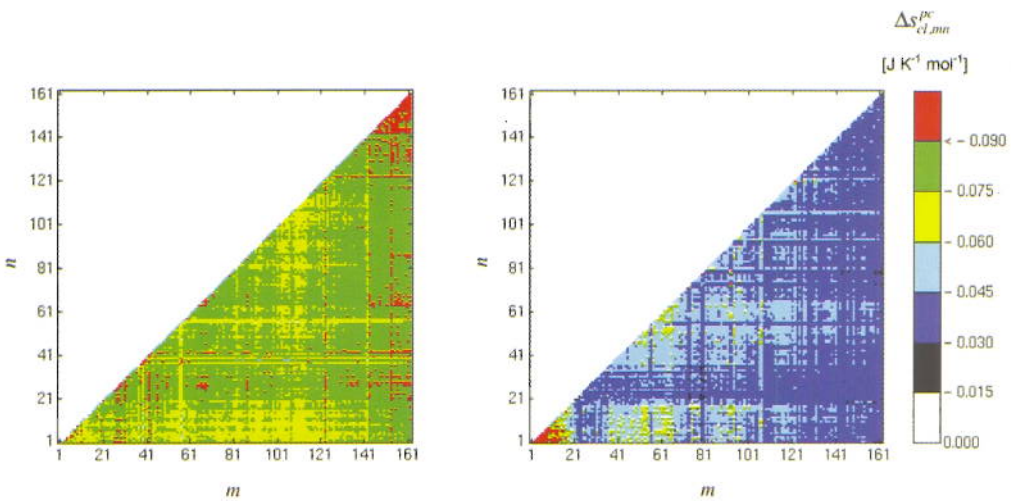

Figure 11. Pairwise (supralinear) correlation correction $\Delta s^{pc}_{cl,mn}$ (eq. (49)) associated with all unique pairs of quasi-harmonic modes m and n for the folded ensemble X_340_F (left panel) and the whole ensemble X_340_A (right panel) of the β-hexapeptide simulation (table 1). The last six eigenvectors with (nearly) vanishing eigenvalues are omitted from the analysis.

in the unfolded state of reversibly folding peptides, neglecting these correlations led to underestimating the magnitude of the solute configurational entropy decrease upon folding [74].

5. CONCLUSION

In the present article, the method to estimate configurational entropies from molecular dynamics simulations based on the quasi-harmonic approximation was investigated critically. The following aspects were considered: (i) the relationship between quasi-harmonic and essential modes; (ii) the requirement of mass-weighting (or metric-tensor-weighting) in quasi-harmonic analysis; (iii) the effect of anharmonicities in the individual modes on the estimated entropy; (iv) the effect of pairwise (supralinear) correlations among the different modes on the estimated entropy. The analysis was carried out using long (hundred nanoseconds) molecular dynamics simulations involving the reversible folding of β-peptides in methanol and considering individually the specific properties of the ensembles corresponding to folded and unfolded configurations. The following observations could be made.

The probability distributions associated with components of the quasi-harmonic coordinates $\boldsymbol{b}$ and of the essential coordinates $\boldsymbol{c}$ only deviate significantly from Gaussians for the first few components. For these components, the probability distributions result from a superposition of clearly distinguishable contributions from the folded and unfolded ensembles. The time series of the components of $\boldsymbol{b}$ and $\boldsymbol{c}$ with close indices are highly correlated. However, the corresponding eigenvectors $\boldsymbol{v}_m$ (from the mass-weighted covariance matrix) and $\boldsymbol{w}_n$ (from the covariance matrix) are not related by a simple mathematical relationship. In addition, the components of these eigenvectors converge slowly with the simulation time (*i.e.* their convergence probably requires more than 100 ns for the systems considered).

It was shown explicitly that mass-weighting of the covariance matrix prior to diagonalization is necessary if entropies are to be estimated from a quasi-harmonic analysis. The cumulative quasi-harmonic entropy estimate converges significantly more slowly upon summation of contributions over successive quasi-harmonic modes compared to the total mean-square fluctuation (MSF) upon summation over successive essential modes. Therefore, the calculation of a configurational entropy based on a reduced number of quasi-harmonic modes should be avoided.

For simulations of reversibly folding β-peptides in methanol, the convergence of the quasi-harmonic entropy estimate with simulation time requires more than 100 ns at 298 K, while it is reached within about 50 ns at 340 K. The influence of the reference structure used to perform the translational and rotational superposition of successive trajectory structures may be sizeable (up to 10 % of the calculated entropy at 298 K and 2 % at 340 K). The heuristic formula of Schlitter for the configurational entropy slightly overestimates the correct quantum-mechanical quasi-harmonic entropy (by 2-3 % for the systems considered here). The contribution of anharmonicity to the entropy estimate is generally small (less than 1.4 % of the quasi-harmonic entropy estimate) for β-peptides in methanol. However, this observation need not be valid in the case of more complex biomolecules (*e.g.* larger proteins). Finally, the (supralinear) pairwise correlation correction to the entropy is large and affects more significantly the entropy of the folded state (53-54 %) than that of the unfolded state (15-19 %). Because pairwise correlation effects are more important in the folded state than in the unfolded one, neglecting these correlations would lead to underestimate the magnitude of the solute configurational entropy decrease upon folding [74]. This underestimation is likely to be even more severe when considering larger proteins composed of tertiary and quaternary structure elements, where motional correlations in the folded state will play a comparatively larger role [9,120-121].

The method presented here permits the estimation of absolute configurational entropies of bound (non-diffusive) systems that take into account the effect of mode anharmonicity and pairwise mode correlations. The inclusion of such corrections is expected to be necessary for estimating entropy

changes for conformational transitions between states characterized by very different extents of correlations in the atomic motions. This is the case, for example, when estimating changes in configurational entropies of macromolecules upon folding, binding or complex formation. A first application to disaccharides in water has been reported [122].

ACKNOWLEDGEMENTS

R.B. would like to thank Prof. Jürgen Schlitter for insightful discussions. The authors thank Prof. Xavier Daura for making the trajectories of the β-heptapeptide available. Financial support from the National Center of Competence in Research (NCCR), Structural Biology, of the Swisse National Science Foundation (SNSF) is gratefully acknowledged.

APPENDIX A

Here, we discuss the postulated maximum-entropy property of the multi-dimensional harmonic oscillator in an arbitrary generalized coordinate system, both at the classical and quantum-mechanical levels.

At the classical level and for a bound (non-diffusive) system involving M (unconstrained) degrees of freedom, the entropy in an arbitrary generalized coordinate system $\boldsymbol{q}$ may be written (eq. (38))

$$S_{cl} = -k_B \int d\boldsymbol{q}\; p(\boldsymbol{q}) \left[\ln p(\boldsymbol{q}) - \frac{1}{2} \ln \left| \underline{\mathbf{A}}_{\boldsymbol{q}}(\boldsymbol{q}) \right| \right] + C \tag{A1}$$

where C is a configuration-independent constant, $p(\boldsymbol{q})$ the configurational probability distribution, and $\underline{\mathbf{A}}_{\boldsymbol{q}}(\boldsymbol{q})$ the metric tensor introduced in eq. (3). We wish to determine the probability distribution $p(\boldsymbol{q})$ that extremizes the entropy with the constraints

$$n = \int d\boldsymbol{q}\; p(\boldsymbol{q}) = 1 \tag{A2}$$

$$\overline{\boldsymbol{q}} = \int d\boldsymbol{q}\; \boldsymbol{q}\; p(\boldsymbol{q}) \tag{A3}$$

and

$$\underline{\mathbf{C}}_{\boldsymbol{q}} = \int d\boldsymbol{q}\; (\boldsymbol{q} \otimes \boldsymbol{q})\; p(\boldsymbol{q}) - \overline{\boldsymbol{q}} \otimes \overline{\boldsymbol{q}} \tag{A4}$$

enforcing the normalization of $p(\boldsymbol{q})$ together with a specified average $\overline{\boldsymbol{q}}$ and covariance matrix $\underline{\mathbf{C}}_{\boldsymbol{q}}$. This can be done by extremizing the functional

$$\begin{aligned} \mathcal{L}[p] = & -k_B^{-1} (S_{cl} - C) \\ & + \gamma n + \boldsymbol{\lambda}^T \overline{\boldsymbol{q}} \\ & + \mathrm{S}\left\{ \left(\underline{\boldsymbol{\mu}} \underline{\boldsymbol{\nu}} \right)^T \underline{\mathbf{C}}_{\boldsymbol{q}} \left(\underline{\boldsymbol{\mu}} \underline{\boldsymbol{\nu}} \right) \right\} \end{aligned} \tag{A5}$$

where the square brackets indicate a functional dependence, the symbol S denotes the sum of all elements of a matrix, and γ (scalar), $\boldsymbol{\lambda}$ (M - dimensional vector), $\underline{\boldsymbol{\mu}}$ (diagonal $M \times M$ - matrix) and $\underline{\boldsymbol{\nu}}$ (orthogonal $M \times M$ - matrix) contain the Lagrange multipliers enforcing the constraints of eqs. (A2), (A3) and (A4). That the last term in eq. (A5) is a valid formulation for the Lagrange multipliers associated with eq. (A4) is seen by observing that: (i) both the scaling by a diagonal matrix and the transformation by an orthogonal matrix preserve the linear independence in a set of equations; (ii) the matrices $\underline{\boldsymbol{\mu}}$ and $\underline{\boldsymbol{\nu}}$ involve $(1/2) M (M+1)$ free coefficients, which is exactly the number of independent equations involved in eq. (A4) due to the symmetry of $\underline{\mathbf{C}}_{\boldsymbol{q}}$.

For any (unconstrained) infinitesimal variation δp of p, one must have

$$\begin{aligned} \delta \mathcal{L}[p, \delta p] = & \int d\boldsymbol{q}\; \delta p(\boldsymbol{q}) \left[\ln p(\boldsymbol{q}) - \frac{1}{2} \ln \left| \underline{\mathbf{A}}_{\boldsymbol{q}}(\boldsymbol{q}) \right| \right] \\ & + \int d\boldsymbol{q}\; \delta p(\boldsymbol{q}) + \gamma \int d\boldsymbol{q}\; \delta p(\boldsymbol{q}) \\ & + \boldsymbol{\lambda}^T \int d\boldsymbol{q}\; \boldsymbol{q}\; \delta p(\boldsymbol{q}) \\ & + \mathrm{S}\left\{ \left(\underline{\boldsymbol{\mu}} \underline{\boldsymbol{\nu}} \right)^T \left[\int d\boldsymbol{q} (\boldsymbol{q} \otimes \boldsymbol{q})\; \delta p(\boldsymbol{q}) \right. \right. \\ & - \overline{\boldsymbol{q}} \otimes \int d\boldsymbol{q}\; \boldsymbol{q}\; \delta p(\boldsymbol{q}) \\ & \left. \left. - \int d\boldsymbol{q}\; \boldsymbol{q}\; \delta p(\boldsymbol{q}) \otimes \overline{\boldsymbol{q}} \right] \left(\underline{\boldsymbol{\mu}} \underline{\boldsymbol{\nu}} \right) \right\} \\ & = 0 \end{aligned} \tag{A6}$$

This leads to the requirement

$$\ln p(q) = -1 - \gamma + \mathrm{S}\left\{\left(\underline{\mu}\underline{\nu}\right)^T (\overline{q} \otimes \overline{q}) \left(\underline{\mu}\underline{\nu}\right)\right\} + \frac{1}{2}\ln\left|\underline{\mathbf{A}}_q(q)\right| - \boldsymbol{\lambda}^T q - \mathrm{S}\left\{\left(\underline{\mu}\underline{\nu}\right)^T [(q-\overline{q}) \otimes (q-\overline{q})] \left(\underline{\mu}\underline{\nu}\right)\right\} \tag{A7}$$

Introducing the symmetric matrix

$$\underline{\alpha} = \left(\underline{\mu}\zeta\right) \otimes \left(\underline{\mu}\zeta\right) \tag{A8}$$

with

$$\zeta_m = \sum_{n=1}^{M} \nu_{mn} \tag{A9}$$

one easily verifies that

$$\mathrm{S}\left\{\left(\underline{\mu}\underline{\nu}\right)^T (x \otimes y) \left(\underline{\mu}\underline{\nu}\right)\right\} = x^T \underline{\alpha}\, y \tag{A10}$$

Using this result and introducing the derived multiplier

$$\tilde{\gamma} = e^{-1-\gamma+\overline{q}^T \underline{\alpha}\, \overline{q}} \tag{A11}$$

leads to

$$p(q) = \tilde{\gamma}\left|\underline{\mathbf{A}}_q(q)\right|^{1/2} e^{-\boldsymbol{\lambda}^T q - (q-\overline{q})^T \underline{\alpha}\, (q-\overline{q})} \tag{A12}$$

Note that eqs. (A8) and (A9) with $\underline{\mu}$ diagonal and $\underline{\nu}$ orthogonal impose some constraints on acceptable $\underline{\alpha}$ matrices (in particular, this matrix must be symmetric). It is easily verified that eq. (A12) corresponds to a maximum in the entropy by calculating the second derivative of this quantity. In the general case where the metric tensor is configuration dependent, the maximum entropy is reached by a function of the form of eq. (A12) with $\tilde{\gamma}$, $\boldsymbol{\lambda}$ and $\underline{\alpha}$ determined by the constraints of eqs. (A2), (A3) and (A4). However, in the special case where the metric tensor is configuration independent (eq. (8)) the maximal entropy distribution satisfying the constraints is the normalized multivariate Gaussian

$$p_o(q) = (2\pi)^{-M/2} \left|\underline{\mathbf{C}}_q\right|^{-1/2} e^{-\frac{1}{2}(q-\overline{q})\underline{\mathbf{C}}_q^{-1}(q-\overline{q})} \tag{A13}$$

with the associated classical entropy (see eqs. (12), (21) and (39))

$$s_{cl,o} = k_B\left[M\left(1 - \frac{1}{2}\ln\beta\hbar^2\right) - \ln\xi + \frac{1}{2}\ln\left|\underline{\mathbf{A}}_q\right| + \frac{1}{2}\ln\left|\underline{\mathbf{C}}_q\right|\right] \tag{A14}$$

That $p_o(q)$ has the correct normalization, average and covariance is easily verified using eqs. (B6) and (B7) as well as the standard result

$$\int dx\,(x \otimes x)\, e^{-\alpha x^T \underline{\mathbf{Y}} x} = \frac{1}{2\alpha}\left(\frac{\pi}{\alpha}\right)^{M/2} \left|\underline{\mathbf{Y}}\right|^{-1/2} \underline{\mathbf{Y}} \tag{A15}$$

Therefore, one may state that at the classical level and based on a generalized coordinate system with a configuration-independent metric tensor, the multivariate Gaussian, *i.e.* the canonical distribution generated by an underlying harmonic model with a Hessian $\underline{\mathbf{H}}_q = \left(\beta \underline{\mathbf{C}}_q\right)^{-1}$, is the maximal entropy distribution compatible with a given average configuration and covariance matrix.

At the quantum-mechanical level, the situation is more complex and we will immediately assume the use of a coordinate system with configuration-independent metric tensor (eq. (8)). We initially focus on the case of a system with one single degree of freedom. This case has already been considered by Schlitter [48]. However, we believe that the maximal-entropy proof provided there is incomplete. For a single quantum-mechanical Cartesian degree of freedom q, associated with a potential energy function $\mathcal{V}(q)$ assumed symmetric with respect to $q = 0$ (setting the average $\overline{q}$ to zero and observing that the variance imposes a symmetric constraint, the maximum entropy can only correspond to a function $\mathcal{V}$ of even symmetry), solving the Schrödinger equation leads to a discrete set of energy levels $\{\epsilon_i\}$ associated with variances $\{c_i\}$. The corresponding populations $\{p_i\}$ in the canonical ensemble are $p_i[\mathcal{V}] = Q^{-1} e^{-\beta(\epsilon_i - \epsilon_o)}$

with

$$Q[\mathcal{V}] = \sum_{i=0}^{\infty} e^{-\beta(\epsilon_i - \epsilon_o)} \tag{A16}$$

The corresponding entropy may be written

$$s_{qm}[\mathcal{V}] = -k_B \sum_{i=0}^{\infty} p_i[\mathcal{V}] \ln p_i[\mathcal{V}] \tag{A17}$$

Here, we wish to determine the form of the potential energy function $\mathcal{V}$ that extremizes this entropy with the constraints

$$n = \sum_{i=0}^{\infty} p_i[\mathcal{V}] = 1 \tag{A18}$$

and

$$C = \sum_{i=0}^{\infty} p_i[\mathcal{V}] c_i[\mathcal{V}] \tag{A19}$$

enforcing the normalization of the population distribution together with a specified variance C (the average $\overline{q}$ is zero because $\mathcal{V}$ is assumed of even symmetry).

Extremizing the quantity $-k_B^{-1} s_{qm}$, using Lagrange multipliers μ and λ to enforce the constraints of eqs. (A18) and (A19), one has to solve

$$\sum_{i=0}^{\infty} \left(\ln p_i[\mathcal{V}] + \lambda c_i[\mathcal{V}] + \mu + 1\right) \delta p_i[\mathcal{V}, \delta\mathcal{V}] + \lambda \sum_{i=0}^{\infty} p_i[\mathcal{V}]\, \delta c_i[\mathcal{V}, \delta\mathcal{V}] = 0 \tag{A20}$$

for any (unconstrained) infinitesimal change $\delta\mathcal{V}$.

At this point, the derivation of Schlitter (eq. (5) in ref. 48) implicitly sets the second term in eq. (A20) to zero. In this case, the latter equation only involves derivatives with respect to the populations of the different energy levels, and may be recast into a series of independent equations, as (using eq. (A16))

$$-\ln Q - \beta\epsilon_i + \lambda c_i + \mu + 1 = 0 \tag{A21}$$

This equation is satisfied for a quantum-mechanical harmonic oscillator, namely (with the constraints of eqs. (A18) and (A19)) for

$$\mathcal{V}(q) = \frac{1}{2} m\omega^2 q^2 \tag{A22}$$

where m is the particle mass and the angular frequency ω is the solution of the equation

$$\begin{aligned} mC &= Q^{-1} \sum_{i=0}^{\infty} \mathrm{m} C_i\, e^{-\beta(\epsilon_i - \epsilon_o)} \\ &= \left(\sum_{i=0}^{\infty} e^{-i\beta\hbar\omega(\epsilon_i - \epsilon_o)}\right)^{-1} \sum_{i=0}^{\infty} \left(i + \frac{1}{2}\right) \hbar\omega^{-1} e^{-i\beta\hbar\omega} \\ &= \hbar\omega^{-1} \left(1 - e^{-\beta\hbar\omega}\right) \left[\frac{1}{2}\left(1 - e^{-\beta\hbar\omega}\right)^{-1} + e^{-\beta\hbar\omega}\right] \\ &= \hbar\omega^{-1} \left[\frac{1}{2} + \left(1 - e^{-\beta\hbar\omega}\right)^{-1} e^{-\beta\hbar\omega}\right] \end{aligned} \tag{A23}$$

Here, one has used the equations $\epsilon_i = \left(i + \frac{1}{2}\right)\hbar\omega$ and $mC_i = \left(i + \frac{1}{2}\right)\hbar\omega^{-1}$ for the harmonic oscillator, together with the standard result for the geometric series.

The associated quantum-mechanical entropy is given by eq. (42). Note, however, that the corresponding probability distribution is not a Gaussian, but a sum of the solutions to the Schrödinger equation for the quantum-mechanical harmonic oscillator (*i.e.* products of Gaussians and Hermite polynomials) weighted by the corresponding populations. However, due to the neglect of the second term in eq. (A20), the proof of Schlitter is either incomplete or incorrect.

It is possible that similar problems arise upon increasing the dimensionality of the system. For example, for a two-dimensional quantum-mechanical system with coordinates characterized by a vanishing average covariance, the underlying potential maximizing the entropy for given average variances along the two coordinates may possibly not be of the form

$$\mathcal{V}(q_1, q_2) = \mathcal{V}_1(q_1) + \mathcal{V}_2(q_2) \tag{A24}$$

as assumed by Schlitter in his analysis of the problem [48].

If the maximal entropy property of the (one- or multi-dimensional) harmonic oscillator is possibly not satisfied exactly at the quantum-mechanical level, it probably remains valid for most practical

applications (excluding systems already very close to harmonicity).

Appendix B

Here, we provide the derivation of eqs. (38) and (39) from eq. (37). For a classical conservative system in the canonical ensemble, the phase-space probability distribution may be written

$$\tilde{p}(q, p_q) = \frac{e^{-\beta \mathcal{H}(q, p_q)}}{\int dq\, dp_q\; e^{-\beta \mathcal{H}(q, p_q)}} \tag{B1}$$

The configurational probability distribution is then defined as

$$\begin{aligned} p(q) &= \int dp_q\; \tilde{p}(q, p_q) \\ &= \frac{\int dp_q\; e^{-\beta \mathcal{H}(q, p_q)}}{\int dq\, dp_q\; e^{-\beta \mathcal{H}(q, p_q)}} \end{aligned} \tag{B2}$$

Using eq. (4), the Hamiltonian may be written

$$\begin{aligned} \mathcal{H}(q, p_q) &= \mathcal{V}(q) \\ &\quad + \frac{1}{2} \underline{p}_q^T\; \underline{\mathbf{A}}_q^{-1}(q)\; \underline{p}_q \end{aligned} \tag{B3}$$

From eqs. (B1) and (B2), one easily shows that the quantity

$$\begin{aligned} \Delta &= \int dq\, dp_q\; \tilde{p}(q, p_q) \ln\left[\xi h^M\; \tilde{p}(q, p_q)\right] \\ &\quad - \int dq\; p(q) \ln p(q) \end{aligned} \tag{B4}$$

is given by

$$\begin{aligned} \Delta &= \ln \xi + \ln h^M \\ &\quad - \frac{\int dq\, dp_q \left\{\beta \mathcal{H}(q, p_q) + \ln\left[\int dp_q\; e^{-\beta \mathcal{H}(q, p_q)}\right]\right\} e^{-\beta \mathcal{H}(q, p_q)}}{\int dq\, dp_q\; e^{-\beta \mathcal{H}(q, p_q)}} \end{aligned} \tag{B5}$$

Using the standard results

$$\int dx\; e^{-\alpha x^T \underline{\mathbf{Y}} x} = \left(\frac{\pi}{\alpha}\right)^{M/2} |\underline{\mathbf{Y}}|^{-1/2} \tag{B6}$$

and

$$\begin{aligned} &\int dx\; x^T \underline{\mathbf{Y}}\, x\; e^{-\alpha x^T \underline{\mathbf{Y}} x} \\ &\quad = \frac{M}{2\alpha} \left(\frac{\pi}{\alpha}\right)^{M/2} |\underline{\mathbf{Y}}|^{-1/2} \end{aligned} \tag{B7}$$

for integrals over a M - variate Gaussian, one easily shows using (B3) that

$$\begin{aligned} &\int dp_q\; e^{-\beta \mathcal{H}(q, p_q)} \\ &\quad = \left(\frac{2\pi}{\beta}\right)^{M/2} |\underline{\mathbf{A}}_q(q)|^{1/2}\; e^{-\beta \mathcal{V}(q)} \end{aligned} \tag{B8}$$

and

$$\begin{aligned} &\int dp_q\; \mathcal{H}(q, p_q)\; e^{-\beta \mathcal{H}(q, p_q)} \\ &\quad = \left(\frac{2\pi}{\beta}\right)^{M/2} |\underline{\mathbf{A}}_q(q)|^{1/2} \left[\mathcal{V}(q) + \frac{M}{2\beta}\right] e^{-\beta \mathcal{V}(q)} \end{aligned} \tag{B9}$$

Inserting these results into eq. (B5), one gets

$$\begin{aligned} \Delta &= \ln \xi + \ln h^M \\ &\quad - \frac{\int dq \left\{\beta\left[\mathcal{V}(q) + \frac{M}{2\beta}\right] + \ln\left[\left(\frac{2\pi}{\beta}\right)^{M/2} |\underline{\mathbf{A}}_q(q)|^{1/2}\, e^{-\beta \mathcal{V}(q)}\right]\right\} |\underline{\mathbf{A}}_q(q)|^{1/2}\, e^{-\beta \mathcal{V}(q)}}{\int dq\, |\underline{\mathbf{A}}_q(q)|^{1/2}\, e^{-\beta \mathcal{V}(q)}} \\ &= \ln \xi + \ln h^M - \frac{M}{2} - \frac{M}{2} \ln \frac{2\pi}{\beta} - \frac{1}{2} \frac{\int dq \ln\left[|\underline{\mathbf{A}}_q(q)|\right] |\underline{\mathbf{A}}_q(q)|^{1/2}\, e^{-\beta \mathcal{V}(q)}}{\int dq\, |\underline{\mathbf{A}}_q(q)|^{1/2}\, e^{-\beta \mathcal{V}(q)}} \\ &= -\frac{M}{2}\left(1 - \ln \frac{\beta h^2}{2\pi}\right) + \ln \xi - \frac{1}{2} \int dq\; p(q)\; \ln |\underline{\mathbf{A}}_q(q)| \end{aligned} \tag{B10}$$

Inserting this result into eq. (37) leads to eq. (38).

Within the harmonic approximation (eq. (5)) and assuming that the metric tensor is nearly configuration-independent (eq. (8); *i.e.* based on the approximate Hamiltonian $\mathcal{H}_o$ of eq. (9)), eq. (38) applied to the generalized normal modes a_q as a particular case of q becomes

$$\begin{aligned} S_{\alpha o} &= k_B \left[\frac{M}{2}\left(1 - \ln \frac{\beta h^2}{2\pi}\right) - \ln \xi \right. \\ &\quad \left. + \frac{1}{2} \ln |\underline{\mathbf{A}}_e| - \sum_{m=1}^{M} \int da_{qm}\; p'_{\alpha m}(a_{qm})\; \ln p'_{\alpha m}(a_{qm})\right] \\ &= k_B \left\{\frac{M}{2}\left(1 - \ln \frac{\beta h^2}{2\pi}\right) - \ln \xi \right. \\ &\quad \left. - \sum_{m=1}^{M} \int da_{qm} \left(2\pi E_{qm}\right)^{1/2}\, e^{-\frac{1}{2} E_{qm}^2 a_{qm}^2} \ln\left[\left(2\pi E_{qm}\right)^{1/2}\, e^{-\frac{1}{2} E_{qm}^2 a_{qm}^2}\right]\right\} \\ &= k_B \left\{\frac{M}{2}\left(1 - \ln \frac{\beta h^2}{2\pi}\right) - \ln \xi \right. \\ &\quad \left. - \sum_{m=1}^{M}\left[-\frac{1}{2} \ln\left(2\pi E_{qm}\right) - \frac{1}{2}\right]\right\} \end{aligned} \tag{B11}$$

where $p'_{\alpha m}$ is the probability distribution along mode m (eq. (17)) and one has used $\underline{\mathbf{A}}_e = \mathbf{1}$ (which follows from eqs. (1), (3) and (11), eq. (B11)) simplifies to eq. (39).

Appendix C

Here, we investigate how the numerical integration of one and two-dimensional probability distributions (occurring in eqs. (46) and (48)) depends on the width of the histogram bins selected to discretize these distributions, and provide a criterion to choose a reasonable value for this parameter.

For a distribution in M dimensions involving variables $\{a_m \mid m = 1..M\}$ that is close to a multivariate Gaussian, it seems reasonable to choose the bin width along each dimension m proportional to the standard deviation $\alpha_m^{1/2}$ of the corresponding coordinate distribution, *i.e.* as

$$\Delta a_m = \kappa_d \, \alpha_m^{1/2} \tag{C1}$$

with

$$\alpha_m = \int da_m \, a_m^2 \, p(a_m) - \left[\int da_m \, a_m p(a_m) \right]^2$$

where κ_d is a parameter that may be optimized for a given dimensionality d of the integral to be evaluated numerically (in present case one for eq. (46) or two for (48)).

For a one-dimensional histogram ($d = 1$) defined by a set of bins k of widths Δa centered at $k\Delta a$, the integral involved in eq. (46) is approximated as

$$\begin{aligned} I &= \int da \; p(a) \ln p(a) \\ &\approx \Delta a \sum_k p(k\Delta a) \ln p(k\Delta a) \end{aligned} \tag{C2}$$

If Δa is chosen too small, each bin is occupied by at most one frame of the simulated trajectory. In this case, the probability evaluates to $(n_f \Delta a)^{-1}$ for all occupied bins and zero otherwise, n_f being the total number of frames in the trajectory, and one has

$$I = -\ln n_f - \ln \Delta a \quad \text{(too small } \Delta a\text{)} \tag{C3}$$

If Δa is chosen too large, all frames are in the same bin. In this case, the probability evaluates to $(\Delta a)^{-1}$ for this single bin and zero otherwise, and one has

$$I = \ln \Delta a \quad \text{(too large } \Delta a\text{)}. \tag{C4}$$

Both limiting equations correspond to incorrect results, showing a dependence of the evaluated integral on the bin size Δa (and, possibly, on the number of frames n_f), but no dependence on the actual distribution of the data. The two equations are easily generalized to the case of a two-dimensional histogram ($d = 2$), in which case Δa should be replaced by the product of the bin widths Δa and $\Delta a'$ along the two modes.

To estimate a reasonable value κ_1^o for the parameter κ_1 in eq. (C1) with $d = 1$, corresponding to the one-dimensional histogram involved in eq. (46), the following procedure may be used: (i) consider the approximate value I^o of the integral in eq. (C2) corresponding to a Gaussian distribution of variance α, *i.e.* $I^o = -(1/2)(1 + \ln 2\pi\alpha)$; (ii) set κ_1^o as the mid-point between the intersections of the horizontal line at I^o and the limiting lines of eqs. (C3) and (C4) in the graph of I as a function of $\ln \kappa_1$, leading to

$$\kappa_1^o = (1/2)\left[1 + \ln\left(\frac{2\pi}{n_f} \right) \right] \tag{C5}$$

The procedure is easily extended to estimate a reasonable value κ_2^o for the parameter κ_2 in eq. (C1) with $d = 2$, corresponding to the two-dimensional histogram involved in eq. (48). In this case $I^o = -(1/2)(2 + \ln 4\pi^2 \alpha\alpha')$ and one finds

$$\kappa_2^o = (1/2)\left[1 + \ln\left(\frac{2\pi}{n_f^{1/2}} \right) \right] \tag{C6}$$

Figure 12(a) shows the values of the one-dimensional integrals involving the probability distributions $p'_m(a_m)$ in eq. (46) for a sample set of eigenvectors (m = 1, 50, and 100), evaluated numerically (eq. (C2)) using different values of κ_1 (eq. (C1)). Both limiting lines of eqs. (C3) and

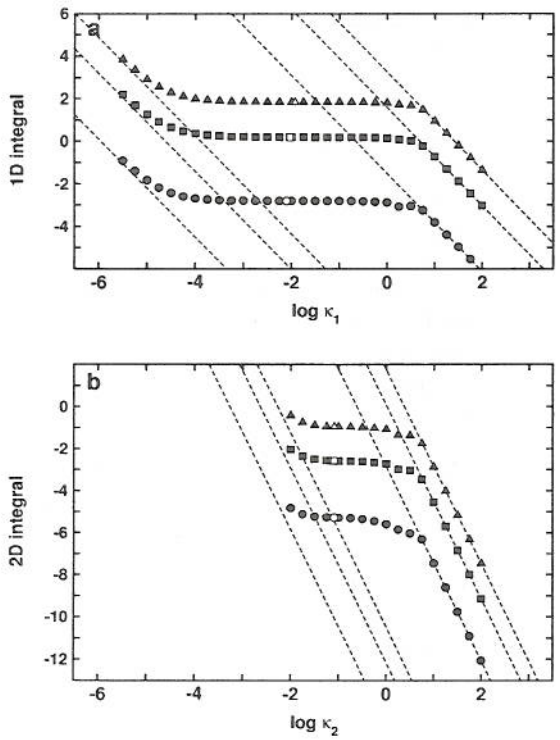

Figure 12. Dependence of the numerical integration of probability distributions on the width of histogram bins (Appendix C) for the β-hexapeptide simulation X_340_A (table 1). (a) Integrals over the one-dimensional distributions involved in eq. (46) are shown for eigenvectors 1 (circles), 50 (squares) and 100 (triangles). (b) Integrals over the two-dimensional distributions involved in eq. (48) are shown for eigenvector pairs 1,2 (circles), 1,50 (squares) and 1,100 (triangles). The results are displayed as a function of $\ln \kappa_1$ (a) or $\ln \kappa_2$ (b), where κ is the ratio of the bin width along each dimension to the corresponding distribution width (eq. (C1)). Unfilled symbols correspond to κ_1^o (eq. (C5)) and κ_2^o (eq. (C6)) values. The limiting lines for too fine (left side) and too coarse (right side) numerical integrations (see *e.g.* eqs. (C3) and (C4) for one dimension) are also displayed (dashed lines).

(C4) are shown, together with the value κ_1^o determined by eq. (C5). For each curve, a clear plateau defines the range of κ_1 values for which the integration result is essentially independent of the bin size. Finite-sampling artifacts affect the integration with smaller values of κ_1 while coarse-binning artifacts affect the integration with larger values.

Figure 12(b) shows the values of the two-dimensional integrals involving the probability distributions $p'_{mn}(a_m, a_n)$ in eq. (48) for a sample group of eigenvector pairs (m,n = 1,2; 1,50; 1,100), evaluated numerically with different values of κ_2 (eq. (C1), together with the corresponding limiting lines and the value κ_2^o determined by eq. (C6). Here, the plateau region is much narrower and the value of κ_2 has to be chosen more carefully. Intuitively, the reason is that the sampling of a two-dimensional histogram requires more points than that of a one-dimensional histogram.

REFERENCES

1. Falcioni, M., Loreto, V., and Vulpiani, A. 2003, in L'héritage de Kolmogorov en physique, Belin, Paris, 96.
2. Jaynes, E. T. 1957, Phys. Rev., 106, 620.
3. Jaynes, E. T. 1957, Phys. Rev., 108, 171.
4. Tanford, C. 1980, Science, 200, 1012.
5. Grunwald, E., and Steel, C. 1995, J. Am. Chem. Soc., 117, 5687.
6. Gallicchio, E., Kubo, M. M., and Levy, R. M. 2000, J. Phys. Chem. B, 104, 6271.
7. Fisicaro, E., Compari, C., and Braibanti, A. 2004, Phys. Chem. Chem. Phys., 6, 4156.
8. Graziano, G. 2005, J. Phys. Chem. B, 109, 12160.
9. Dill, K. A. 1990, Biochemistry, 29, 7133.
10. Privalov, P. L., and Makhatadze G. I. 1996, Protein Sci., 5, 507.
11. Tamura, A., and Privalov, P. L. 1997, J. Mol. Biol., 273, 1048.
12. Bicout, D. J., and Szabo, A. 2000, Protein Sci., 9, 452.
13. Roccatano, D., Di Nola, A., and Amadei, A. 2004, J. Phys. Chem. B, 108, 5756.
14. Chavez, L. L., Onuchic, J. N., and Clementi, C. 2004, J. Am. Chem. Soc., 128, 8426.

15. Daura, X. 2006, Theor. Chem. Acc., 116, 297.
16. Karplus, M., Ichiye, T., and Pettitt, B. M. 1987, Biophys. J., 52, 1083.
17. Lee, A. L., and Wand, A. J. 2001, Nature, 411, 501.
18. Stone, M. J. 2001, Acc. Chem. Res., 34, 379.
19. Trebbi, B., Dehez, F., Fowler, P. W., and Zerbetto, F. 2005, J. Phys. Chem. B, 109, 18184.
20. Searle, M. S., and Williams, D. H. 1992, J. Am. Chem. Soc., 114, 10690.
21. Tidor, B., and Karplus, M. 1994, J. Mol. Biol., 238, 405.
22. Wang, J., Morin, P., Wang, W., Kollman, P. A. 2001, J. Am. Chem. Soc., 123, 5221.
23. Gilson, M. K., Given, J. A., Bush, B. L., and McCammon, J. A. 1997, Biophys. J., 72, 1047.
24. Gallicchio, E., Kubo, M. M., and Levy, R. M. 1998, J. Am. Chem. Soc., 120, 4526.
25. Zidek, L., Novotny, M. V., and Stone, M. J. 1999, Nat. Struct. Biol., 6, 1118.
26. Sham, Y. Y., Chu, Z. T., Tao, H., and Warshel, A. 2000, Proteins Struct. Funct. Genet., 39, 393.
27. Gohlke, H., and Klebe, G. 2002, Angew. Chem. Int. Ed., 41, 2645.
28. Carlsson, J., and Åqvist, J. 2005, J. Phys. Chem. B, 109, 6448.
29. Homans, S. W. 2005, Chem. BioChem., 6,1585.
30. Villà, J., Štrajbl, M., Glennon, T. M., Sham, Y. Y., Chu, Z. T., and Warshel, A. 2000, Proc. Natl. Ac. Sci. USA, 97, 11899.
31. Dunitz, J. D. 1995, Chem. Biol., 2, 709.
32. Cooper, A. 1999, Curr. Opin. Chem. Biol., 3, 557.
33. Beveridge, D. L., and DiCapua, F. M. 1989, Annu. Rev. Biophys. Biophys. Chem., 18, 431.
34. Straatsma, T. P., and McCammon, J. A. 1992, Annu. Rev. Phys. Chem., 43, 407.
35. King, P. M. 1993, Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson (Eds.), vol.2, Escom Science, Leiden, 267.
36. Kollman, P. 1993, Chem. Rev., 93, 2395.
37. van Gunsteren, W. F., Beutler, T. C., Fraternali, F., King, P. M., Mark, A. E., and Smith, P. E. 1993, in Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, W. F. Gunsteren, P. K. Weiner, and A. J. Wilkinson (Eds.), vol. 2, Escom Science, Leiden, 315.
38. Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., and van Gunsteren, W. F. 1994, Chem. Phys. Lett., 222, 529.
39. Straatsma, T. P. 1996, in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd (Eds.), vol. 9, VCH Publishers, New York, 81.
40. Mark, A. E. 1998, in Encyclopedia of Computational Chemistry, P. Ragué Schleyer (Ed.), vol. 2, Wiley, New York, 1070.
41. Reinhardt, W. P., Miller, M. A., and Amon, L. M. 2001, Acc. Chem. Res., 34, 607.
42. Chipot, C., and Pearlman, D. A. 2002, Mol. Simul., 28, 1.
43. van Gunsteren, W. F., Daura, X., and Mark, A. E. 2002, Helv. Chim. Acta, 85, 3113.
44. Karplus, M., and Kushick, J. 1981, Macromolecules, 17, 325.
45. Di Nola, A., Berendsen, H. J. C., and Edholm, O. 1984, Macromolecules, 17, 2044.
46. Edholm, O., and Berendsen, H. J. C. 1984, Mol. Phys., 51, 1011.
47. Rojas, O. L., Levy, R. M., Szabo, A. 1986, J. Chem. Phys., 85, 1037.
48. Schlitter, J. 1993, Chem. Phys. Lett., 215, 617.
49. Schäfer, H., Mark, A. E., and van Gunsteren, W. F. 2000, J. Chem. Phys. 113, 7809.
50. Andricioaei, I., and Karplus, M. 2001, J. Chem. Phys., 115, 6289.
51. Lin, S. T., Blanco, M., and Goddard, W. A. 2003, J. Chem. Phys. 119, 11792.
52. Peter, C., Oostenbrink, C., van Dorp, A., and van Gunsteren, W. F. 2004, J. Chem. Phys., 120, 2652.
53. Chang, C-E., Chen, W., and Gilson, M. K. 2005, J. Chem. Theory Comput., 1, 1017.
54. Zwanzig, R. W. 1954, J. Chem. Phys., 22, 1420.

55. Smith, D. E., and Haymet, D. J. 1993, J. Chem. Phys., 98, 644556. Lu, N., Kofke, D. A., and Woolf, T. B. 2003, J. Phys. Chem. B, 107, 5598.
56. Lu, N., Kofke, D. A., and Woolf, T. B. 2003, J. Phys. Chem. B, 107, 5598.
57. Guillot, B., Guissani, Y., and Bratos, S. 1991, J. Chem. Phys., 95, 3643.
58. Guillot, B., and Guissani, Y. 1993, J. Chem. Phys., 99, 8075.
59. Torrie, G. M., and Valleau, J. P. 1977, J. Comput. Phys., 23, 187.
60. Fleischman, S. H., and Brooks, C. L. 1987, J. Chem. Phys., 87, 3029.
61. Levy, R. M., Karplus, M., Kushick, J. N., and Perahia, D. 1984, Macromolecules, 17, 1370.
62. Irikura, K. K., Tidor, B., Brooks, B. R., and Karplus, M. 1985, Science, 229, 571.
63. Samson, C., Lüthi, H. P., and Hünenberger, P. H., in preparation.
64. Head, M. S., Given, J. A., and Gilson, M. K. 1997, J. Phys. Chem. A, 101, 1609.
65. Smith, P. E., and van Gunsteren, W. F. 1994, J. Mol. Biol., 236, 629.
66. Eckart, C. 1935, Phys. Rev., 47, 552.
67. Wilson Jr., E. B., Howard, J. B. 1936, J. Chem. Phys., 4, 260.
68. Sayvetz, A. 1939, J. Chem. Phys., 6, 383.
69. Schieborr, U., and Rüterjans, H. 2001, Proteins Struct. Funct. Genet., 45, 207.
70. Darian, E., Hnizdo, V., Fedorowicz, A., Singh, H., and Demchuck, E. 2005, J. Comput. Chem., 26, 651.
71. Kabsch, W., Kabsch, H., and Eisenberg, D. 1976, J. Mol. Biol., 100, 283.
72. McLachlan, A. D. 1979, J. Mol. Biol., 128, 49.
73. Amadei, A., Chillemi, G., Ceruso, M. A., Grottesi, A., and Di Nola, A. 2000, J. Chem. Phys., 112, 9.
74. Schäfer, H., Daura, X., Mark, A. E., van Gunsteren, W. F. 2001, Proteins Struct. Funct. Genet., 43: 45.
75. Baron, R., Bakowies, D., and van Gunsteren, W. F. 2005, J. Peptide Sci., 11, 74.
76. Mu, Y., and Stock, G. 2002, J. Phys. Chem. B, 106, 5294.
77. Baron, R., de Vries, A. H., Hünenberger, P.H., and van Gunsteren, W. F. 2006, J. Phys. Chem. B, 110, 8464.
78. Schäfer, H., Smith, L. J., Mark, A. E., van Gunsteren, W. F. 2002, Proteins Struct. Funct. Genet., 46, 215.
79. Bakowies, D., and van Gunsteren, W. F. 2002, J. Mol. Biol., 315, 713.
80. Hsu, S. D., Peter, C., van Gunsteren, W. F., and Bonvin, A. M. J. J. 2004, Biophys. J., 88, 15.
81. Dixit, S. B., Andrews, D. Q., and Beveridge, D. L. 2005, Biophys. J., 88, 3147.
82. Dolenc, J., Baron, R., Oostenbrink, C., Koller, J., and van Gunsteren, W. F. 2006, Biophys. J., 91, 1460.
83. Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. 1993, Proteins Struct. Funct. Gent., 17, 412.
84. van Aalten, D. M. F., Amadei, A., Linssen, A. B. M., Eijsink, V. G. H., Vriend, G., Berendsen, H. J. C. 1995, Proteins Struct. Funct. Genet., 22, 45.
85. Amadei, A., Linssen, A. B. M., De Groot, B. L., van Aalten, D. M. F., and Berendsen, H. J. C. 1996, J. Biomol. Str. Dynamics, 13, 615.
86. van Aalten, D. M. F., De Groot, B. L., Findlay, J. B. C., Berendsen, H. J. C., and Amadei, A. 1997, J. Comput. Chem., 18, 169.
87. Gō, N., Noguti, T., and Nishikawa, T. 1983, Proc. Natl. Acad. Sci. USA, 80, 6571.
88. Brooks, B., and Karplus, M. 1983, Proc. Natl. Acad. Sci. USA, 80, 6571.
89. Levitt, M., Sander, C., and Stern, P. S. 1983, Int. J. Quant. Chem., 10, 181.
90. Levitt, M., Sander, C., and Stern, P. S. 1985, J. Mol. Biol., 181, 423.
91. Nishikawa, T., and Gō, N. 1987, Proteins Struct. Funct. Genet., 2, 308.
92. Perahia, D., Levy, R. M., and Karplus M. 1990, Biopolymers, 29, 645.
93. Case, A. 1994, Curr. Opin. Struct. Biol., 4, 285.
94. Hayward, S., and Gō, N. 1995, Annu. Rev. Phys. Chem., 46, 233.
95. Hayward, S., Kitao, A., and Berendsen, H. J. C. 1997, Proteins Struct. Funct. Genet., 27, 425.
96. Reiher, M., and Neugebauer, J. 2003, J. Chem. Phys., 118, 1634.
97. Cui, Q., Li, G., Ma, J., and Karplus, M. 2004, J. Mol. Biol., 340, 345.

98. Daura, X., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. 1998, J. Mol. Biol., 280, 925.
99. Daura, X., van Gunsteren, W. F., and Mark, A. E. 1999, Proteins Struct. Funct. Genet., 34, 269.
100. Daura, X., Mark, A. E., and van Gunsteren, W. F. 1999, Comput. Phys. Comm. 123, 97.
101. Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. 1999, Angew. Chem. Int. Ed., 38, 236.
102. Daura, X., Gademann, K., Schäfer, H., Jaun, B., Seebach, D., and van Gunsteren, W. F. 2001, J. Am. Chem. Soc., 123, 2393.
103. Baron, R., Bakowies, D., van Gunsteren, W. F., and Daura, X. 2002, Helv. Chim. Acta, 85, 3872.
104. Hünenberger, P. H., Mark, A. E., and van Gunsteren, W. F. 1995, J. Mol. Biol., 252, 492.
105. Garcia, A. E. 1992, Phys. Rev. Lett., 68, 2696.
106. Hess, B. 2000, Phys. Rev. E, 62, 8438.
107. Kitao, A., and Gō, N. 1999, Curr. Opin. Struct. Biol., 9, 164.
108. de Groot, B. L., Daura, X., Mark, A. E., and Grubmüller, H. 2001, J. Mol. Biol., 309, 299.
109. Hess, B. 2002, Phys. Rev. E, 65, Art. No. 031910.
110. Tournier, A. L., and Smith, J. C. 2003, Phys. Rev. Lett., 91, Art. No. 208106.
111. Fixman, M. 1978, J. Chem. Phys., 69, 1527.
112. van Gunsteren, W. F. 1980, Mol. Physics, 40, 1015.
113. van Gunsteren, W. F., and Karplus, M. 1982, Macromolecules, 15, 1528.
114. van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P., and Tironi, I. G. 1996, The GROMOS96 Manual and User Guide, Vdf Hochschulverlag AG, Zürich.
115. Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Krüger, P., and van Gunsteren, W. F. 1999, J. Phys. Chem. A, 103, 3596.
116. van Gunsteren, W. F., Daura, X., and Mark, A. E. 1998, in Encyclopedia Comput. Chem., vol. 2, 1211.
117. Daura, X., Mark, A. E., and van Gunsteren, W.F. 1998, J. Comput. Chem., 19, 535.
118. Walser, R., Mark, A. E., van Gunsteren, W. F., Lauterbach, M., and Wipff, G. 2000, J. Chem. Phys., 112, 10450.
119. Ichiye, T., and Karplus, M. 1987, Proteins Struct. Funct. Genet. 2, 236.
120. Dill, K. A., Fiebig, K. M., and Chan, H. S. 1993, Proc. Natl. Acad. Sci. USA, 90, 1942.
121. Chan, H. S., Bromberg, S., and Dill, K. A. 1995, Phil. Trans. R. Soc. Lond. B, 348, 61.
122. Pereira, C. S., Kony, D., Baron, R., Müller, M., van Gunsteren, W. F., and Hünenberger, P. H. 2006, Biophys. J., 90, 4337.