Adaptive anisotropic kernels for nonparametric estimation of absolute configurational entropies in high-dimensional configuration spaces

Ulf Hensen,[1] Helmut Grubmüller,[1] and Oliver F. Lange[2,*]
[1]*Department of Theoretical Biophysics, Max-Planck Institut für biophysikalische Chemie, 37070 Göttingen, Germany*
[2]*Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA*
(Received 29 November 2008; revised manuscript received 19 March 2009; published 20 July 2009)

The quasiharmonic approximation is the most widely used estimate for the configurational entropy of macromolecules from configurational ensembles generated from atomistic simulations. This method, however, rests on two assumptions that severely limit its applicability, (i) that a principal component analysis yields sufficiently uncorrelated modes and (ii) that configurational densities can be well approximated by Gaussian functions. In this paper we introduce a nonparametric density estimation method which rests on adaptive anisotropic kernels. It is shown that this method provides accurate configurational entropies for up to 45 dimensions thus improving on the quasiharmonic approximation. When embedded in the minimally coupled subspace framework, large macromolecules of biological interest become accessible, as demonstrated for the 67-residue coldshock protein.

DOI: 10.1103/PhysRevE.80.011913 PACS number(s): 87.15.H−, 65.40.gd, 05.70.−a

I. INTRODUCTION

Entropies are key quantities in physics, chemistry, and biology. While free energy changes govern the direction of all chemical processes including reaction equilibria, entropy changes are the underlying driving forces of ligand binding, protein folding or other phenomena driven by hydrophobic forces. Atomistic simulations, e.g., molecular dynamics, in principle provide all the information needed for calculating both, free energies and entropies.

Generally, one can distinguish three types of approaches to obtain these quantities from computer simulations. One type of methods evaluates the partition sum directly to obtain *absolute* values. The second type of methods obtains *differences* by special purpose perturbation simulation techniques. Third, scanning procedures, introduced by Meirovitch [1,2], aim at reconstructing molecular probability distributions in a step-wise fashion.

For free-energy differences, methods of the perturbation type are commonly used, e.g., free-energy perturbation or thermodynamic integration which allow to exploit cancellation of terms in the difference of the partition sums of the two states. Thus, it suffices to sample those degrees of freedom which are affected by the perturbation which are usually few (e.g., regions around binding sites). If the two states differ substantially, however, sampling suffers from slow convergence.

For entropy differences, however, the full Hamiltonian contributes to the result, such that here complete sampling of the full phase space is indeed required [3,4]. Thus, perturbation approaches do not share the same fundamental advantage over a direct evaluation of the partition sum to obtain entropies as they do for free-energy differences.

Direct approaches have additional advantages. First, such methods are independent from finding suitable perturbation pathways between the states of interest and/or analytical tractable reference states. Furthermore, to gain a detailed understanding of the molecular mechanisms, one would like to separate various contributions to the entropy (changes), e.g., side-chain, backbone or ligands. Using a direct method, all these analyses can be carried out on the same computational ensembles.

However, it remains notoriously difficult to directly obtain accurate entropies even from well-sampled computational ensembles [3,5–8]. The difficulty arises from the necessary evaluation of the integral for the configurational part of the entropy,

$$S_c \sim -\int \rho(\mathbf{x}) \ln \rho(\mathbf{x}) d\mathbf{x}, \tag{1}$$

in the $3N$-dimensional configurational space of the system, where N is the number of atoms. Since N is usually large, e.g., on the order of several hundred or thousands for proteins, this integral cannot be evaluated directly.

Thus, for larger molecules one usually has to resort to a harmonic or quasiharmonic (QH) approximation of the system [9,10], which renders the integral Eq. (1) separable. Accordingly, the (quasi) harmonic entropy $S_{c,\mathrm{harm}}=\Sigma_i^{3N} S_c(i)$ is given as a sum of entropies $S_c(i)$ of the individual (quasi) harmonic modes. Using the partition sum of the quantum-mechanical oscillator for the entropies $S_c(i)$ of individual modes, this estimate has been shown to provide a rigorous upper limit of the true entropy [9,10]. However, for macromolecules the true configurational entropy is typically considerably smaller due to coupling of the principal components and anharmonicity, i.e., multimodality. Although this problem has already been pointed out quite early [9,10], little is known about how large the overestimation actually is. Indeed, model calculations on systems as the ideal gas [7], Lennard-Jones fluids [7], small linear alkane molecules [11], or small peptides [12,13], however, reveal drastic differences between the quasiharmonic entropy estimates and the actual entropy. For complex macromolecules, the extent of this effect is unknown.

*olange@u.washington.edu

By construction, methods based on nonparametric density estimation include couplings, anharmonicities as well as multimodalities and should therefore be capable of providing accurate results. However, in the high-dimensional spaces of macromolecules the "curse of dimensionality" [14] prevents density estimation. In order to render density estimation applicable in the context of macromolecular entropies, nonetheless, we recently suggested [15] a novel hierarchical approach consisting of three components; first, full correlation analysis (FCA) to find an orthogonal coordinate transformation that optimally decouples the configurational space into lower dimensional subspaces that have negligible coupling between them. The configurational entropy of the overall system is thus well approximated by the sum of the contributions from these subspaces. The contributions of each of these subspaces are obtained by, second, suitable density estimation with non-negligible couplings, third, taken into account via a mutual information expansion (MIE) [16,17], which bears close resemblance to the inclusion-exclusion principle/sieve formula known from set theory [18].

For biological macromolecules, it must be expected that this hierarchical approach requires accurate density estimates for increasingly high-dimensional spaces. Already for the relatively small protein calmodulin [15], the minimally coupled subspaces derived from FCA are up to 100 dimensional and, therefore, roughly 50-dimensional density estimates are required to keep the number of terms for MIE sufficiently small. However, current density estimators have been shown to yield accurate results for only up to 10 internal degrees of freedom, whereas failure was reported for a molecular system with 23 internal degrees of freedom [19]. In this study, we therefore develop a density estimator which can be used in our hierarchical approach. Full correlation analysis as applied here has been reported elsewhere [20]. The full approach has been benchmarked on a variety of test systems [15], and has been applied to calmodulin [15] and insulin [21].

To derive our density estimator, note that the intricate mix of stiff and soft degrees of freedom [22] typically found in macromolecules such as proteins yields a configurational density that is *threaded*, i.e., the density is extended in spatial directions that correspond to "soft" degrees of freedom, whereas it is confined to thin regions of space in the "stiff" directions. Accordingly, the established methods which average over isotropic regions in space will considerably blur such density threads locally perpendicular to the thread.

As an alternative, estimates in dihedral space have been considered [9,23] to ease the problem with stiff and soft degrees of freedom. These however suffer from a number of complications such as very large gradients in the energy landscape, complex topology of the dihedral variables or metric tensor effects due to necessarily approximate Jacobians [12]. Due to these severe drawbacks of dihedral coordinates we abandoned this approach in favor of Cartesian space.

We show here that nonparametric entropy estimates in Cartesian space are considerably improved by using *anisotropic* kernels which are locally adapted to the threads in the configurational density. The increase of entropy due to regularization employed during the estimation process is corrected for by an empirical term parameterized by the average volume of the used regularization kernel. Finally, generalizing the quasiharmonic Schlitter formula [10], we account for the quantum mechanical nature of the stiffest degrees of freedom. To this aim, we employ here a size limit to each principal axis of the elliptic averaging kernel to prohibit negative entropy contributions from the stiffest degrees of freedom.

II. THEORY

A. Thermodynamic entropy

We first sketch the conceptual framework to clarify notation. The molecular dynamics of an isolated macromolecule with N atoms is described by the Hamiltonian

$$H(\mathbf{x},\mathbf{p}) = \frac{1}{2}\sum_{i=1}^{3N} \mathbf{p}_i^2 + V(\mathbf{x}),$$

where $\mathbf{x}$ and $\mathbf{p}$, are the mass-weighted $3N$-dimensional position and momentum vectors, respectively. Their Cartesian components x_i and p_i are related to the non-mass-weighted components $\tilde{x}_i$ and $\tilde{p}_i$ by $x_i = m_i^{1/2}\tilde{x}_i$ and $p_i = m_i^{-1/2}\tilde{p}_i$, respectively, where m_i denotes the mass of the respective atom. The entropy of the system is given by

$$S = \frac{\langle H \rangle}{T} + k_B \ln Z,$$

where the angular brackets $\langle . \rangle$ denote an ensemble average, T the temperature, and Z is the classical partition function

$$Z = \frac{1}{h^{3N}} \int e^{-\beta H} d\mathbf{p} d\mathbf{x},$$

with h being Planck's constant, and $\beta = 1/k_B T$.

For conservative systems, Z separates into two-dimensionless factors, $Z = Z_p Z_x$, with

$$Z_p = \frac{1}{\kappa^{3N}} \left(\frac{2\pi}{\beta} \right)^{3N/2} \quad \text{and}$$

$$Z_x = \frac{1}{\ell^{3N}} \int e^{-\beta V(\mathbf{x})} d\mathbf{x}.$$

Here, we have defined a convenient characteristic length $\ell = 1$ nm $\mathrm{u}^{1/2}$ and, similarly, $\kappa = h/\ell$. Accordingly, the entropy of the system falls into a kinetic and a configurational part,

$$S = S_p + S_c,$$

with

$$S_p = \frac{3}{2} N k_B \left(1 + \ln \frac{2\pi}{\kappa^2 \beta} \right) \quad \text{and}$$

$$S_c = \frac{\langle V(\mathbf{x}) \rangle}{T} + k_B \ln Z_x. \tag{2}$$

Using the configurational probability density function, $\rho(\mathbf{x}) = \exp[-\beta V(\mathbf{x})]/Z_x$, we express the configurational entropy as

$$S_c = -\frac{k_B}{\ell^{3N}} \int \rho(\mathbf{x}) \ln \rho(\mathbf{x}) d\mathbf{x}, \tag{3}$$

which closely resembles Shannon's information entropy.

B. Quasiharmonic approximation

To estimate S_c from a given (finite) ensemble of structures $\{\mathbf{x}\}$ obtained, e.g., from a molecular dynamics or Monte Carlo simulation, the quasiharmonic approximation is commonly used with

$$\rho(\mathbf{x}) \approx \tilde{\rho}(\mathbf{x}) = \exp\left[-\frac{1}{2} (\mathbf{x} - \langle \mathbf{x} \rangle)^{\mathbf{T}} \mathbf{A} (\mathbf{x} - \langle \mathbf{x} \rangle) \right],$$

with $\mathbf{A} = \mathbf{C}^{-1}$, where $\mathbf{C}$ denotes the covariance matrix

$$\mathbf{C} = \langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^T \rangle.$$

In this approximation the configurational density factorizes,

$$\rho_{\mathrm{QH}}(\mathbf{y}) = \prod_{i=1}^{3N} \rho_i(\mathbf{y}_i), \tag{4}$$

with marginal densities $\rho_i(y_i) \propto \exp(-\beta y_i^2 / 2\lambda_i)$, where $\mathbf{y} = \mathbf{T}(\mathbf{x} - \langle \mathbf{x} \rangle)$ are the principal coordinates derived from diagonalization of $\mathbf{C}$, i.e., $\mathbf{T}^T \mathbf{C} \mathbf{T} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{3N})$. Using the orthonormality of $\mathbf{T}$, and combining Eqs. (2) and (3), the quasiharmonic entropy estimate,

$$S_{\mathrm{QH}} = \frac{1}{2} k_B \sum_{i=1}^{3N} \ln \left[\frac{e^2 \lambda_i}{\beta \hbar^2} \right], \tag{5}$$

is obtained.

We note that this estimate holds only for the molecular frame, with rigid body motions properly removed. Equation (5) has been elegantly generalized [10] to account for the quantum mechanical nature of the vibration of stiff degrees of freedom,

$$S_{\mathrm{qm,harm}} = \frac{1}{2} k_B \sum_{i=1}^{3N} \ln \left(1 + \frac{e^2}{\beta \hbar^2} \lambda_i \right). \tag{6}$$

This approximation, which will be used below, in particular avoids divergencies for $\lambda_i \to 0$, which is an artifact of the classical treatment, Eq. (5).

C. Locally adapted nonparametric entropy estimation

We will use this framework to develop a adaptive anisotropic nonparametric density estimation based on the k-nearest-neighbor (NN) density estimate of n sample points $\{\mathbf{x}\}$ which we will apply to $d \leq 3N$ dimensions,

$$\hat{S}_k = \frac{k_B}{n} \sum_{i=1}^{n} \ln \frac{n \pi^{d/2} r_{i,k}^d}{k \Gamma\left(\frac{1}{2} d + 1\right) \ell^d} + \frac{d}{3N} S_p. \tag{7}$$

Here, $r_{i,k}$ is the Euklidean distance between the sample point $\mathbf{x}_i$ and its k nearest neighbors, and Γ denotes the gamma function.

The k-NN estimator rests on the straightforward assumption of the local density at sample point $\mathbf{x}_i$

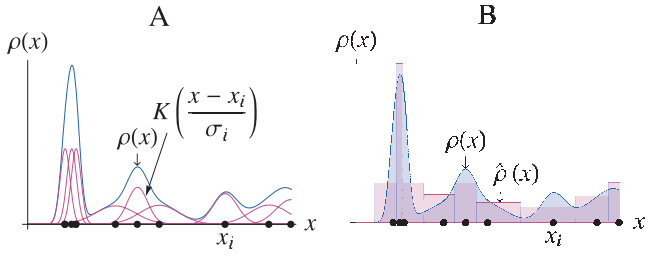

FIG. 1. (Color) Principle of kernel density estimation with Gaussian kernels. Other kernels were also tested. (a) The configurational space density $\rho(\mathbf{x})$ vs the local kernel approximation at sample points $\mathbf{x}_i$, indicated as black dots. (b) The configurational space density is approximated as sum, $\hat{\rho}(\mathbf{x})$, of locally constant densities of Voronoi cells, $\hat{\rho}(\mathbf{x}_i)$, around $\mathbf{x}_i$.

$$\rho_k(\mathbf{x}_i) \approx \frac{k \ell^d}{n V_d[r_i(k)]}, \tag{8}$$

where $V_d(r_{i,k})$ denotes the volume of the d-dimensional sphere with radius $r_i(k)$ which is chosen such that the d-dimensional sphere centered at $\mathbf{x}_i$ contains k sample points.

1. Soft degrees of freedom

A generalisation of Eq. (7) to arbitrary kernel functions K for given k and for each point $\mathbf{x}_i$ is obtained by requiring $\sigma_{i,k}$ to be chosen such that

$$k = \sum_{j=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\sigma_{i,k}} \right). \tag{9}$$

This yields a potentially improved local (smoothed) density estimate at $\mathbf{x}_i$,

$$\rho(\mathbf{x}_i) \approx \rho_k(\mathbf{x}_i) = \frac{k \ell^d}{n \sigma_{i,k}^d Z_d}, \tag{10}$$

where $Z_d = \ell^{-d} \int K(\mathbf{x} - \mathbf{x}_i) d\mathbf{x}$. The probability density distribution $\rho(\mathbf{x})$ is then approximated as a sum of these local densities [Fig. 1(b)],

$$\rho(\mathbf{x}) \approx \hat{\rho}(\mathbf{x}) := \sum_{i=1}^{n} \theta_i(\mathbf{x} - \mathbf{x}_i) \rho_k(\mathbf{x}_i),$$

where the Voronoi function θ_i is unity for all $\mathbf{x}$ that are closer to $\mathbf{x}_i$ than to any other $\mathbf{x}_j$, $i \neq j$, and zero otherwise. This approximation dissects the entropy integral, Eq. (3),

$$S_c \approx -k_B \int \hat{\rho}(\mathbf{x}) \ln \hat{\rho}(\mathbf{x}) d\mathbf{x} \approx -k_B \sum_{i=1}^{n} \ln \hat{\rho}(\mathbf{x}_i),$$

where the volume of the Voronoi cells was assumed to be given by the inverse local density.

Accordingly, the entropy contribution from the d considered degrees of freedom is estimated by

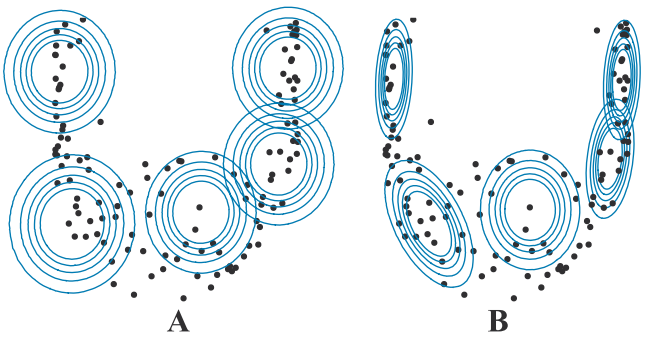

FIG. 2. (Color) Isotropic vs adaptive anisotropic kernels. For threaded configurational space densities (sampled by the black dots), isotropic kernels (a) fail to provide accurate density approximations. In contrast, adaptive anisotropic kernels (b) improve the approximation particular perpendicular to the threads.

$$\hat{S}_{\text{g-NN}} = \frac{k_B}{n}\sum_{i=1}^{n} \ln \frac{nZ_d\sigma_{i,k}^d}{k\ell^d} + \frac{d}{3N}S_p, \tag{11}$$

which generalizes Eq. (7) and, thus, may be denoted in the following as g-NN. The k-NN formula is recovered by using the rectangular and isotropic kernel function

$$K_{k\text{-NN}}\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\sigma_{i,k}}\right) = \begin{cases} 1 & \|\mathbf{x}_i - \mathbf{x}_j\| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and setting $\sigma_{i,k}=r_{i,k}$, since for this kernel $Z_d=V_d(1)$.

The Gaussian function framework underlying the quasiharmonic approach sketched above suggests to use instead a Gaussian kernel [24]

$$K(\mathbf{x}) = \exp(-\|\mathbf{x}^2\|). \tag{12}$$

However, as illustrated in Fig. 2, for such isotropic kernels, the approximation Eq. (11) tends to fail for the thin "threads" typically encountered for the configurational density of biological macromolecules, i.e., where the width of the "thread" is smaller than the average distance between adjacent sample points (A). To address this issue, we propose to locally adapt the kernel function to the stiff degrees of freedom using for each $\mathbf{x}_i$ an anisotropic Gaussian kernel (B),

$$K_{\text{loc,gauss}}(\mathbf{x}) = \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T\mathbf{A}_{\text{loc}}(\mathbf{x}-\mathbf{x}_i)\right], \tag{13}$$

with

$$\mathbf{A}_{\text{loc}}^{-1} = \mathbf{C}_{\text{loc}} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^T,$$

given by the k-local covariance matrix $\mathbf{C}$ and the sum runs over the k nearest neighbors of $\mathbf{x}_i$. Elliptic hard kernels are obtained analogously

$$K_{\text{loc,ell}}(\mathbf{x}) = \begin{cases} 1 & \|(\mathbf{x}-\mathbf{x}_i)^T\mathbf{A}(\mathbf{x}-\mathbf{x}_i)\| \le 1 \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

2. Stiff degrees of freedom—quantum correction

The adaptive anisotropic Gaussian kernels naturally define stiff degrees of freedom in a canonical way, and, for those, it is possible to obtain physically appropriate quantum corrections. The classical treatment of those stiff degrees of freedom yields unphysical entropies by allowing unlimited sharpness of the "threads." In contrast, the Schlitter formula, Eq. (6), effectively requires stiff degrees of freedom to assume density distributions of a minimal width $\sigma_{\text{qm}}=\beta\hbar^2/e^2$. Schlitter's treatment can be generalized to arbitrary densities by computing the entropy from

$$\rho_\sigma(\mathbf{x}) = \int \rho(\mathbf{x})K(\mathbf{y}-\mathbf{x}/\sigma_{\text{qm}})dy, \tag{15}$$

rather than from the true density $\rho(\mathbf{x})$, where $K(\mathbf{y})$ denotes a d variate smoothing kernel of unit width. Applying this convolution to the quasiharmonic density Eq. (4) with a Gaussian smoothing kernel K (Eq. (12)) of width σ_{qm} yields a smeared out multivariate Gaussian density whose widths are given by $\lambda_i+\sigma_{\text{qm}}$. Accordingly, one retrieves Schlitter's formula, Eq. (6).

In contrast to Schlitter's original treatment, our generalisation also holds for arbitrary densities. As an illustration consider, e.g., a bimodal density of two nonoverlapping Gaussian functions,

$$\rho(\mathbf{x}) = \frac{\ell}{2\varepsilon}\sqrt{\frac{1}{2\pi}}\left\{\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_0)^2}{2\varepsilon^2}\right] + \exp\left[-\frac{(\mathbf{x}+\mathbf{x}_0)^2}{2\varepsilon^2}\right]\right\},$$

with $x_0 \gg \varepsilon$. Convolution according to Eq. (15) yields the entropy $S(\varepsilon)=k_B \ln 2$ for $\varepsilon \to 0$, demonstrating that this generalisation yields meaningful results beyond the quasiharmonic approximation.

The smoothing inherent to the locally adapted kernel density estimation, therefore, simultaneously provides an approximate and canonical treatment of the stiff degrees of freedom. Accordingly, we restrict the width $\sigma_{j,i}$ of the g-NN smoothing kernel i in the local direction j to σ_{qm} by $\sigma_{j,i,\text{qm}} = \max(\sigma_{i,j}, \sigma_{\text{qm}})$, where the minimal width is set to

$$\sigma_{\text{qm}} = \frac{2\pi\hbar^2\beta}{eZ_d^{2/d}},$$

such that if $\sigma_{i,k}=\sigma_{\text{qm}}$,

$$\hat{S}_{\text{g-NN}} = \frac{k_B}{n}\sum_{i=1}^{n}\ln\frac{n}{k}.$$

Thus, $\hat{S}_{\text{g-NN}} \to 0$ for $k \to n$, in the special case that the density is concentrated within a single region below the quantum resolution limit and $\hat{S}_{\text{g-NN}}=k_B \log M$ if the density is distributed equally between M nonoverlapping areas in phase space that are all narrower than the quantum resolution limit.

3. Empirical smoothing correction

A correction for the smoothing of the configurational density implied by the convolution with the g-NN smoothing kernel functions within our nearest-neighbor approximation was derived from Eq. (5). Inverting Eq. (5) yields an effec-

tive width λ_{eff} of a hypothetical isotropic d-dimensional Gaussian density function that would correspond to a given entropy S,

$$\lambda_{\text{eff,qh}} = \frac{\hbar^2\beta}{e^2}\exp\left(\frac{2}{dk_B}S\right),$$

A deviation of the entropy estimated by the g-NN method $S_{\text{g-NN}}$ from the true entropy S can thus be expressed as $\Delta\lambda = \lambda_{\text{eff,qh}} - \lambda_{\text{eff,g-NN}}$. Convolution of a Gaussian density with a kernel function of width σ results in a density of width $\lambda_{\text{eff,qh}} + \sigma$. We therefore assume a functional relationship of $\Delta\lambda$ with the average kernel width $\bar{\sigma}$. To test this hypothesis in high-dimensional space, we generated multivariate Gaussian densities with varying widths. The corresponding entropy was computed analytically and compared to $S_{\text{g-NN}}$. We found that for the used set of test functions a good approximation to $\Delta\lambda$ is given by the linear relationship $\Delta\lambda = m \max(0, \bar{\sigma} - \sigma_{\text{qm}}) + b$, where

$$m = -1.015 \times 10^{-3} d + 0.079$$

$$b = -5.4 \times 10^{-8} d + 8.5 \times 10^{-7}.$$

Thus, a corrected entropy is given by

$$S_{\text{corr}} = \frac{dk_B}{2}\ln\left[\frac{e^2}{\hbar^2\beta}(\lambda_{\text{eff,g-NN}} + \Delta\lambda)\right].$$

Because of all functions with given variance, the Gauss function has the largest entropy, S_{corr} is guaranteed to provide an upper bound for the true entropy.

III. METHODS

A. Simulation setup

The test systems that were compared with a thermodynamic integration reference (butane to decane and dialanine, see below) were set up as follows. Force field parameterizations were obtained from the Dundee Prodrug server [25] based on the GROMOS united-atom force field [26]. Stochastic dynamics simulations were performed using the molecular simulations package GROMACS [27] *in vacuo* at 400 K (alkanes and dialanine) and respectively 300 K (dialanine only) with friction constant γ set to 10, dielectric constant $\varepsilon = 1$, integration step size of 0.0005 ps and no bond constraints. Positional restraints were applied to three adjacent terminal heavy atoms.

The coldshock protein (protein database entry 1CSP) was simulated using the OPLS all atom force field [23] in explicit TIP4P solvent [28] and periodic boundary conditions. NpT ensembles were simulated, with the protein and solvent coupled separately to a 300 K heat bath ($\tau = 0.1$ ps) [29]. The systems were isotropically coupled to a pressure bath at 1 bar ($\tau = 1.0$ ps) [29]. Application of the Lincs [30] and Settle [31] algorithms allowed for an integration time step of 2 fs. Short-range electrostatics and Lennard-Jones interactions were calculated within a cutoff of 1.0 nm, and the neighbor list was updated every 10 steps. The particle mesh Ewald (PME) method was used for the long-range electrostatic interactions [32], with a grid spacing of 0.12 nm.

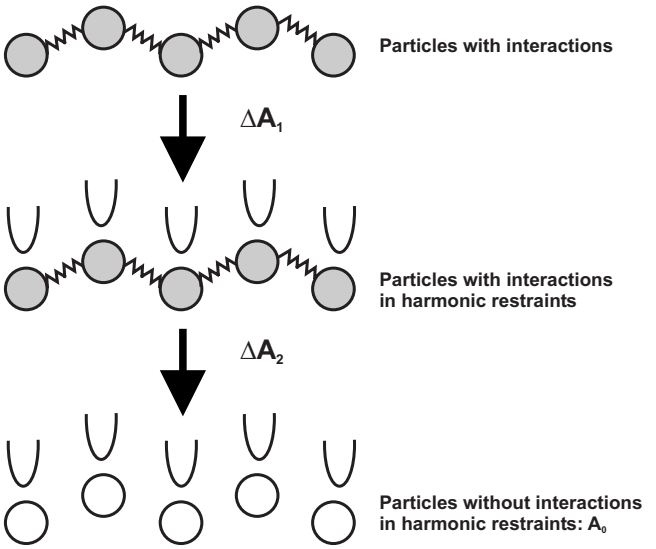

FIG. 3. Thermodynamic integration scheme used for obtaining reference entropies. Grey circles depict interacting particles, white circles noninteracting particles, zigzags represent chemical bonds, and local harmonic potentials are sketched as parabolas. Partial charges were removed in a separate step not illustrated here. A similar scheme has been used by Tyka *et al.* [33].

B. Reference entropies by Thermodynamic Integration

Absolute free energies for the test systems butane to decane and dialanin were calculated by thermodynamic integration (TI). No reference values were obtained for the CSP, because the system is too large for the TI to converge. Simulation parameters cf. above. The TI scheme we have chosen to obtain the Helmholtz free energy A of the fully interacting particles consists of two phases (also shown in Fig. 3). Harmonic position restraints with a force constant $k = 25000$ kJ/(mol nm^2) were slowly switched on for each atom in the first phase, and in the second phase all force field components were gradually switched off. The system then consisted of noninteracting dummy particles with mass m oscillating in their respective harmonic position restraint potentials, i.e.,

$$V = \frac{1}{2}k\sum_{j=1}^{N}(\mathbf{x} - \mathbf{x}_j)^2.$$

The free energy of this harmonic system can be obtained analytically,

$$A_0 = -\beta^{-1}\frac{3}{2}\sum_{j=1}^{N}\left[\ln\left(\frac{1}{\hbar^2\beta^2 k_j}\right)\right]$$

where $k_j = \tilde{k}_j / m_j$ denotes the mass-weighted force constant. Hence, the thermodynamic integration yields the absolute free energy

$$A = A_0 - \Delta A_2 - \Delta A_1$$

and the entropy by $S = (A - \langle V \rangle)/T$, where $\langle V \rangle$ denotes the ensemble average of the potential energy.

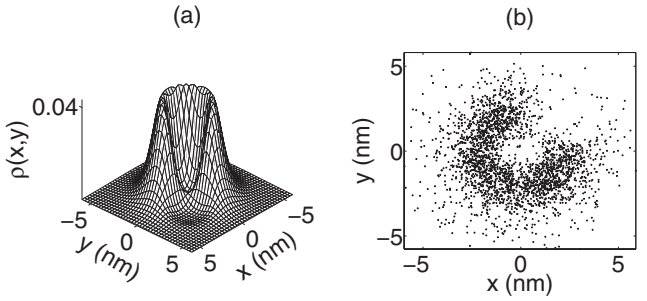

FIG. 4. (a) Synthetic test two-dimensional density whose entropy, 97.8 J K^{-1} mol^{-1}, was computed numerically by integrating on a grid. (b) 5000 points drawn according to this density function were used to estimate an entropy of 97.2 J K^{-1} mol^{-1}.

For the TI between the systems given by V_s (start) and V_f (end), 18 intermediate steps $V_i(\lambda)=\lambda V_s+(1-\lambda)V_f$, $i=1,\ldots,18$ were used, and the intermediate values of λ_i=0, 1e-6, 5e-6, 1e-5, 5e-4, 1e-4, 1e-3, 1e-2, 2e-2, 3e-2, 5e-2, 7e-2, 9e-2, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1 were distributed unevenly to obtain approximately balanced ΔA_i values. For each value of λ a trajectory of 12.5 ns (alkanes) or 100 ns (dialanine), respectively, was generated.

C. Efficient implementation

To compute the entropy via g-NN (Eq. (11)) as described is computationally more expensive than for k-NN (Eq. (7)), because to solve Eq. (9) until convergence requires a large number of iterations. Note, however, that the density estimate Eq. (10) is for all practical purposes invariant for small changes of k, because both k and $\sigma_{i,k}$ appear in Eq. (10) and scale similarly. Accordingly, convergence of k to 10% was considered sufficient to achieve accurate results, thus drastically reducing computational cost to a level similar to k-NN.

The *kd*-tree implementation of the nearest-neighbor library ANN [34] has been used for fast look-up of the k neighboring sample points.

IV. RESULTS AND DISCUSSION

A. Example: Simple density distributions

In a first step, our nonparametric entropy estimation was tested with synthetic probability density distributions whose entropy is accessible either analytically (e.g., Gaussians) or through grid-based numerical approximation (e.g., densities in two- and three-dimensional space). Figure 4 shows one of several test densities which were designed to exhibit typical features, e.g., a curved "thread," also seen in the configurational densities of macromolecules. All reference entropies were reproduced by the estimator to within 1 J K^{-1} mol^{-1} or better (results not shown). A similar test density was studied in three dimensions, as well as 42-dimensional checkerboardlike densities.

In Sec. II C 3, we derived an empirical relationship of the entropy overestimation due to the smoothing of the soft degrees and the average width of the estimation kernel. Here we show that this can exploited to correct for excess smoothing in more challenging high-dimensional cases, where this effect is aggravated due to increased surface effects. We considered Gaussian densities ranging from 12 to 42 dimensions. Figure 5(a) shows analytically computed reference entropies S_{ana} versus values obtained from our density estimator, both with (gray circles) and without (black crosses) empirical smoothing correction. Figure 5(b) shows the same data as a function of dimensionality. As can be seen, the correction improves the obtained entropy values considerably and provides accurate values independent of dimensionality. In contrast, the uncorrected values systematically overestimate the entropy with increasing dimension. Note, that this correction is different from the QM correction described in Sec. II C 2. The empirical smoothing correction accounts for an excess of smoothing of the soft degrees of freedom, whereas the QM correction enforces a minimum of smoothing for the otherwise unphysical sharp stiff degrees of freedom.

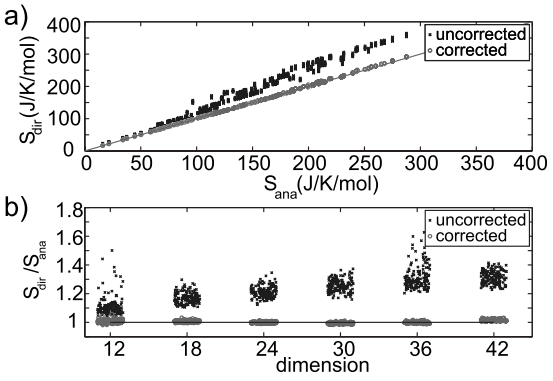

FIG. 5. Effect of the empirical smoothing correction described in the text. (a) Systematic overestimation seen as deviations from the diagonal. Black crosses: uncorrected values; gray circles: corrected values. (b) The same data as a function of dimensionality.

B. Example: Alkanes

We next applied our estimation method to a number of more realistic molecular test systems. Here, the accuracy of the estimate may be affected by two sampling effects; first, insufficient simulation sampling due to unvisited configurational space regions, second, locally too sparse sampling, which affects the accuracy of the NN density estimates. These two sampling effects are largely independent; whereas the sparse sampling problem, also called the "curse of dimensionality" [14], aggravates with increasing dimensionality and is largely independent of the particular system studied, the simulation sampling problem will depend on the size of the accessible configurational space as well as the slowest relaxation times of that system. Consider, for example, a single harmonic well. Whereas there is no simulation sampling problem for high-dimensional wells, NN-density estimations will inevitably suffer from sparse sampling for high-dimensional wells—a problem which, due to the regularization assumptions, is much less pronounced for the quasiharmonic approximation.

As test systems we chose (i) n-alkanes ranging from butane to decane, as a model of protein sidechain behavior, and

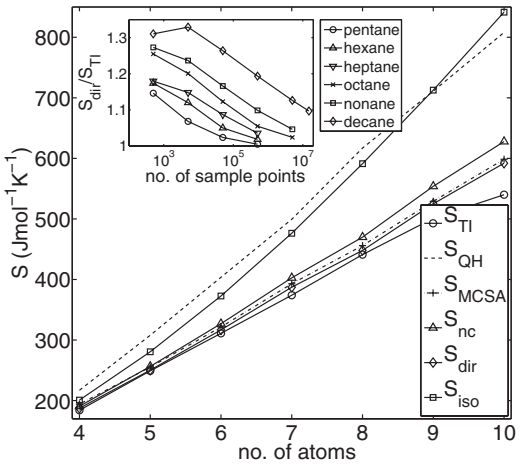

FIG. 6. Entropies for the *n*-alcanes from butane (N=4) to decane (N=10). S_{dir}: direct adaptive kernel density estimation without any subspace clustering; S_{nc}: such as S_{dir} but without empirical smoothing correction (Sec. II C 3); S_{iso}: direct isotropic kernel density estimation without any subspace clustering; S_{MCSA}: sum of density estimates after subspace clustering; *clust*: size of largest cluster; S_{QH}: entropy estimate according to quasiharmonic approximation. The errors of respectively S_{dir}, S_{nc}, S_{iso}, and S_{MCSA} were below 0.23 J mol^{-1} K^{-1} in all cases and not shown in the figure, since they are smaller than the symbols. The inset shows the same data as the relative derivation to the reference TI value, indicating convergence behavior.

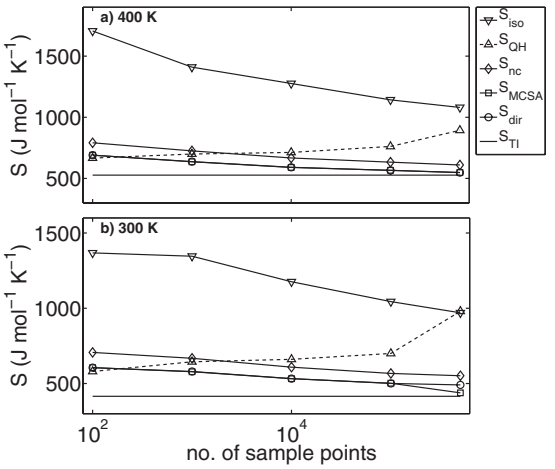

FIG. 7. Calculated dialanine entropies vs number of used sample points. (a) Simulation at 400 K and, respectively, (b) at 300 K The entropy labels shown are those defined in the caption of Fig. 6.

(ii) dialanine, as a minimal model for a protein backbone. The flexible wormlike shape of alkanes is likely to aggravate the two sampling problems discussed above sufficiently to explore the limits of our density estimator.

To obtain reference entropy values for these systems, a thermodynamic integration scheme was used which gradually perturbed the systems toward an analytically tractable state consisting of noninteracting particles in harmonic wells, as described in methods, Sec. III B. In each case, convergence was reached (data not shown).

The seven *n*-alkanes ranging from butane to decane were described by a united-atom force field, such that the systems comprised four to ten atoms, respectively. Figure 6 shows the entropies obtained with density estimation for these systems [35] as well as the reference obtained by thermodynamic integration and the QH entropy for comparison. Each of the listed density estimates has been obtained by averaging five independent trajectories. As can be seen, the QH entropy (dashed line) overestimates the configurational entropy by a wide margin for all systems (15–50%). The adaptive kernel density estimation, in contrast, yielded considerably improved results. For butane to octane, a deviation below 3% deviation from reference was achieved. For the largest alkanes considered, nonane and decane, a slightly reduced accuracy of 5 and 10%, respectively, was obtained.

It is instructive to consider the density estimation results for the conventional isotropic kernels, S_{iso}, also listed in the supplementary material [35]. Here, apparently due to the curse of dimensionality [14], drastic overestimates are seen for all alkanes. For butane (12 degrees of freedom) the entropy is overestimated by 9%, and for decane performs even worse than the QH approximation.

This result shows that it is the adaptive anisotropic kernels (ellipsoids rather than spheres, whose principal axes are determined from the local density by a principal component analysis, cf. methods), which provide the observed accuracy. Anisotropic kernels according to Eq. (14) rather than according to Eq. (13) turned out to yield more accurate results (data for the latter not shown).

As can also be seen in the figure, the empirical smoothing correction, which accounts for the inevitable blurring during the averaging process of the density estimation (Sec. II C 3), markedly improved the density estimates.

The inset of Fig. 6 highlights the convergence behavior of the adaptive anisotropic kernel estimates for alkanes with increasing simulation length. Clear convergence is seen for all alkanes up to octane, i.e., up to 24-dimensional densities. Whereas for these alkanes, 500.000 sample points sufficed for an accurate density estimate, 5.000.000 were required for nonane, and even more for decane, where sufficient sampling was not reached. In all cases, four sample points per picosecond simulation time were recorded. Although it is unclear at this point whether this lack of convergence is due to insufficient sampling of the simulation or the curse of dimensionality due to sparse sampling described above, the high flexibility of decane suggests the former. In the next section, we will address this question by considering a less flexible, but larger system.

C. Example: Dialanine

Dialanine, as a minimal model of a protein backbone comprising 45 atoms, served as our largest test system. The molecule was simulated with implicit solvent at temperatures 300 and 400 K using a united-atom force field. Figure 7 shows entropies S_{dir} and S_{QH} obtained for simulation lengths ranging from 100 ps to 500 ns. The TI reference entropy is shown as a horizontal line. Similarly as seen above for the alkanes, the QH entropy S_{QH} overestimates the TI reference. Interestingly, and in contrast to the other estimates, the QH-

estimate error increases with the length of the simulation. As a consequence, the longest, 500 ns, trajectory yields the most inaccurate result with an error of 250%. We attribute this peculiar behavior to a decreasing quality of the harmonic approximation with increasing complexity of the sampled configurational space. In contrast, the entropy obtained with density estimation S_{dir} converges toward the correct value with increasing simulation length. For the 500 ns simulation, the TI reference is reached to within 4.0 and 18% for 400 and 300 K, respectively.

Again, density estimates from isotropic kernels yielded even lower accuracy than the QH estimate in all cases. Further, compared to the alkanes, the smoothing correction improved density estimates to an even larger extent; the uncorrected estimates overestimated the reference by 16 and 33% for 400 and 300 K, respectively.

This result corroborates the previous observation that the achieved accuracy was mainly due to the adaptive anisotropic kernels. Moreover, even though the dimension of the configurational space of dialanine is almost twice as large as that of nonane, the obtained accuracy is similar despite the fact that ten times less sample points were used for dialanine. Apparently, the curse of dimensionality is not yet severe here, thus underscoring the robustness of the adaptive anisotropic kernels even for high dimensionality. A second conclusion is that, as expected, lack of simulation sampling limited the accuracy of the decan entropy estimates. This interpretation is further supported by the fact that more accurate estimates are obtained for the high-temperature simulations, which are likely to provide enhanced simulation sampling, but sparser local sampling. In this sense, alkanes proved to be particular hard test systems.

D. Application: Coldshock protein

The previous sections established that adaptive anisotropic kernel density estimation is very well feasible in configurational space of up to 45 dimensions. Although this is still far from the high dimensionality typically found for biological macromolecules, it holds the promise to provide the missing building block for a reliable application of the minimally coupled subspace framework (MCSA) derived in detail in Ref. [15]. Briefly, full correlation analysis [20] (FCA) is used to obtain nearly independent orthogonal configurational subspaces, which allow to approximately factorize the configurational space density. Subsequently, adaptive anisotropic kernels are used to estimate the entropy contributions for each of these subspaces.

To test the feasibility and accuracy of this approach, we apply this framework to the coldshock protein, a 67-residue soluble protein (protein database entry 1CSP) which was simulated in explicit solvent, as described in methods. Of particular interest was the question of how many (soft) degrees of freedom should be subjected to FCA and to the adaptive anisotropic kernel density estimation, leaving the remaining (stiff) degrees of freedom to the Schlitter QH approximation. For simplicity and computational efficiency, only backbone contributions to the configurational entropy were considered. FCA was carried out on the first

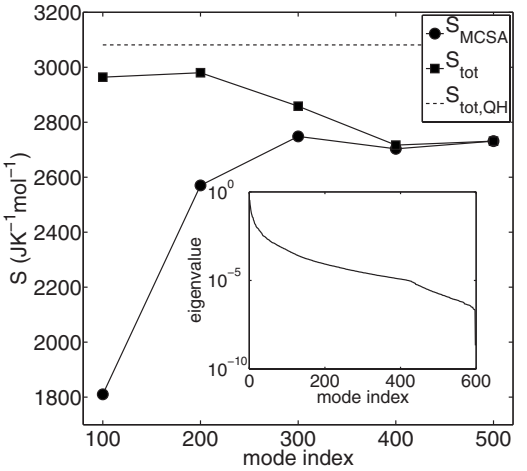

FIG. 8. Estimated entropies for the coldshock protein (1CSP, backbone only). The total estimate, S_{tot}, as a function of MCSA subspace size (100,...,500 modes), was calculated as sum of S_{MCSA}, the estimate obtained by application of adaptive anisotropic kernels to minimally coupled subspaces obtained by FCA, and the QH estimate for the remaining modes, as described in the text; S_{PCA}: cumulative QH estimate; $S_{\mathrm{tot,QH}}$: global QH reference.

100,200,...,500 modes of the configurational ensemble, and nearly noncorrelated subspaces were defined as described in Ref. [15]. Our adaptive anisotropic kernel estimate was subsequently applied to each of the resulting minimally coupled subspaces yielding an improved entropy estimate S_{MCSA}. The remaining modes were considered via the Schlitter QH approximation, Eq. (6). The sum of these two yields the desired entropy estimate S_{tot}, also shown in Fig. 8. Due to the large size of this protein no converged TI entropy reference could be obtained; hence, we here resort to comparison with the QH estimate of the total configurational space density $S_{\mathrm{tot,QH}}$.

Figure 8 shows that the combination of FCA and adaptive anisotropic kernels improves on the QH estimate in all cases. In particular, a marked improvement on the QH estimate is seen for the adaptive anisotropic kernel estimate, which shows that the neglect of nonlinear and higher-order correlations, and of anharmonicities by the established QH method implies an overestimation of the entropy by at least 14% (QH: 3081 J K^{-1} mol^{-1} vs S_{FCA}: 2710 J K^{-1} mol^{-1}). It turns out that for the coldshock protein 400 modes need to be subjected to the minimally coupled subspace approach for converged results. Obviously only the 200 stiffest modest are already sufficiently decoupled by PCA and can be well approximated by a quasiharmonic density.

This empirical distinction between stiff and soft modes is nicely reflected in the eigenvalue spectrum shown in the inset of Fig. 8. Whereas eigenvalues of modes 1 to 400 decrease smoothly, an abrupt drop is seen at mode 400. Similar observations for calmodulin [15] suggest that such abrupt drop can be used to identify the stiff modes that can be subjected to the QH approximation. Our results also suggest that, generally, a large fraction of degrees of freedom needs to be subjected to MCSA to obtain reliable estimates.

V. CONCLUSIONS

We have introduced a kernel-based density estimation method and showed that it provides accurate results in even the high-dimensional and quite complex configurational space density generated by the dynamics of biological macromolecules. Whereas established *k*-nearest-neighbor estimators have been reported to fail to converge for 23-dimensional configurational space with 15 million sample points, the adaptive anisotropic kernels developed here provide robust results in up to 45 dimensions with only 500.000 sample points. We attribute this improvement to the occurrence of typically sparsely sampled configurational density "threads" ubiquitous in protein dynamics, to which, apparently, our locally multivariate approach provides enhanced approximation.

Used within the framework of the minimally coupled subspace approach [15], this density estimator serves to overcome the three limitations inherent in the conventional quasiharmonic approximation, i.e., neglect of nonlinear and higher-order correlations, of anharmonicity, and of multimodality. So far we have applied the method to the proteins insulin (51 residues) and calmodulin (146 residues). In both cases the method yielded significantly lower entropy than QH and was able to reconcile experimental data previously found to be inconsistent with the QH entropy. Although encouraging, further studies will have to corroborate these results, and investigate possible limitations in more detail.

ACKNOWLEDGMENTS

U.H. was supported by the Deutsche Forschungsgemeinschaft (research training group 782). O.F.L. was supported by the Human Frontiers of Science Foundation

[1] S. Cheluvaraja and H. Meirovitch, Proc. Natl. Acad. Sci. U.S.A. **101**, 9241 (2004).
[2] S. Cheluvaraja and H. Meirovitch, J. Chem. Phys. **125**, 024905 (2006).
[3] T. P. Straatsma and J. A. McCammon, Annu. Rev. Phys. Chem. **43**, 407 (1992).
[4] C. Peter, C. Oostenbrink, A. van Dorp, and W. F. van Gunsteren, J. Chem. Phys. **120**, 2652 (2004).
[5] D. L. Beveridge and F. M. DiCapua, Annu. Rev. Biophys. Biophys. Chem. **18**, 431 (1989).
[6] P. Kollman, Chem. Rev. **93**, 2395 (1993).
[7] H. Schäfer, A. E. Mark, and W. F. van Gunsteren, J. Chem. Phys. **113**, 7809 (2000).
[8] H. Meirovitch, Curr. Opin. Struct. Biol. **17**, 181 (2007).
[9] J. N. Karplus and Martin Kushick, Macromolecules **14**, 325 (1981).
[10] J. Schlitter, Chem. Phys. Lett. **215**, 617 (1993).
[11] C. Chang, W. Chen, and M. Gilson, J. Chem. Theory Comput. **1**, 1017 (2005).
[12] R. Baron, W. van Gunsteren, and P. Hünenberger, Trends Chem. Phys. **11**, 87 (2006).
[13] R. Baron, A. deVries, P. Hünenberger, and W. van Gunsteren, J. Phys. Chem. B **110**, 8464 (2006).
[14] R. E. Bellman, *Adaptive Control Processes* (Princeton University Press, Princeton, NJ, 1961).
[15] U. Hensen, H. Grubmüller, and O. F. Lange (unpublished).
[16] H. Matsuda, Phys. Rev. E **62**, 3096 (2000).
[17] A. Baranyai and D. J. Evans, Phys. Rev. A **40**, 3817 (1989).
[18] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions* (Reidel, Dordrecht, Boston, 1974), Vol. 1.
[19] V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh, J. Comput. Chem. **28**, 655 (2007).
[20] O. F. Lange and H. Grubmüller, Proteins **70**, 1294 (2008).
[21] J. Haas, E. Vöhringer-Martinez, A. Bögehold, D. Matthes, U. Hensen, A. Pelah, B. Abel, and H. Grubmüller, ChemBioChem **10**, 1816 (2009).
[22] N. Go and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).
[23] G. Kaminski, R. Friesner, J. Tirado-Rives, and W. Jorgensen, J. Phys. Chem. B **105**, 6474 (2001).
[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley-Interscience, New York, 2000).
[25] A. W. Schüttelkopf and D. M. F. van Aalten, Acta Crystallogr., Sect. D: Biol. Crystallogr. **60**, 1355 (2004).
[26] W. F. van Gunsteren, X. Daura, and A. E. Mark, *GROMOS Force Field* (1998), pp. 1211–1216.
[27] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[28] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[29] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[30] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[31] S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).
[32] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[33] M. Tyka, A. Clarke, and R. Sessions, J. Phys. Chem. B **110**, 17212 (2006).
[34] URL http://www.cs.umd.edu/~mount/ANN/.
[35] See EPAPS Document No. E-PLEEE8-80-094907 for detailed test results. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.