Received: 2 June 2023 | Revised: 27 August 2023 | Accepted: 13 September 2023

DOI: 10.1002/wcms.1691

SYSTEMATIC REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

The *k*th nearest neighbor method for estimation of entropy changes from molecular ensembles

Federico Fogolari[1,2] | **Roberto Borelli**[1] | **Agostino Dovier**[1] | **Gennaro Esposito**[2,3]

[1]Dipartimento di Scienze Matematiche, Informatiche e Fisiche (DMIF), University of Udine, Udine, Italy

[2]Istituto Nazionale Biostrutture e Biosistemi, Roma, Italy

[3]Science and Math Division, New York University at Abu Dhabi, Abu Dhabi, United Arab Emirates

Correspondence
Federico Fogolari, Dipartimento di Scienze Matematiche, Informatiche e Fisiche (DMIF), University of Udine, Via delle Scienze 206, 33100 Udine, Italy.
Email: federico.fogolari@uniud.it

Funding information
European Commission, Grant/Award Number: G25F21003390007

Edited by: Peter R. Schreiner, Editor-in-Chief

Abstract

All processes involving molecular systems entail a balance between associated enthalpic and entropic changes. Molecular dynamics simulations of the end-points of a process provide in a straightforward way the enthalpy as an ensemble average. Obtaining absolute entropies is still an open problem and most commonly pathway methods are used to obtain free energy changes and thereafter entropy changes. The *k*th nearest neighbor (kNN) method has been first proposed as a general method for entropy estimation in the mathematical community 20 years ago. Later, it has been applied to compute conformational, positional–orientational, and hydration entropies of molecules. Programs to compute entropies from molecular ensembles, for example, from molecular dynamics (MD) trajectories, based on the kNN method, are currently available. The kNN method has distinct advantages over traditional methods, namely that it is possible to address high-dimensional spaces, impossible to treat without loss of resolution or drastic approximations with, for example, histogram-based methods. Application of the method requires understanding the features of: the *k*th nearest neighbor method for entropy estimation; the variables relevant to biomolecular and in general molecular processes; the metrics associated with such variables; the practical implementation of the method, including requirements and limitations intrinsic to the method; and the applications for conformational, position/orientation and solvation entropy. Coupling the method with general approximations for the multivariable entropy based on mutual information, it is possible to address high dimensional problems like those involving the conformation of proteins, nucleic acids, binding of molecules and hydration.

This article is categorized under:

Molecular and Statistical Mechanics > Free Energy Methods
Theoretical and Physical Chemistry > Statistical Mechanics
Structure and Mechanism > Computational Biochemistry and Biophysics

KEYWORDS

conformational entropy, entropy, *k*th nearest neighbor, translation/rotation entropy

1 | INTRODUCTION

Many important processes involving biomolecules, like protein folding, conformational transitions, or non-covalent ligand binding, display fine thermodynamic properties that involve a balance between different driving forces. Such a balance is functionally important for processes to proceed in one or the reverse direction, depending on changes in environmental variables (like temperature and pH) and/or in molecular concentrations,[1] or upon chemical changes in the biomolecules, like amino acid mutations in proteins.[2]

Molecular simulations have been playing a crucial role in our current ability to predict and rationalize the thermodynamics of biomolecular systems and characterize their free energy changes. Enthalpy is easily obtained from simulations as the average of the forcefield energy. Although the latter is subject to large fluctuations, equilibrated simulations can be performed long enough as to reduce the standard error of the mean. Entropy estimation is more difficult because it requires knowledge of the probability distribution function of the system, dependent on many correlated variables. Many reviews focusing on different aspects of free energy and entropy estimation from molecular simulations are available.[3–15]

Wereczinski and McCammon[10] identify two major techniques for computing free energies from simulation: thermodynamic pathway and end-point methods. In the former class the difference in free energy between two states is specifically addressed, and sampling along a pathway between them is used to compute the free energy. For instance, in alchemical transformations[16] it is the system Hamiltonian which is changed from H_0 (initial state) to H_1 (end state). Free Energy Perturbation[17] and Thermodynamic Integration[18] implement this approach. In another scenario the initial and end states are different states of the same system, for example, the folded and unfolded states of a protein, or the free and bound states of two molecules. In this case, a biasing potential is added to the potential, to sample the path between the two end states, along one or more "collective variables," choosing specific points along the path, like in Umbrella Sampling,[19] or adaptively modifying the potential to let the system diffuse freely in collective variable space, like in Metadynamics,[20] or in the Adaptive Biasing Force method,[21] or enhancing sampling with no specific choice of collective variables like in accelerated Molecular Dynamics.[22] Reweighting is necessary to recover free energy differences for the unbiased ensembles. In this respect schemes have been devised, like the Weighted Histogram Analysis Method[23] and the Multistate Bennett Acceptance Ratio Method,[24] to optimally use all the samples. Non-equilibrium simulations can also provide equilibrium free energy changes through use of Jarzynski equality.[25,26]

Calculating free energies (sometimes called "absolute" free energies) from end-point simulations with respect to a given reference state, involves on the contrary equilibrium simulations, in which the Hamiltonian is not changed, and the analysis of the thermodynamic ensemble of conformations, or other quantities from the simulations.

The application of end-point methods poses several problems, including the difficulties in sampling and treatment of a very large dimensional space where many dimensions are strongly correlated.

Suarez and Diaz[11] have thoroughly reviewed the subject focusing on single-molecule (conformational) entropy and reviewing the main methods that have been developed and are currently used.

Early models have considered fairly rigid molecular arrangements and applied the rigid rotor harmonic oscillator (RRHO) model where the six (external) degrees of freedom for translation and rotation are decoupled from internal degrees of freedom which are treated by harmonic or quasi-harmonic analysis.[27–32]

Such methods were further refined and exact methods were formulated to take into account anharmonicity and supralinear correlations[33,34] or to improve the accuracy with respect to quasi-harmonic analysis.[35]

Other approaches have been formulated considering multiple energy wells in the multidimensional energy surface and summing the average entropy associated with each well and the entropy due to the well population probability.[8,36] The approach is adopted in the Mining Minima (M2) approach by Gilson et al.[37,38] and forms the basis of many other different approaches (see, e.g., References [39, 40]).

The main difficulty common to this and other methods is the partitioning of the large dimensional space in low-dimensional subspaces and how to treat correlations among the variables in different subspaces.

In the last two decades methods to estimate the entropy from end-point simulations which are based on the *k*th nearest neighbor method[41,42] have been increasingly used and implemented in publicly available software programs.

The original articles detailing the theory were not meant for an audience of physicists and chemists, thus we review here the methods highlighting the most interesting results in the context of entropy estimation from molecular simulations. Although the method may be applied in general to any kind of molecule, the discussion will be focused on molecules whose conformation is mainly described by correlated torsion angles and whose environment is mostly aqueous.

Here we do not review general methods for estimation of entropy from molecular dynamics simulations, which are covered by other excellent reviews,[8,11] but rather address the *k*th nearest neighbor method, its theory, and implementation in the context of molecular entropy estimation.

2 | METHODS

2.1 | A brief summary on entropy in classical statistical mechanics[43,44]

We consider a system consisting of N atoms, in a volume V_{3D} at temperature T. The 3-dimensional ordinary volume of the system has been indicated explicitly by the subscript 3D to avoid confusion with the volume in multidimensional coordinates space which will be introduced later. The state of the system is described by $3N$ Cartesian atomic coordinates and $3N$ momentum coordinates. Its energy is given by the sum of a potential energy function $U\left(\vec{x}_1,\vec{x}_2,...,\vec{x}_N\right)$ depending only on the positional coordinates $\left\{\vec{x}_1,\vec{x}_2,...,\vec{x}_N\right\}$ and the sum of atomic kinetic energies $K=\sum_{i=1,N}\frac{\left|\vec{p}_i\right|^2}{2m_i}$, with $\vec{p}_i$ the momentum of atom i.

The classical partition function for this system is:

$$Q=\frac{\int e^{-\frac{U+K}{k_BT}}d\vec{x}_1d\vec{x}_2...d\vec{x}_Nd\vec{p}_1d\vec{p}_2...d\vec{p}_N}{h^{3N}}, \tag{1}$$

where k_B and h are the Boltzmann's and Planck's constants, respectively. Introduction of h^{3N} at the denominator, which is clearly coming from quantum statistical mechanics, serves two purposes: making Q adimensional and making it consistent with the partition function derived in quantum statistical mechanics for simple systems.[43] As it is apparent below, the choice of normalization factor, here h^{3N} but often implicitly $1\times(\text{unit of momentum}\times\text{unit of length})^{3N}$, sets the partition function of the zero free energy reference state and it is inconsequent on free energy differences. For molecular processes in condensed phase, for which changes of volume are typically rather limited, we can consider the Helmholtz free energy A as an appropriate thermodynamic potential:

$$A=-k_BT\log\left(\frac{1}{h^{3N}}\int e^{-\frac{U+K}{k_BT}}d\vec{x}_1\vec{x}_2...d\vec{x}_Nd\vec{p}_1d\vec{p}_2...d\vec{p}_N\right). \tag{2}$$

The momentum integral can be calculated and factored out leaving us with the equation:

$$A=-k_BT\log\left(\int e^{-\frac{U}{k_BT}}d\vec{x}_1d\vec{x}_2,...,d\vec{x}_N\right)-\frac{3}{2}k_BT\sum_{i=1,N}\log(2\pi m_ik_BT)+3Nk_BT\log(h) \tag{3}$$

The second right-hand term comes from the momentum integral and is typically safely neglected when comparing states of the same system (i.e., two different conformations, or free vs. bound state) at the same temperature, leading to its cancellation. The third term is related to the choice of normalization factor of the partition function. The first term is a configurational term and leads to the differences in free energy when we consider different macrostates of the same system at the same temperature. Different macrostates are defined by restricting the integral to different regions of the configurational space. For example, for a molecular complex the space could be partitioned in "free" and "bound" states based on one or more intermolecular distances. Entropy is obtained by the derivative:

$$S = -\left(\frac{\partial A}{\partial T}\right)_{N,V_{3D}}. \tag{4}$$

Explicit derivation leads to the following expression for entropy:

$$\begin{aligned} S &= -k_B \int \rho\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\right) \log\left(\rho\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\right)\right) d\vec{x}_1 d\vec{x}_2, ..., d\vec{x}_N \\ &\quad + \frac{3}{2}\left(N + \sum_{i=1,N} \log(2\pi m_i k_B T)\right) - 3Nk_B \log(h) \\ &= S_{\text{conf}} + S_{\text{mom}} + S_0 \end{aligned} \tag{5}$$

where S_0 is related to the choice of normalization factor for the partition function, S_{mom} is the entropy associated with the momentum integral and S_{conf} is the entropy associated with the configurational freedom of the molecule which is expressed in terms of the probability density:

$$\rho\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\right) = \frac{e^{-\frac{U\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\right)}{k_B T}}}{\int e^{-\frac{U\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\right)}{k_B T}} d\vec{x}_1 d\vec{x}_2, ..., d\vec{x}_n}. \tag{6}$$

Note that after partitioning the entropy in three terms, the argument of the logarithm in all terms, is a dimensional quantity. When each term is considered in isolation the choice of the units of measure implies the choice of a reference state for each term. This subject and the change of entropy upon change of variables are thoroughly discussed by Hnizdo and Gilson.[44]

From the above Equation (5) it is seen that S_{conf} is the average of $-k_B \log(\rho)$ over the thermodynamic ensemble, that is:

$$S_{\text{conf}} = -k_B \int \rho\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\right) \log\left(\rho\left(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\right)\right) d\vec{x}_1 d\vec{x}_2 ... d\vec{x}_n, \tag{7}$$

which is the thermodynamic ensemble average of $-k_B \log(\rho)$, that is:

$$S_{\text{conf}} = -k_B \langle \log(\rho) \rangle. \tag{8}$$

The fundamental Equation (7) was introduced by Boltzmann in terms of the number of what he called complexions, instead of probability densities (see the introduction to the translation of the original article by Sharp and Matschinsky[45]). The present form involving the probability density function was later introduced by Gibbs.[46] The definition of entropy in terms of probabilities later laid the foundation of information theory.[47] The term "entropy" in that context was suggested by von Neumann who had defined in a similar way the entropy for quantum mechanical systems.[48]

Equation (8) is explicitly used in the *k*th nearest neighbor method.

It is worth noting that all thermodynamic quantities expressed as ensemble averages may be estimated from the configurations sampled by equilibrium molecular dynamics simulations according to the probability density of Equation (6).

2.2 | The *k*th nearest neighbor method for entropy estimation

Our task is to estimate the entropy of a probability density function (pdf) $\rho\left(\vec{x}\right)$ from a set of n independent identically distributed samples $\left\{\vec{x}_i\right\}$ in a general d-dimensional space. A reasonable guess of the probability density at point $\vec{x}_i$

may be obtained by taking a d-dimensional volume V centered at $\vec{x}_i$ and counting how many of the n samples ($n_V\left(\vec{x}_i\right)$) are found inside the volume. Assuming a constant pdf within the volume, the probability of any sample to be found inside V centered at $\vec{x}_i$ would be estimated by

$$\widehat{p}_{V\left(\vec{x}_i\right)} = \frac{n_V\left(\vec{x}_i\right)}{n}, \tag{9}$$

which is a discrete estimator (dependent on point $\vec{x}_i$ and chosen volume V) of a continuous probability function of space and chosen volume V.

The probability density, assumed to be uniform inside V, is consequently estimated by:

$$\widehat{\rho}\left(\vec{x}_i\right) = \frac{n_V\left(\vec{x}_i\right)}{nV}. \tag{10}$$

With an estimate of the probability density at each sample, the entropy can be thus naively estimated by:

$$\widehat{S}_{\text{naive}} = < -k_B \log\left(\widehat{\rho}\left(\vec{x}\right)\right) > = -\frac{k_B}{n}\sum_{i=1,n} \log\left(\frac{n_V\left(\vec{x}_i\right)}{nV}\right), \tag{11}$$

with k_B the Boltzmann's constant.

Since V is set equal for each sample $\vec{x}_i$ and the probability distribution function is not uniform, then the estimates of the probability density function at each sample $\vec{x}_i$ will result from a different number of samples. Moreover to have samples also in low probability regions, V would be chosen unnecessarily large for more dense regions, resulting in a loss of resolution, that is, in an overestimate of the entropy.

In the kth nearest neighbor method instead of the volume, the number of samples k inside each volume is chosen equal for all samples, and the volume (which is therefore dependent on the sample $\vec{x}_i$) $V\left(\vec{x}_i\right)$ is adapted as to include just k samples.

Although the volume could be chosen of any shape, a reasonable choice (in the absence of a detailed analysis on the distribution of samples) is to take a ball whose radius excludes the kth nearest neighbor sample of sample $\vec{x}_i$, in such a way that exactly k samples are found inside the volume. Therefore the volume of the ball centered at each sample depends on the sample i and on the value of k chosen (identical for all samples). This fact is made explicit in the following by the subscripts on V: $V_{i,k}$ (Figure 1). The problem with the naive approach is that in general the logarithm of an unbiased estimator of the pdf is not an unbiased estimator of the logarithm of the pdf.

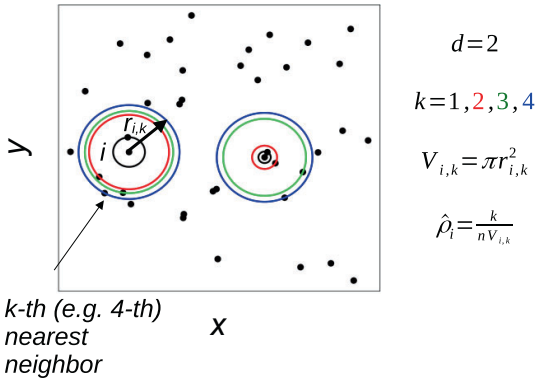

FIGURE 1 Illustration of the kth nearest neighbor method in two dimensions ($d=2$). In the example the sample i and its distance ($r_{i,k}$) to the kth (in the example, the fourth) nearest neighbor is highlighted. The example involves a two-dimensional Euclidean space where the volume of a ball with radius $r_{i,k}$ is $V_{i,k}=\pi r_{i,k}^2$, that is, the area of the disk with radius $r_{i,k}$. $\widehat{\rho}_i$ is the estimated probability density at sample i. The balls up to the first (black), second (red), third (green), and fourth (blue) neighbors are shown for sample i and for another sample.

Kozachenko and Leonenko[41] in their seminal paper (unfortunately available only in Russian) elaborated the naive entropy estimation provided by Equation (11) and showed that, for $k=1$, save for the k_B factor added in the present context, the estimator:

$$\widehat{S}_{KL} = -\frac{k_B}{n}\sum_{i=1,n}\log\left(\frac{1}{(n-1)V_{i,1}}\right)+\gamma, \tag{12}$$

where γ is the Euler-Mascheroni constant (0.5722 ...), is an unbiased estimator of the entropy as $n\to\infty$. They further studied the mean square error of the estimate, without reporting the complete cumbersome demonstration.

Demchuk et al.[42] developed and extended the idea originally proposed by Kozachenko and Leonenko,[41] and worked out the theory providing useful formulae and similarly, Goria et al. developed essentially the same theory[49] and Berrett et al. considered a weighted version of the approach.[50] Starting from the work of Kozachenko and Leonenko, Grassberger et al.[51] proposed an implementation of the kNN method, focusing on mutual information. Dealing with different (single and product) spaces and dimensions the choice of a fixed value of k for all spaces corresponds to largely different resolutions. Their algorithms address specifically this problem, using the maximum norm instead of Euclidean norm and choosing different k for each one of the single and product spaces.

We summarize hereafter the theory following closely Demchuk et al.,[42] simplifying and generalizing to non-Euclidean metrics. As above, we start considering the probability density estimator

$$\widehat{\rho}\left(\vec{x}_i\right) = \frac{k}{nV_{i,k}}. \tag{13}$$

We further consider the entropy estimator according to the naive reasoning above and based on nearest neighbors:

$$\widehat{S}_{\text{naive,nn}} = <-k_B\log\left(\widehat{\rho}\left(\vec{x}\right)\right)> = -\frac{k_B}{n}\sum_{i=1,n}\log\left(\frac{k}{nV_{i,k}}\right), \tag{14}$$

and we work out the probability density distribution of the quantity:

$$T_i = \log\left(\frac{nV_{i,k}}{k}\right), \tag{15}$$

The volume $V_{i,k}$ is the volume of a ball of radius $r_{i,k}$, where $r_{i,k}$ is the distance of sample i to its kth nearest neighbor.

Let us also distinguish variables $\vec{X}_1,\vec{X}_2,...$ representing the vector of variables of samples 1, 2, ..., respectively, from their actual samples $\vec{x}_1,\vec{x}_2,...$ and similarly $R_{i,k}$ from its actual sample $r_{i,k}$.

The probability that $T_i > t$

$$\Pr\left(T_i > t|\vec{X}_i = \vec{x}_i\right), \tag{16}$$

is equal to the probability:

$$\Pr\left(R_{i,k} > r|\vec{X}_i = \vec{x}_i\right), \tag{17}$$

where, based on the definition of T_i, r is the radius of a ball of volume:

$$V_t = \left(\frac{k}{n}e^t\right). \tag{18}$$

Note that all T_i are identically distributed, because all samples are drawn from an identical distribution, so that the probability that $\vec{X}_i = \vec{x}$ is exactly the same as for any other sample, say j, that $\vec{X}_j = \vec{x}$. Let us consider therefore one specific sample, say the first, replace $\vec{x}_1$ with the generic position $\vec{x}$ and consider T_1.

The probability (p) that sample j (with $j \neq 1$) is in the volume V_t centered at sample $\vec{X}_1 = \vec{x}$ is the integral of the probability density function over the ball of radius t centered at sample $\vec{X}_1 = \vec{x}$:

$$p = \Pr\left(\vec{X}_j \text{ in } V_t\left(\vec{x}\right)\right) = \int_{V_t\left(\vec{x}\right)} \rho\left(\vec{x}'\right) d\vec{x}'. \tag{19}$$

With this notation we can express the probability in Equation (16) as 1 minus the sum of the probabilities that only $0,1,2,...,k-1$ samples out of the total number of the samples other than 1, that is, $n-1$, occur inside the volume $V_t\left(\vec{x}\right)$, that is,

$$Pr\left(T_1 > t | \vec{X}_1 = \vec{x}\right) = 1 - \sum_{i=0,k-1} \binom{n-1}{i} p^i (1-p)^{n-1-i}. \tag{20}$$

When $n \to \infty$ the volume $V_t \to 0$ for any finite value of t and we can approximate

$$p = V_t\left(\vec{x}\right)\rho\left(\vec{x}\right). \tag{21}$$

In the same limit the binomial probability distribution tends to the Poisson distribution:

$$Pr\left(T_1 > t | \vec{X}_1 = \vec{x}\right) \to 1 - \sum_{i=0,k-1} \frac{\lambda^i e^{-\lambda}}{i!}, \tag{22}$$

with $\lambda = np$. The probability density function $h_{T_1,k,n}\left(\vec{x}\right)$ of the variable T_1 is obtained by taking the derivative with respect to t of Equation (22), applying the chain rule because λ depends on p which in turn depends on V_t which in turn depends on t. After some straightforward steps we obtain:

$$h_{T_1,k,n}\left(\vec{x}\right) = \frac{e^{-\lambda}\lambda^k}{(k-1)!}. \tag{23}$$

The expectation value of T is thus given by the integral:

$$E(T_1) = \int_{-\infty}^{\infty} t\, h_{T_1,k,n}\left(\vec{x}\right) dt = \int_{-\infty}^{\infty} t \frac{e^{-k\rho\left(\vec{x}\right)e^t}\left(k\rho\left(\vec{x}\right)e^t\right)^k}{(k-1)!} dt. \tag{24}$$

By a change of variable $z = k\rho\left(\vec{x}\right)e^t$ the integral may be written as:

$$E(T_1) = \int_0^{\infty} \left(\log(z) - \log(k) - \log\left(\rho\left(\vec{x}\right)\right)\right) \frac{e^{-z} z^{k-1}}{(k-1)!} dz, \tag{25}$$

$$= \left(\int_0^{\infty} \frac{\log(z) e^{-z} z^{k-1}}{(k-1)!}\right) - \log(k) - \log\left(\rho\left(\vec{x}\right)\right), \tag{26}$$

$$= L_{k-1} - \gamma - \log(k) - \log\left(\rho\left(\vec{x}\right)\right), \tag{27}$$

where L_{k-1} is defined iteratively by $L_0 = 0$, $L_i = L_{i-1} + \frac{1}{i}$, and γ is, as above, the Euler–Mascheroni constant. The last equality stems from recognition that the integral is the definition of the digamma function, that is, the derivative of the logarithm of the gamma function.

The above equations show that the variable

$$\log(V_{i,k}) + \log(n) - L_{k-1} + \gamma, \tag{28}$$

is an unbiased estimator of $-\log\left(\rho\left(\vec{x}\right)\right)$ and thus the average over the samples of this estimator is an unbiased estimator of the entropy of the distribution:

$$\widehat{S}_{n,k} = k_B \sum_{i=1,n} \frac{(\log(V_{i,k}) + \log(n) - L_{k-1} + \gamma)}{n}, \tag{29}$$

$$= k_B \sum_{i=1,n} \frac{\log(V_{i,k})}{n} + \log(n) - L_{k-1} + \gamma. \tag{30}$$

Note the similarity of Equation (30) with Equation (14), the only difference lies in the term $\log(k)$ in the naive estimator which is replaced in the exact estimator by $L_{k-1} - \gamma$. The difference becomes negligible for large values of k.

By similar, but definitely more complex treatment one can compute the variance of the unbiased estimator (Equation (30)) which is done in the article by Demchuk et al.[42] The variance of the sum in Equation (30) is the sum of different contributions entailing the variance of the pdf itself, the variance associated with kth nearest neighbor method and the covariances between pairs of estimates. Both Leonenko and Kozachenko[41] and Demchuk et al.[42] demonstrate that in the limit $n \to \infty$ the latter covariance term is zero, that is, samples are expected to be uncorrelated, and thus the variance of the unbiased entropy may be approximated by the sum of the variances of the T_i estimators, resulting in:

$$\mathrm{var}\left(\widehat{S}_{n,k}\right) \approx k_B^2 \frac{1}{n}\left(\sum_{j=k,\infty} \frac{1}{j^2} + \mathrm{var}\left(\log\left(\rho\left(\vec{X}\right)\right)\right)\right) = k_B^2 \frac{1}{n}\left(\left(\frac{\pi^2}{6} - \sum_{j=1,k-1} \frac{1}{j^2}\right) + \mathrm{var}\left(\log\left(\rho\left(\vec{X}\right)\right)\right)\right). \tag{31}$$

Demchuk et al. report all the details of the above approximation.[42] We remark that the approximations done involve three key aspects:

- n must be large to be able to use the Poisson distribution and to guarantee all limiting behaviors;
- in particular n must be such that, within the ball including $k-1$ nearest neighbors, the pdf must be constant. This means that all the pdf features at lengths shorter than the average radius of the balls built about all samples will be lost (and the entropy will be consequently overestimated). The average radius of the balls is the effective resolution of the kNN entropy estimator;
- in particular n must be large to guarantee that the covariance between any pair of single sample estimators is zero.

The above considerations may be confronted with Equation (31) which shows that increasing the number of neighbors inside the ball, that is, increasing k, has the effect of reducing the variance of the entropy estimate. For small k, depending on the number of samples n the variance part due to sampling may be large.

The practical application of the method must find a compromise between increasing the resolution (large n and small k) and reducing the variance of the entropy estimate (large k).

It is worth also remarking that the treatment above applies to samples in both Euclidean and non-Euclidean spaces. The only requirements are that it is possible to define a metric and that the volume of a ball can be computed as a function of its radius.

For this reason it is essential to describe the metric and ball volume for spaces, like those in which rotations and position–orientations are described, which are not Euclidean.

2.3 | Independence of samples

Before addressing metrics and volumes relevant for molecular entropy calculation, we remark that one of the key assumptions of the kNN method is that important correlations are properly sampled and that samples are independent of each other. For molecular dynamics simulations this translates into important requirements on the length of simulation and/or frequency of sampling.

If samples (e.g., MD snapshots) are taken very frequently, snapshots adjacent in sampling time will describe conformations close to each other, which will result in short distances and therefore in high densities in conformational space. Unless the same conformational region is sampled several times after memory of the starting configuration has been lost, the closest samples, that is, those on which entropy estimation is based in the kNN method, will not be independent of each other. They will be thus closer than expected for random sampling and the estimated entropy will be therefore largely negative, just as a result of inappropriate sampling.

One way to detect the non-independence of the samples due to too frequent sampling is to check correlation between conformational and time distances. If such correlation is found it means memory of the preceding configurations has not been lost and samples are not independent. The estimated entropy will be therefore lower than the true entropy. As an extreme case, if two identical configurations are present in the ensemble, for example, after a Montecarlo rejected move, the distance will be zero and the estimated entropy will be negative and infinite. Such cases must be treated ad hoc, but in general, judicious choices for sampling must be taken, as discussed in the Section 3.

2.4 | Metrics relevant for biomolecular processes

The kNN method is based on distances between configurational samples, so it is necessary to choose a molecular representation amenable to define a metric. Moreover, the large number of variables calls for a choice that reduces as far as possible correlations among variables, because it will be very difficult to treat, a posteriori, the mutual information among a large number of variables.

Atomic coordinates are not well suited for this task, because of the direct correlations within groups where atoms are in constrained reciprocal positions. In a sense, normal mode analysis, which has been used to describe motions and associated entropies at different time and length scales in proteins, addresses correlations among coordinates. However, the analysis is based on a linear model which is justified for small vibrations about mean positions, but not for the complex motions of macromolecules, in particular for entropy estimation, although many developments have been proposed to extend the method to anharmonic motions.[34]

The bond, angle, torsion (BAT) representation has been widely used instead, because obvious correlations are controlled by torsion angles. Bond and angles are defined as "hard" degrees of freedom compared with torsions defined as "soft" degrees of freedom, because the potentials restraining covalent geometry are stronger than those depending on dihedral angles.

The effect of interactions on bonds and angles is typically very small, so these degrees of freedom are often ignored. The entropy changes associated with bonds and angles are mostly about 10%–20% of those for torsional degrees of freedom.[52,53] Motions associated with "hard" degrees of freedom are dominated by the force field (typically harmonic) potentials, and are therefore referred to as vibrations. Note that in the literature the term "vibration" has been used including also fluctuations within minima wells for dihedral angles.

2.4.1 | Change of variables

Hnizdo and Gilson have thoroughly addressed the issue of entropy calculation under a change of variables.[44] Here we focus on how the issue is related to the kNN method. When changing variables $\{x\} \rightarrow \{y\}$, volumes change as $dx \rightarrow Jdy$ where J is the determinant of the Jacobian of the transformation. Densities change accordingly $\rho(x) \rightarrow \frac{\rho(x(y))}{J(y)}$. Entropy is estimated as

$$\widehat{S} = -k_B < \log(\rho(x)) > = -k_B \left\langle \log\left(\frac{\rho(y)}{J(y)}\right)\right\rangle. \tag{32}$$

The latter ensemble average is estimated according to the theory described above. The volume is computed in new coordinates as $\int Jdy$. The change of variables is correctly taken into account by the change of measure density in the variables' spaces, to recover the configurational ("spatial") entropy that would be calculated in the starting coordinates.

2.4.2 | Cartesian coordinates change to external and BAT coordinates

We consider BAT representation, with external coordinates to describe overall translation and rotation. It is important to study the change of variables and the determinant of the Jacobian of the transformation because the latter enters the integrals (in particular the integral that defines the volume of a ball in the transformed coordinates).

A detailed analysis of the conversion from Cartesian coordinates to BAT coordinates for a linear chain, was reported by Gō and Scheraga[54] who showed that the determinant of the Jacobian of the transformation is independent of torsional degrees of freedom. In that work the first atom coordinates defined the overall positional coordinate, two imaginary atoms were considered allowing for the definition of BAT coordinates for all atoms in the chain. The first bend angle and the first two torsion angles, defined using the two imaginary atoms, served as overall orientational coordinates.

The coordinates typically used for the overall position and orientation of a molecule in processing molecular trajectories are the center of geometry of a set of selected atoms and three angular coordinates to define the orientation about the latter center of geometry, after optimal superposition on the same set of selected atoms on a reference structure.

These considerations do not apply, in a straightforward way, to molecules which have more than one set of fairly rigid parts mobile with respect to each other. In the latter case it would be difficult to find a single position and rotation identifying the state of the molecule. It is worth noting that the same problem has been faced in the past and solutions have been provided that minimize in principle the dependence of external coordinates on internal coordinates.[55–57]

In the following we restrict ourselves to the single rigid core situation and adhere to the most common practice of identifying atoms in such rigid core and use the latter to define the overall position and orientation of the system.

The approach of Gō and Scheraga must be modified for this procedure to describe a different referencing for position and orientation.

Let us remark that the disadvantage of considering the first atoms in the linear chain to define position and orientation is that if that part of the molecule is very flexible compared with the rest of the molecule, orientational and positional coordinates will be strongly correlated to internal degrees of freedom, which makes entropy estimation more difficult.

The change of variables may be performed in four steps as shown in Figure 2 and detailed in Section SI.1. At the end of the change from atomic Cartesian coordinates to external and BAT coordinates, the molecule is described by:

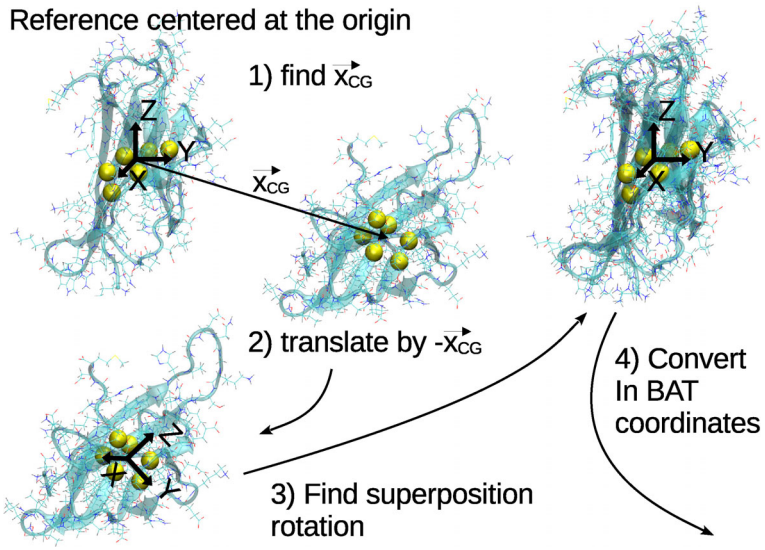

FIGURE 2 Steps for change from atomic Cartesian coordinates to external and internal BAT coordinates. The structure in the top right is the reference one with atoms selected for optimal superposition shown as yellow spheres. (1) The center of geometry of selected atoms, with respect to the same atoms in the reference structure, is found; (2) the structure is translated to the origin; (3) the optimal rotation to superimpose selected atoms is found; and (4) Cartesian coordinates are converted in BAT coordinates.

- three coordinates ($\vec{x}_{CG}$) for the center of geometry of the set of atoms chosen for superposition on a reference system
- three angular coordinates (ϕ,ψ,θ) to describe the rotation for optimal superposition on a reference system. ϕ and ψ identify the axis of rotation and θ is the rotation angle
- n_b variables describing bond lengths
- n_a variables describing bend angles
- n_t variables describing torsion angles

Assuming rigid bonds and bend angles and nearly rigid superposed atoms, we are left with three degrees of freedom for overall position ($\vec{x}_{CG}$), three degrees of freedom for overall orientation (ϕ,ψ,θ) and nt degrees of freedom for independent torsional angles named $\chi_1,\chi_2,...,\chi_{nt}$.

Under the same assumption the determinant of the Jacobian of the transformation:

$$\left\{\vec{x}_1,\vec{x}_2,...,\vec{x}_n\right\}\rightarrow\left\{\vec{x}_{CG},\phi,\psi,\theta,d_2,d_3,b_3,...,b_{nb},\xi,\xi_2,\xi_2,...,\xi_{na},\chi_1,\chi_2,...,\chi_{nt}\right\}, \tag{33}$$

may be written as

$$J=J_{\text{const}}\sin(\phi)(1-\cos(\theta)), \tag{34}$$

where J_{const} is the part of the determinant of the Jacobian depending only on bond lengths and bending angles, which are by assumption nearly constant and will be not further considered.

2.4.3 | Distances in external coordinates and BAT space

The kth nearest neighbor method is based on distances (defined according to a metric), therefore it is necessary to define the distance between any two samples of the system studied, in the space where it is represented. The space involves spatial and angular coordinates and we must define a distance that takes into account both kinds of coordinates.

A general metric that can be applied in multivariable spaces where a metric is defined for each variable or group of variables is the so-called Euclidean product metric.

It is easy to show that if d_1 is a metric on space X and d_2 is a metric on space Y, then $d=\sqrt{d_1^2+d_2^2}$ is a metric on the product space $X\bigotimes Y$.[58] Using Euclidean product metric we can define distances in complex spaces with different kinds of coordinates, and in particular it is convenient to breakdown large multivariable spaces like those of biomolecules in subspaces with obvious correlations among variables, for example, for rotations, translations, position–orientations and rotations about contiguous torsional angles. In these subspaces we can make use of the Euclidean product metric, and treat correlations among different variables or group of variables with specific tools (vide infra).

We review in the following the subspaces relevant for molecules and define for each subspace the distances between sample a and b.

Translation

The distance between the centers of geometry which identify the translation state of the molecule is the straightforward Euclidean distance in three dimensions

$$d_{ab}=\left\|\vec{x}_{CG,b}-\vec{x}_{CG,a}\right\|, \tag{35}$$

and the volume of a ball of radius d in three dimensions is the volume of the sphere:

$$V_t=\frac{4\pi}{3}d^3. \tag{36}$$

Rotation

There are many possible definitions of a distance in rotational space.[59,60] Based on Euler's theorem any rotation can be represented by an angle θ rotation about an axis $\vec{v}$. If the rotational state of a rigid body is described in the axis-angle representation, a natural definition of distance in rotational space is given by the angle needed to superpose the two samples of the same rigid body,[59–61] that is, the angle θ of the superposing rotation, independent therefore on the axis of rotation.

The rotation matrix may be expressed in terms of $\vec{v}$ and θ as:

$$R = \vec{v}\vec{v} + \left(1 - \vec{v}\vec{v}\right)\cos(\theta) + \text{skew}\left(\vec{v}\right)\sin(\theta), \tag{37}$$

where $\text{skew}(\vec{v})$ is the antisymmetric matrix whose element ij is $-\sum_k \varepsilon_{ijk} v_k$, with ε_{ijk} the Levi–Civita tensor.

When expressed in terms of the rotation matrices R_a and R_b corresponding to the rotation states a and b with respect to a common reference system, the distance is thus:

$$d_{ab} = \arccos\left(\frac{\text{Tr}\left(R_a^{-1}R_b\right) - 1}{2}\right), \tag{38}$$

where Tr indicates the trace operator. Note that the rotation angle is $0 \le d_{ab} \le \pi$.

The volume of a ball with radius d in rotational space is the integral over possible axis orientations (4π) and $d\theta$ from 0 to d, of the determinant of the Jacobian written above, save for constant terms, $J = \sin(\phi)(1 - \cos(\theta))$, that is:

$$V_r = \int_0^{2\pi} d\psi \int_0^{\pi} d\phi \sin(\phi) \int_0^{d} (1 - \cos(\theta)) d\theta = 4\pi(d - \sin(d)). \tag{39}$$

Position–orientation

The six degrees of freedom for translations and rotations can be considered together and the Euclidean product metric can be used to define a distance in the position–orientation space as first proposed heuristically by Huggins[62] and used later by Liedl et al.,[63] and Heinz and Grubmueller.[64] The approach was justified theoretically[65] computing the volume of a ball in the six-dimensional position–orientation space endowed with non-Euclidean metric and providing analytical approximations to the volume. In general it is useful to multiply the angular distance by a length, say f, to have common units for position and orientation distances and to weight the two distances in such a way that they contribute similarly to the Euclidean product metric distance in the product space.[65] Similar considerations and numerical tests have been recently presented by Heinz and Grubmueller.[64]

The distance in position–orientation space is defined as:

$$d_{ab} = \sqrt{\left\|\vec{x}_{CG,a} - \vec{x}_{CG,b}\right\|^2 + f^2 \left\|\arccos\frac{\text{Tr}\left(R_a^{-1}R_b\right) - 1}{2}\right\|^2}. \tag{40}$$

The volume of a ball of radius d, for small distances, in position–orientation space is approximated by a series expansion[65]:

$$V_{tr} \approx (\pi)^3 \sum_{l=1}^{n} a_l{}^* \frac{d^{4+2l}}{f_1^{2l+1}}, \tag{41}$$

with:

$$a_l = (-1)^{l+1} \frac{(2l+2)}{4^l (l+1)!(l+2)!}, \tag{42}$$

that is, for the first terms:

$$V_{tr} = \pi^3 \left(\frac{1}{12} \frac{d^6}{f^3} - \frac{1}{384} \frac{d^8}{f^5} + \frac{1}{23,040} \frac{d^{10}}{f^7} - \frac{1}{2,211,840} \frac{d^{12}}{f^9} + \frac{1}{309,657,600} \frac{d^{14}}{f^{11}} + \ldots \right). \tag{43}$$

The Euclidean product metric approach has been extended to treat the positional–orientational space of two molecules.[66] If we use a different scaling factor for each of the two molecules, that is, f_1 and f_2, the volume of the ball in the 12-dimensional space with radius d is approximated by:

$$V_{tr,12} \approx (\pi)^6 \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{a_l}{f_1^{2l+1}} \frac{c_{kl}}{f_2^{2k+1}} d_{kl} d^{2l+2k+8}, \tag{44}$$

with

$$c_{kl} = (-1)^{k+1} \frac{4^{l+2}(2k+2)(l+2)!(l+k+3)!}{(k+1)!(2(l+k+3))!}, \tag{45}$$

$$d_{kl} = \frac{(2(l+k+3))!}{4^{l+k+2}(l+k+3)!(l+k+4)!}. \tag{46}$$

The explicit form of the expansion up to the first terms is:

$$\begin{aligned} V_{tr,12} \approx (\pi)^6 \Bigg(& \frac{1}{2880} \frac{d^{12}}{f_1^3 f_2^3} \\ & - \frac{1}{161,280} \frac{d^{14}}{f_1^3 f_2^5} - \frac{1}{161,280} \frac{d^{14}}{f_1^5 f_2^3} \\ & + \frac{1}{15,482,880} \frac{d^{16}}{f_1^3 f_2^7} + \frac{1}{10,321,920} \frac{d^{16}}{f_1^5 f_2^5} + \frac{1}{15,482,880} \frac{d^{16}}{f_1^7 f_2^3} \\ & - \frac{1}{2,229,534,720} \frac{d^{18}}{f_1^3 f_2^9} - \frac{1}{1,114,767,360} \frac{d^{18}}{f_1^5 f_2^7} - \frac{1}{1,114,767,360} \frac{d^{18}}{f_1^7 f_2^5} - \frac{1}{2,229,534,720} \frac{d^{18}}{f_1^9 f_2^3} \\ & + \ldots) \end{aligned} \tag{47}$$

Torsion angles

The distance between two torsions is defined taking into account the periodicity of the rotation. If the periodicity is $\frac{2\pi}{n}$ the distance between two samples of the same torsion angle is:

$$d_{tors,ab} = \min_k \left\{ |\chi_a - \chi_b \pm k \frac{2\pi}{n}| \right\}, \tag{48}$$

and for m torsions:

$$d_{tors,m,ab} = \sqrt{\sum_{i=1,m} d_{ab,i}^2}, \tag{49}$$

where $d_{ab,i}$ is the distance between the two samples for the ith torsion angle.

The volume of a ball of radius d in an m-dimensional Euclidean space is:

$$V_{tors,m} = \frac{\pi^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}+1\right)} d^m, \tag{50}$$

where $\Gamma()$ is the gamma function.

2.5 | Implementation of the kth nearest neighbor method

We first discuss how the method is straightforwardly implemented. Provided that n samples are available, the distance d between any two samples is computed and the volume of a ball in the sample space with radius d, corresponding to each distance, is computed.

Then the method is implemented through the following steps:

1. the matrix (d_{ij}) of the distances between all pairs of samples (i and j) is computed;
2. each row i (corresponding to all distances from sample i) is sorted;
3. for each sorted distance (d_{ik}, where k now indicates the kth shortest distance, discarding the distance from each sample to itself) the volume ($V_{i,k}$) of a ball with that distance as radius is computed;
4. the logarithm of the volume is substituted in Equation 30 resulting in the mean over each column (corresponding to the first, the second, ..., the kth nearest neighbor).

Each value of k provides an estimate of the entropy.

A straightforward implementation presents two main problems.

First a choice must be made as to which value of k is used for entropy estimation. A compromise must be sought between low variance (i.e., large values of k) and fine resolution (small values of k).

Second, implementation of the method may be very time-consuming, because in principle all pairwise distances must be computed for all pairs of samples, and distances for each sample must be sorted, so that if we have n samples, computation is $O(n^2\log(n))$.

A large number of samples is necessary to increase resolution, on the other hand the mean operation, implied in the term $\sum_{i=1,n} \frac{\log(V_{i,k})}{n}$ in Equation (30), could be performed using a much smaller number of samples than available.

In one of the next sections we examine data structures and algorithms which greatly reduce the expected computational time.

To illustrate resolution and variance considerations we examine a simple 1D example. The probability density function (pdf) is the sum of four Gaussian functions with weights 4, 2, 1, 1, means 10, 25, 50, 54 and standard deviations 3, 5, 1, 1, respectively (Figure 3). This pdf has minor features which are resolved but close. We expect that a low number

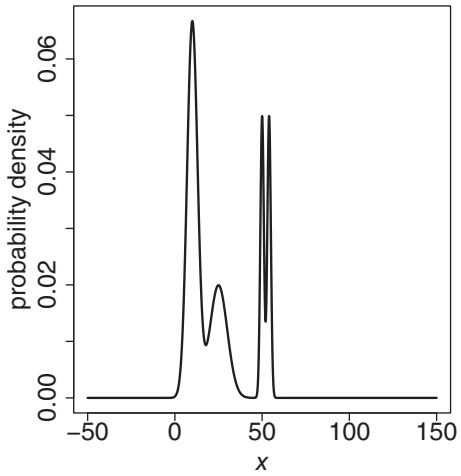

FIGURE 3 Example of a probability density function with features on different length scales.

of samples will result in limited resolution and in overestimation of the entropy. Two sets of 30 and 1000 samples were drawn from the distribution and the entropy was estimated using Equation (30).

To illustrate the effect of resolution on entropy estimation we graph the estimated entropy versus the average distance for each kth nearest neighbor (Figure 4). A clear increase of entropy estimates with the mean neighbor distance is observed. It is reasonable therefore to choose the first nearest neighbors (the number is arbitrary as long as the resolution is comparable or less than the length scales of the distribution) and fit the data, weighted by the inverse of the variance, to a line and use the intercept on the entropy axis to provide an estimate of entropy.

The effect of resolution is apparent in the two first rows of Figure 4. In the last row all n neighbors are considered for a limited number m of samples, in such a way that the time required for the computation is $O(mn\log(n))$ with a reduction in computations by a factor m/n. n may also be reduced, at the price of losing resolution.

Another important technical point concerns implementation of the method for multivariable samples spanning largely different ranges, for example, one or more orders of magnitude, possibly with different units of measures. When this happens, in a straightforward implementation, nearest neighbors distances will be dominated by the variables spanning the largest ranges and the effect of other variables will be negligible, unless the resolution is very fine, that is, the number of samples is very large. The consequence is that the space sampled has variations in fewer than the true dimensions. This is illustrated by the following example, variable x is drawn from a uniform pdf ranging between 0 and 1, then y is drawn from a Gaussian distribution with standard deviation equal 10 times x (Figure 5).

It is advisable in such cases to scale variables in such a way that they all have similar variances and thus contribute equally to the distance. The entropy change for scaling must be corrected afterwards, which might not be trivial if the metric is not Euclidean.

In the example (Figure 5) the analysis of unscaled data results in an average overestimation of the true entropy by 0.4 k_B, because the unscaled distribution is essentially one dimensional as far as distances are concerned at low resolution and mutual information between the two variables is completely overlooked. The two effects produce in this case

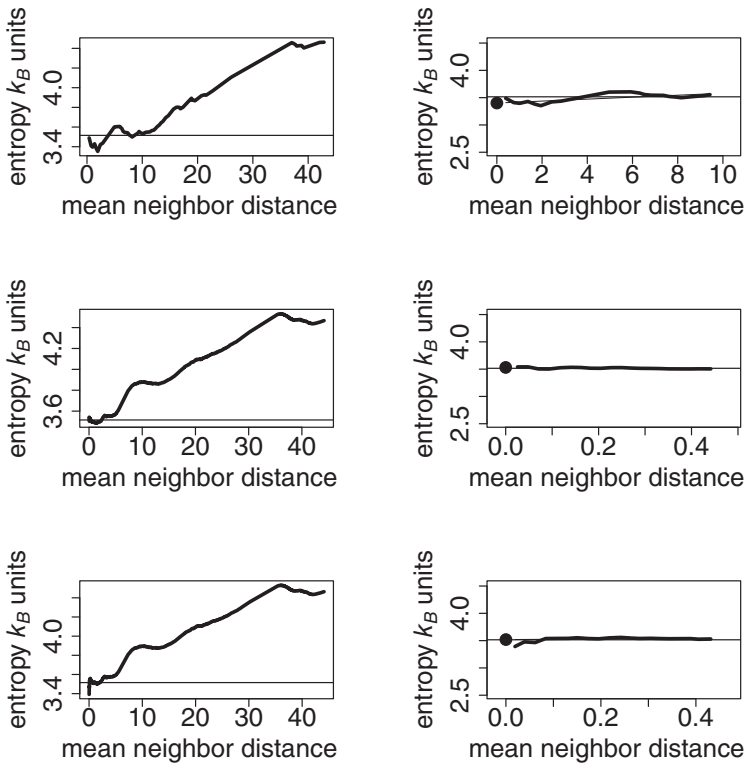

FIGURE 4 The kth nearest neighbor entropy estimates plotted versus kth nearest neighbor mean distance for the probability density function of Figure 3. On the left column data for all nearest neighbors are show, on the right column data for the first 20 neighbors and fitting line is shown. The entropy estimate obtained extrapolating the fitting line to zero mean nearest neighbor distance is shown as a filled sphere. The horizontal thin line is the exact value of the entropy for the probability density function. In the top row only 30 samples are analyzed. In the middle row 1000 samples are analyzed. In the bottom row 1000 samples are used, but the logarithm of the volume entering Equation (30) is averaged only over 100 samples.

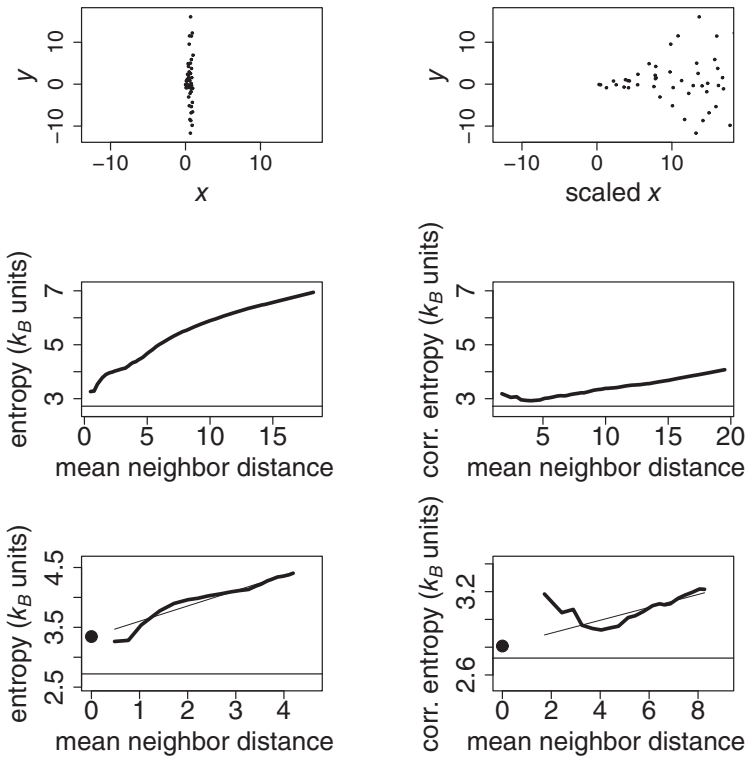

FIGURE 5 On the left column data and analysis are shown for raw data, on the right column corresponding data and analyses are shown for scaled data in such a way that the variances are the same in both dimensions. The displayed entropy on the right column is corrected for scaling. In the top row the 100 samples are shown, in the middle row the computed entropy versus mean neighbor distance for all values of k is shown, and in the bottom row line fitting, with points weighted by the inverse of the variance, for the first 20 neighbors is shown. The estimated entropy is indicated by a filled symbol. The true entropy is indicated by a thin horizontal line.

a small effect in entropy estimate. As can be seen estimates improve when scaling is applied to the first variable greatly reducing the average error in entropy.

2.6 | Data structures and algorithms for kNN searches

One of the major drawbacks of the kNN method is that, as implied by Equation 30, a naive implementation requires the computation of all the $\Theta(n^2)$ distances between samples. As noted by Vaidya,[67] the problem of finding the kth nearest neighbors for each point in a set of n samples has a lower bound of $\Omega(n\log(n))$ in the algebraic decision tree model of computation. To avoid the $\Omega(n^2)$ run-time complexity, in a recent review Brown et al.[68] consider appropriate data structures which may reduce computational times for nearest neighbor searches, both for periodic and non-periodic variables.

In particular, K-D and Vantage Point (VP) tree structures that partition the space into various regions are well suited for this purpose because, for each sample, its neighbors are found by visiting only some (but in the worst case, all) of them. Adopting such data structures implies that algorithms for computing samples' neighborhood will consist of two phases: first the tree is built from a set of samples, and then the tree is used to perform n searches, one for each of the n samples (Section SI.2).

Borelli et al.[69] analyzed the performance of K-D Trees, VP Trees, and Naive algorithms both in Euclidean and non-Euclidean spaces in realistic biophysical scenarios, demonstrating the efficiency of the K-D Tree method in Euclidean spaces and the good performance of the VP Tree method in non-Euclidean spaces, with time-saving of almost an order of magnitude with respect to the Naive method even in high-dimensional spaces.

K-D Tree construction and nearest neighbor search algorithms have also been implemented on GPUs with significant acceleration of kNN computations, in particular in scenarios involving high dimensions[70,71] (Section SI.2). It is also worth mentioning that the algorithms proposed by Hu et al.[71] can be easily adjusted to work with VP Trees.

3 | APPLICATIONS

The kNN method has been applied considering different degrees of freedom relevant for different problems in molecular simulations. The metric and the kind of mutual information among degrees of freedom are specific for each application. In the following we summarize contexts where the kNN method has been used in molecular sciences and provide examples from the literature.

3.1 | Conformational entropy

Conformations of molecule may be described using Cartesian or internal, for example, BAT, coordinates. The space of Cartesian coordinates is rather problematic for computing configurational entropy because of the extensive correlations among variables. However one of the most widely used approach for entropy estimation, due to its efficiency in computational time is the quasi-harmonic approach[27,32] and its modifications[12] using Cartesian coordinates. The main limitations of the quasi-harmonic approach are the non-Gaussian distribution of amplitudes of linearly independent motions and neglection of non-linear relationships among linearly independent motions, calling in turn for treatment of a large dimensional space.[34,72]

Ways to circumvent these limitations have been devised and have been reviewed.[12] In this context Knapp, Grubmueller et al.[73–76] used the kNN method to estimate corrections to the approach considering both non-Gaussian behavior of principal component motions, and the non-linear dependence of pairs of principal components, via the estimation of mutual information. The final entropy estimate was obtained summing corrected single variable entropies and pairwise mutual information. The approach of Numata et al.[76] was also used by Mukherjee et al. to calculate changes of solute entropy upon binding.[77]

The results confirmed that the upper bound on the entropy,[31,78] provided by the quasi harmonic analysis greatly overestimates the entropy and accounting for pairwise mutual information and, to a lesser extent, the anharmonicity of the distributions lowers importantly the upper bound on entropy.[76]

Recently, Marx et al. have used internal coordinates bond lengths and azimuthal and polar angle to describe the conformation of a fluxional molecule (protonated acetylene) and used the kNN method to assess its entropy.[79,80] The authors in particular assessed the two-, three-, and four-variable correlation using interaction information, expressed in terms of *n*-variable entropies, similar to the MIE expansion.[81]

Hnizdo et al. reported first the application of the kNN method for estimating the configurational entropy[82] of molecules, using only torsional degrees of freedom. Their work highlights also the need to deal in an efficient way with the entropy of more variables, necessary to estimate the entropy of the full-variable distribution. In particular they use the kNN method to estimate pairwise mutual information and use it to define a matrix of association coefficients, used in turn to cluster sets of variables. Then the entropies are computed for each set and summed up to provide an upper bound of the configurational entropy. To deal with the large dimensionality of the problem the kNN method has been used in combination with two main approaches based on mutual information: (i) a truncated mutual information expansion (MIE) approach showing its potential in molecular applications where important correlations are mostly involving torsion pairs[81,83,84]; the maximum information spanning tree (MIST) method proposed by Tidor et al.[85,86] involving also torsion pairs.[84] The MIST approach considers pairwise mutual informations and builds a tree corresponding to a chain of conditional entropies, which provides the lowest upper bound to the entropy, obtainable using mutual informations involving at most two variables. The approach has been used also grouping together variables and considering the mutual information between pairs of groups of variables.[53] The approach has been used to assess the effect of correlation on conformational entropy for a host-guest system,[84] to find the entropy of unfolded protein chains,[87] the change of entropy upon folding[87] and binding,[88–91] the change of entropy upon chemical modifications,[92] mutations[93] and to assess the relationship of the conformational entropy of proteins with NMR measured HN order parameter,[87] which has been proposed as an experimental "proxy" for conformational entropy.[94,95] In all these works the periodic Euclidean metric was used, but other metrics could be used as well as detailed by Hnizdo et al.[96]

The kNN method for conformational entropy is currently implemented in a few software programs. PDB2ENTROPY[53] uses trajectories in standard PDB format to compute torsion angles, according to user definitions. A file containing standard torsion angle definitions for proteins and nucleic acids is provided, and it can be easily be modified to define other torsion angles. Two schemes may be adopted:

i. each residue is treated independently from others, that is, only correlations within each residue are considered;
ii. torsion angles within each residue are grouped in fixed (small) size sets. The entropy is computed for each set and the mutual information for pairs of sets which have at least one atom within a chosen cutoff is computed. Finally the MIST approach is used to provide a (possibly strict) upper bound on entropy. If higher order correlations may be neglected the bound is close to the true entropy. For easier interpretation mutual informations are divided between the sets and entropies are listed also by residue.

In general, for practical consideration, simulations should be equilibrated, which means that the conformational space of correlated variables should be thoroughly explored during the simulation. In practice simulations of hundreds of nanoseconds are typically enough to sample torsional angles and pairs of correlated torsional angles in such a way that each conformational region is explored several times. If this is the case the sampling frequency should be such that the number of samples is large enough to have adequate resolution, for example, 5000–10,000 samples for pairs of torsion angles, and, on the other hand, such that the nearest neighbors of each sample are not correlated in time with it. Typically 10–20 ps is an appropriate choice. The time required, for example, on a 16-core 11th Gen Intel(R) Core(TM) i7-11850H @ 2.50 GHz to analyze a system entailing 866 torsional angles and 8745 torsional angle pairs for MIST analysis, is 120, 1290, and 15,770 s, using 1000, 3000, and 10,000 samples, respectively. Running time scales, in a straightforward implementation, with $n^2 \log(n)$ where n is the number of samples.

As an example of application we consider here the amino acid phenylalanine which is described in the BAT representation by four torsional angles (ϕ, ψ, χ_1, and χ_2, and bonds and bend angles which are neglected here, Figure 6). A sample of the conformations explored in the unfolded state is provided by a set of 5000 conformations obtained from a non-redundant set of 3600 crystallographic protein structures in the culled PDB dataset.[97]

For such ensemble, the entropies associated with each torsional angle, with respect to a uniform distribution, are: -1.1, -0.8, -1.2, and -0.6 k_B for ϕ, ψ, χ_1, and χ_2, respectively. The maximum information spanning tree includes the mutual informations between ϕ and ψ (-0.6 k_B), ψ and χ_1 (-0.5 k_B), and χ_1 and χ_2 (-0.5 k_B). The total entropy is estimated as -5.3 k_B. As an ensemble representative of a specific phenylalanine in a folded state we considered 2000

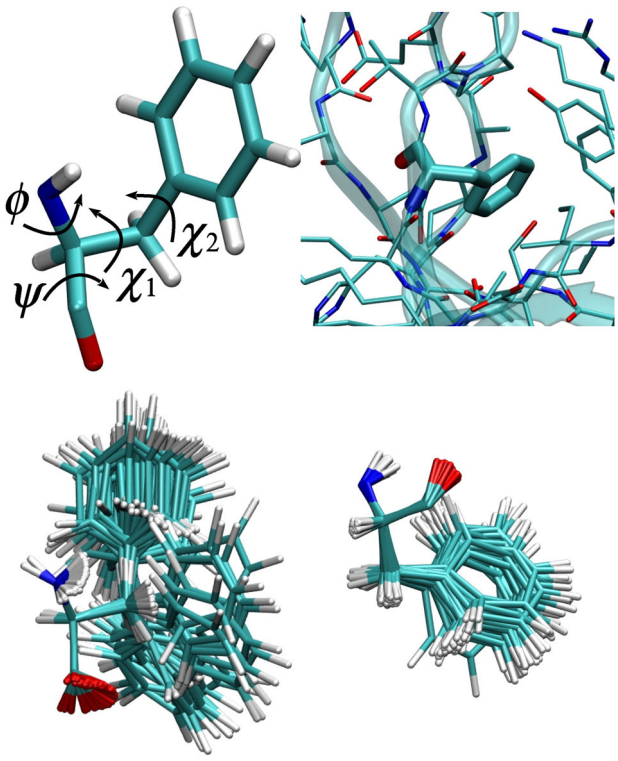

FIGURE 6 Entropy of a phenylalanine residue in an unfolded and folded protein. Top left panel: Phenyalanine residue torsion angles; top right panel: Phenylalanine 70 in the folded β2-microglobulin protein; bottom left panel: 50 coil conformations from a non-redundant data set as representative of the unfolded state; bottom right panel: 50 conformations from a 200 ns molecular dynamics simulation of the folded protein.

snapshots of Phe70 from a 200 ns molecular dynamics simulation of β2-microglobulin. As expected, and apparent in Figure 6 entropies associated with each torsional angle in the folded state, with respect to a uniform distribution, are lower: −1.9, −2.0, −2.4 and −1.2 k_B for ϕ, ψ, χ_1, and χ_2, respectively. The maximum information spanning tree includes the mutual informations between ϕ and ψ (−0.1 k_B), ϕ and χ_1 (−0.1 k_B), and χ_1 and χ_2 (−0.1 k_B). The mutual information of rather restricted torsion angles is lesser than for the wider conformational space explored in the unfolded state. The total entropy is estimated as −7.8 k_B. The reduction in conformational freedom for β2-microglobulin Phe 70 upon folding results therefore in 2.5 k_B entropy loss. The example of analysis reported here is extended to all torsional angles and mutual informations between close torsional angles, to obtain the whole protein conformational entropy.

3.2 | Positional–orientational entropy

Another outstanding field of application of the kNN method is the calculation of the positional–orientational entropy changes upon binding. Following McCammon et al.[4] the term positional–orientational entropy is preferred over translational–rotational entropy to avoid confusion with the entropy associated with linear and angular momenta. Notwithstanding the caveat about decomposition of entropy in independent contributions, if the positional–orientational state description of the molecule is properly chosen, it may be reasonably decoupled from internal degrees of freedom, as discussed above. Thermodynamics of binding has been described in excellent reviews.[4,7,8,10,98] The calculation of positional–orientational entropy is relevant for binding, where two molecules associate, but also, as discussed in the next subsection, for solvation, where waters' positions and orientations are changed by the presence of the solute and by other waters.

The kNN method for positional–orientational entropy, using the theory reported above has been implemented in the program PDB2TRENT.[53] PDB2TRENT combines orientational and positional distances to define distances in the orientation/position space of the molecules and to compute entropies based on the kNN method, which is particularly useful to define the entropy changes upon binding of molecules.

Applications have been reported for the entropy of protein association with hydrophobic surfaces,[88] with proteins,[65] protein–ligand binding,[87] and peptide liquid–liquid phase separation.[99,100]

For practical considerations, sampling in the six dimensional space of position/orientation would require a very large number of samples. On the other hand, typically, the relative position of a ligand with respect to its target is restricted in a space of $\approx 1 \text{Å}^3$ and $\approx 1 \text{degree}^3$. In practice 10,000 samples are typically sufficient to achieve the necessary resolution.

As an example, we consider the association of two proteins, in the present case two transthyretin monomers. The reference state is the ideal freely-rotating, 1 M concentration state of both proteins. The associated state has been simulated for 250 ns and 25,000 snapshots have been collected to provide a representative ensemble. The positional–orientational entropy of the associated proteins is computed considering the first of the two monomers in the ideal freely rotating, 1 M concentration state, and considering the configurational restriction of the second monomer with respect to the first one. To do this the complex is first superimposed using the backbone atoms in secondary structure regions of the first monomer, then the positional–orientational entropy of the second monomer, now restricted in position and orientation is computed, using again the backbone atoms in secondary structure regions (Figure 7). The whole calculation (superposition and entropy estimation) takes about 130 s on a 16-core 11th Gen Intel(R) Core(TM) i7-11850H @ 2.50 GHz. The results show that the positional entropy is −6.0 k_B and the orientational entropy is −9.7 k_B. The mutual information between the orientational and positional degrees of freedom lowers the entropy by an additional 1.8 k_B, totaling −17.5 k_B positional–orientational entropy.

3.3 | Solvation entropy

Hydration is perhaps one of the most important fields where the kNN method is currently applied. The problem involves the positional–orientational entropy of each water molecule and its correlations. The large dimensionality of the space precludes the possibility of using histograms to estimate probabilities. An additional complication arises from the indistinguishability of solvent molecules, which calls for specific treatment of molecular dynamics trajectories where molecules and their symmetric parts bear a specific label.

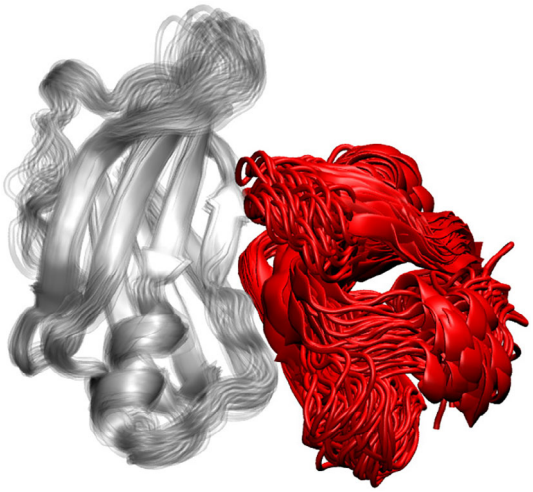

FIGURE 7 Ensemble of 50 configurations of a transthyretin dimer shown in cartoon. The backbone atoms of secondary structure elements of the leftmost monomer (represented in white transparent material) have been used to superimpose the whole complex. The positional–orientational entropy is computed for the rightmost monomer (in red) using the backbone atoms of secondary structure elements.

Berne et al. were the first to recognize the distinctive advantages of the kNN method[101] over other more traditional methods to estimate hydration entropies. In their approach the excess entropy of water due to pair correlation function, was treated using a single radial variable and five orientational variables, following Lazaridis and Karplus.[102] A number of approximations on the multivariable distribution, including recasting the oxygen–oxygen distance distribution in a three-value distribution and the Generalized Kirkwood superposition approximation, made the problem treatable and the results were in good agreement with more costly and accurate pathway methods.

Starting from inhomogeneuos solvation theory (IST),[103,104] Gilson et al. used a grid approach to the estimate the entropy of waters around a solute neglecting water–water correlations.[105] The orientational part of the single-water entropy was estimated using the kNN method. A further step was using the kNN method for both the orientational and positional entropy and also for full positional–orientational entropy[106,107] in the software GIST, using the Euclidean product metric[62] and later in the software SSTMap.[108] Liedl et al. made a GPU implementation of the GIST approach (GIGIST) available.[63,109] The same authors have been extending recently the GIST approach to salt-water solutions and other, rigid, solvents.[110,111] The method is implemented in the software GIST,[106,107] SSTMap,[108] PME-GIST,[112] GIGIST,[63,109] which are based on single molecule entropy computed on grid cells. A different approach is implemented in the software Per|Mut,[64,113] which uses permutation of water labels to increase single water sampling, and the mutual information expansion[83] to deal with the large dimensionality.

It must be noted that it is the kNN method which enables the treatment of the high-dimensionality position–orientation space of single waters.

The entropy arising from water–water correlation is typically neglected in these approaches. Based on the observation that for many solutes solute–water and water–water interaction energy contributions to solvation energy are in a ratio of $-\frac{1}{2}$ Huggins[62] proposed to set the entropy change in water–water position–orientation equal to $-\frac{1}{2}$ of the single water molecule entropy. Apparently such heuristic approximation gives very good results for small molecules hydration entropy. Very recently Kurtzman et al. reported a detailed study of the solvation entropy of 32 small molecules computed with a novel version of GIST, PME-GIST,[112] and showed that higher order water entropy terms can be estimated as -0.4 times the first order water entropy.[112] This value is suggested by linear fitting computed entropies versus the reference value provided by thermodynamic integration and is fully justified by the authors based on theoretical ground. PME-GIST, which uses the kNN method, is implemented in the software CPPtraj.[114] The accuracy their approach obtains on test molecules sampling almost continuously a 20 kcal/mol solvation free energy range, is 0.4 kcal/mol (unsigned mean error).

It is worth noting here that comparison of kNN based solvation entropies with those obtained, for example, by pathway method provides a test, which is difficult to perform for conformational and positional–orientational entropies, unless in gas-phase. In solution, indeed, pathway methods provide free energy which include also solvation, which makes comparison not straightforward.

The positional part of the entropy arising from water–water correlation was addressed using the kNN method[115] and the full two-molecule entropy was estimated by Huggins[116] considering the 12-dimensional space of positions and orientations of two waters, restricting the analysis within a cutoff of 4 Å.

Most approaches mentioned above and others face the problem of indistinguishability of solvent molecules by using water density on small voxels and/or permutation of labels[117] as proposed by Grubmuller et al.[118,119] Recently the same authors have used permutation of labels to compute hydration entropies[64,113] using the kNN method coupled with mutual information expansion up to third order for orientational entropies (using quaternion metric on single orientational space and Euclidean product metric for pairs and triplets of orientations) and later for positional–orientational entropies with excellent results.

The specific linkage of kNN method and optimal relabeling of water molecules with hydration thermodynamics has been reviewed by us recently.[120] The approximations involved in the approach imply very small errors on the estimation of single-molecule entropies. The exact treatment of the full positional–orientational space of two molecules[66] enables in principle the computation of entropy terms due to water–water correlations.

For practical considerations, due to the fast reorientation of bulk water a sampling frequency of 1 ps may be used, and simulations, considering also more hindered surficial waters, may be in the range of tens of ns. The running time for a thorough analysis depends on many factors, like the extension of the solvent region considered for the analysis, preprocessing of the trajectory, the spacing of the grid if a grid is used, and so on. Considerations similar to those for the computation of positional–orientational entropy however apply also here.

4 | CONCLUSIONS

The kNN method has distinctive advantages over other methods for estimating the entropy of biomolecular processes:

- it does not depend on a specific model for the probability density function of variables;
- it allows for decomposition of entropy in terms of single and groups of variables describing configurations;
- it adapts the neighborhood of each sample to the estimated probability density at that sample, providing fine resolution in dense configurational space, that is, in the regions most contributing to the entropy;
- not requiring predefined intervals for estimating the density of the distribution it is suitable to address large dimensional spaces.

The disadvantages are:

- the large computational load for which however straightforward parallelization is possible. Moreover specific data structures and algorithms may be used largely reducing the expected computational time;
- samples should be independent on each other, which means that samples must be spaced in time, which calls for long simulations.

The recent availability of software performing the kNN entropy estimation makes this method a useful tool that allows the analysis for the complex probability density function of complex molecules in solution.

AUTHOR CONTRIBUTIONS

Federico Fogolari: Conceptualization (lead); methodology (lead); writing – original draft (lead); writing – review and editing (equal). **Roberto Borelli:** Conceptualization (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (equal). **Agostino Dovier:** Conceptualization (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (equal). **Gennaro Esposito:** Conceptualization (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (equal).

FUNDING INFORMATION

This work was partly supported by a grant by the European Union – Next Generation EU (University of Udine CUP: G25F21003390007).

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Federico Fogolari https://orcid.org/0000-0001-9862-250X
Roberto Borelli https://orcid.org/0000-0003-2586-8183

RELATED WIREs ARTICLES

Direct methods for computing single-molecule entropies from molecular simulations

REFERENCES

1. Atkins P, de Paula J. Physical chemistry for the life sciences. Oxford: Oxford University Press; 2006.
2. Goldenzweig A, Fleishman SJ. Principles of protein stability and their application in computational design. Annu Rev Biochem. 2018; 87:105–29.
3. Beveridge DL, diCapua L. Free energy via molecular simulation: applications to chemical and biomolecular systems. Annu Rev Biophys Chem. 1989;18:431–92.
4. Gilson MK, Given JA, Bush BL, McCammon JA. The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys J. 1997;72:1047–69.
5. Roux B, Simonson T. Implicit solvent models. Biophys Chem. 1999;78:1–20.
6. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res. 2000;33:889–97.
7. Gilson MK, Zhou HX. Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct. 2007;36:21–42.
8. Zhou HX, Gilson MK. Theory of free energy and entropy in noncovalent binding. Chem Rev. 2009;109:4092–107.
9. Polyansky AA, Zubac R, Zagrovic B. Estimation of conformational entropy in protein-ligand interactions: a computational perspective. Methods Mol Biol. 2012;819:327–53.
10. Wereszczynski J, McCammon JA. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. Q Rev Biophys. 2012;45:1–25.
11. Suarez D, Diaz N. Direct methods for computing single-molecule entropies from molecular simulations. WIREs Comput Mol Sci. 2015; 5:1–26.
12. Kassem S, Ahmed M, El-Sheikh S, Barakat KH. Entropy in bimolecular simulations: a comprehensive review of atomic fluctuations-based methods. J Mol Graph Model. 2015;62:105–17.
13. Chong SH, Chatterjee P, Ham S. Computer simulations of intrinsically disordered proteins. Annu Rev Phys Chem. 2017;68:117–34.
14. Fogolari F, Corazza A, Esposito G. Free energy, enthalpy and entropy from implicit solvent end-point simulations. Front Mol Biosci. 2018;5:11.
15. Chipot C. Free energy methods for the description of molecular processes. Annu Rev Biophys. 2023;52:113–38.
16. Straatsma T, McCammon J. Computational alchemy. Annu Rev Phys Chem. 1992;43:407–35.
17. Zwanzig RW. High temperature equation of state by a perturbation method. I Nonpolar gases. J Chem Phys. 1954;22:1420–6.
18. Straatsma TP, McCammon JA. Multiconfiguration thermodynamic integration. J Chem Phys. 1954;95:1175–88.
19. Torrie GM, Valleau JP. Nonphyisical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comp Phys. 1977;23:187–99.
20. Laio A, Parrinello M. Escaping free energy minima. Proc Natl Acad Sci U S A. 2002;99:12562–6.
21. Darve E, Pohorille A. Calculating free energies using average force. J Chem Phys. 2001;115(20):9169–83.
22. Hamelberg D, Mongan J, McCammon JA. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys. 2004;120:11919–29.
23. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The weighted histogram analysis method for free-energy calculations on biomolecules. I The method. J Comp Chem. 1992;13(8):1011–21.
24. Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys. 2008;129(12): 124105.
25. Jarzynski C. Nonequilibrium equality for free energy differences. Phys Rev Lett. 1997;78(14):2690–3.
26. Crooks GE. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. J Stat Phys. 1998;90:1481–7.
27. Karplus M, Kushick JN. Method for estimating the configurational entropy of native macromolecules. Macromolecules. 1981;14: 325–32.
28. Levy RM, Karplus M, Kushick JN, Perahia D. Method for estimating the configurational entropy of native macromolecules. Macromolecules. 1984;17:1370–4.

29. Karplus M, Ichiye T, Pettitt BM. Configurational entropy of native proteins. Biophys J. 1987;52(6):1083–5.
30. Tidor B, Karplus M. The contribution of cross-links to protein stability: a normal mode analysis of the configurational entropy of the native state. Proteins. 1993;15:71–9.
31. Schlitter J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chem Phys Lett. 1993;215: 617–21.
32. Andricioaei I, Karplus M. On the calculation of entropy from covariance matrices of the atomic fluctuations. J Chem Phys. 2001;115: 6289–92.
33. Baron R, van Gunsteren WF, Hünenberger PH. Estimating the configurational entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the quasi-harmonic approximation. Trends Phys Chem. 2006;11:87–122.
34. Baron R, Hünenberger PH, McCammon JA. Absolute single-molecule entropies from quasi-harmonic analysis of microsecond molecular dynamics: correction terms and convergence properties. J Chem Theory Comput. 2009;5(12):3150–60.
35. Hikiri S, Yoshidome T, Ikeguchi M. Computational methods for configurational entropy using internal and Cartesian coordinates. J Chem Theory Comput. 2016;12(12):5990–6000.
36. Suarez E, Diaz N, Suarez D. Entropy calculations of single molecules by combining the rigid–rotor and harmonic–oscillator approximations with conformational entropy estimations from molecular dynamics simulations. J Chem Theory Comput. 2011;7(8):2638–53.
37. Chen W, Chang CE, Gilson MK. Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. Biophys J. 2004;87(5):3035–49.
38. Chang CA, Chen W, Gilson MK. Ligand configurational entropy and protein binding. Proc Natl Acad Sci U S A. 2007;104(5):1534–9.
39. Pereira GP, Cecchini M. Multibasin quasi-harmonic approach for the calculation of the configurational entropy of small molecules in solution. J Chem Theory Comput. 2021;17(2):1133–42.
40. Diaz N, Suarez D. Toward reliable and insightful entropy calculations on flexible molecules. J Chem Theory Comput. 2022;18:7166–78.
41. Kozachenko LF, Leonenko NN. Sample estimates of entropy of a random vector. Probl Inf Transm. 1987;23:95–101.
42. Singh H, Misra N, Hnizdo V, Fedorowicz A, Demchuk E. Nearest neighbor estimate of entropy. Am J Math Manag Sci. 2003;23:301–21.
43. McQuarrie DA. Statistical mechanics. Sausalito: University Science Books; 2000.
44. Hnizdo V, Gilson MK. Thermodynamic and differential entropy under a change of variables. Entropy. 2010 Mar;12(3):578–90.
45. Sharp K, Matschinsky F. Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, pp 373–435 (Wien. Ber. 1877, 76: 373–435). Reprinted in Wiss. Abhandlungen, Vol. II, reprint 42, p. 164–223, Barth, Leipzig, 1909. Entropy. 2015;17(4):1971–2009.
46. Gibbs JW. Elementary principles in statistical mechanics: developed with special reference to the rational foundation of thermodynamics. New York: Dover Publications; 1902.
47. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.
48. von Neumann J. Mathematical foundations of quantum mechanics: new edition. Vol 53. Princeton: Princeton University Press; 2018.
49. Goria MN, Leonenko NN, Mergel VV, Inverardi PLN. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. J Nonparam Stat. 2005;17:277–97.
50. Berrett TB, Samworth RJ, Yuan M. Efficient multivariate entropy estimation via k-nearest neighbour distances. Ann Stat. 2019;47: 288–318.
51. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004;69:066138.
52. Killian BJ, Yundenfreund Kravitz J, Somani S, Dasgupta P, Pang YP, Gilson MK. Configurational entropy in protein-peptide binding: computational study of Tsg101 ubiquitin E2 variant domain with an HIV-derived PTAP nonapeptide. J Mol Biol. 2009;389:315–35.
53. Fogolari F, Maloku O, Dongmo Foumthuim CJ, Corazza A, Esposito G. PDB2ENTROPY and PDB2TRENT: conformational and translational–rotational entropy from molecular ensembles. J Chem Inf Model. 2018;58:1319–24.
54. Go N, Scheraga HA. On the use of classical statistical mechanics in the treatment of polymer chain conformation. Macromolecules. 1976;9:535–42.
55. Eckart C. Some studies concerning rotating axes and polyatomic molecules. Phys Rev. 1935;47:552–8.
56. Sayvetz A. The kinetic energy of polyatomic molecules. J Chem Phys. 1939;7:383–9.
57. Szalay V. Eckart- Sayvetz conditions revisited. J Chem Phys. 2014;140:234107.
58. Ò Searcòid M. Metric spaces. London: Springer-Verlag; 2007.
59. Huynh DQ. Metrics for 3D rotations: comparison and analysis. J Math Imaging Vis. 2009;35:155–64.
60. Huggins DJ. Comparing distance metrics for rotation using the k-nearest neighbors algorithm for entropy estimation. J Comput Chem. 2014;35:377–85.
61. Miles RE. On random rotations in R^3. Biometrika. 1965;52:636–9.
62. Huggins DJ. Estimating translational and orientational entropies using the k-nearest neighbors algorithm. J Chem Theory Comput. 2014;10:3617–25.
63. Kraml J, Hofer F, Kamenik AS, Waibl F, Kahler U, Schauperl M, et al. Solvation thermodynamics in different solvents: water–chloroform partition coefficients from grid inhomogeneous solvation theory. J Chem Inf Model. 2020;60(8):3843–53.
64. Heinz LP, Grubmüller H. Per|Mut: spatially resolved hydration entropies from atomistic simulations. J Chem Theory Comput. 2021;17: 2090–1.
65. Fogolari F, Dongmo Foumthuim CJ, Sr F, Soler MA, Corazza A, Esposito G. Accurate estimation of the entropy of rotation–translation probability distributions. J Chem Theory Comput. 2016;12:1–8.

66. Fogolari F, Esposito G, Tidor B. Entropy of two-molecule correlated translational-rotational motions using the kth nearest neighbour method. J Chem Theory Comput. 2021;17:3039–51.
67. Vaidya PM. An O(n log n) algorithm for the all-nearest neighbors problem. Discrete Comput Geom. 1989;4:101–15.
68. Brown JM, Bossomaier T, Barnett L. Review of data structures for computationally efficient nearest-neighbour entropy estimators for large systems with periodic boundary conditions. J Comput Sci. 2017;23:109–17.
69. Borelli R, Dovier A, Fogolari F. Data structures and algorithms for k-th nearest neighbours conformational entropy estimation. Biophysica. 2022;2(4):340–52.
70. Garcia V, Debreuve E, Barlaud M. Fast k nearest neighbor search using GPU. In: 2008 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE; 2008. p. 1–6.
71. Hu L, Nooshabadi S, Ahmadi M. Massively parallel KD-tree construction and nearest neighbor search algorithms. In: 2015 IEEE international symposium on circuits and systems (ISCAS). IEEE; 2015. p. 2752–2755.
72. Chang CE, Chen W, Gilson MK. Evaluating the accuracy of the quasiharmonic approximation. J Chem Theory Comput. 2005;1:1017–28.
73. Lange OF, Grubmüller H. Generalized correlation for biomolecular dynamics. Proteins. 2006;62:1053–61.
74. Lange OF, Grubmüller H. Full correlation analysis of conformational protein dynamics. Proteins. 2008;70:1294–312.
75. Hensen U, Lange OF, Grubmuller H. Estimating absolute configurational entropies of macromolecules: the minimally coupled subspace approach. PLoS One. 2010;5:e9179.
76. Numata J, Wan M, Knapp EW. Conformational entropy of biomolecules: beyond the quasi-harmonic approximation. Genome Inform. 2007;18:192–205.
77. Mukherjee A. Entropy balance in the intercalation process of an anti-cancer drug daunomycin. J Phys Chem Lett. 2011;2:3021–6.
78. Jaynes ET. Information theory and statistical mechanics. Phys Rev. 1957;106:620–30.
79. Topolnicki R, Brieuc F, Schran C, Marx D. Deciphering high-order structural correlations within fluxional molecules from classical and quantum configurational entropy. J Chem Theory Comput. 2020;16(11):6785–94.
80. Beckmann R, Topolnicki R, Marx D. Deciphering the impact of helium tagging on flexible molecules: probing microsolvation effects of protonated acetylene by quantum configurational entropy. J Phys Chem A. 2023;127(11):2460–71.
81. Killian BJ, Yundenfreund Kravitz J, Gilson MK. Extraction of configurational entropy from molecular simulations via an expansion approximation. J Chem Phys. 2007;127:024107.
82. Hnizdo V, Darian E, Fedorowicz A, Demchuk E, Li S, Singh H. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. J Comput Chem. 2007;28(3):655–68.
83. Hnizdo V, Tan J, Killian BJ, Gilson MK. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. J Comput Chem. 2008;29:1605–14.
84. Fenley AT, Killian BJ, Hnizdo V, Fedorowicz A, Sharp DS, Gilson MK. Correlation as a determinant of configurational entropy in supramolecular and protein systems. J Phys Chem B. 2014;118:6447–55.
85. King BM, Tidor B. MIST: maximum information spanning trees for dimension reduction of biological data sets. Bioinformatics. 2009;25:1165–72.
86. King BM, Silver NW, Tidor B. Efficient calculation of molecular configurational entropies using an information theoretic approximation. J Phys Chem B. 2012;116:2891–904.
87. Fogolari F, Corazza A, Fortuna S, Soler MA, VanSchouwen B, Brancolini G, et al. Distance-based configurational entropy of proteins from molecular dynamics simulations. PloS One. 2015;10:e0132356.
88. Dongmo Foumthuim CJ, Corazza A, Esposito G, Fogolari F. Molecular dynamics simulations of β2-microglobulin interaction with hydrophobic surfaces. Mol Biosyst. 2017;13:2625–37.
89. Verteramo ML, Stenström O, Ignjatović MM, Caldararu O, Olsson MA, Manzoni F, et al. Interplay between conformational entropy and solvation entropy in protein-ligand binding. J Am Chem Soc. 2019;141:2012–26.
90. Qaisrani MN, Belousov R, Rehman JU, Goliaei EM, Girotto I, Franklin-Mergarejo R, et al. Phospholipids dock SARS-CoV-2 spike protein via hydrophobic interactions: a minimal in-silico study of lecithin nasal spray therapy. Eur Phys J E Soft Matter. 2021;44(11):132.
91. Baweja L, Wereszczynski J. Conformational and thermodynamic differences underlying wild-type and mutant eleven-nineteen-leukemia YEATS domain specificity for epigenetic marks. J Chem Inf Model. 2023;63(4):1229–38.
92. Sgrignani J, Chen J, Alimonti A, Cavalli A. How phosphorylation influences E1 subunit pyruvate dehydrogenase: a computational study. Sci Rep. 2018;8:14683.
93. Ponleitner M, Szollosi D, El-Kasaby A, Koban F, Freissmuth M, Stockner T. Thermal unfolding of the human serotonin transporter: differential effect by stabilizing and destabilizing mutations and cholesterol on thermodynamic and kinetic stability. Mol Pharmacol. 2022;101(2):95–105.
94. Li Z, Raychaudhuri S, Wand AJ. Insights into the local residual entropy of proteins provided by NMR relaxation. Protein Sci. 1996;5:2647–50.
95. Frederick KK, Marlow MS, Valentine KG, Wand AJ. Conformational entropy in molecular recognition by proteins. Nature. 2007;448:325–9.
96. Misra N, Singh H, Hnizdo V. Nearest neighbor estimates of entropy for multivariate circular distributions. Entropy. 2010 Mar;12(3):1125–44.
97. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003;19:1589–91.
98. Decherchi S, Cavalli A. Thermodynamics and kinetics of drug-target binding by molecular simulation. Chem Rev. 2020;120:12788–833.

99. Workman RJ, Pettitt BM. Thermodynamic compensation in peptides following liquid-liquid phase separation. J Phys Chem B. 2021; 125(24):6431–9.
100. Workman RJ, Gorle S, Pettitt BM. Effects of conformational constraint on peptide solubility limits. J Phys Chem B. 2022;126(49): 10510–8.
101. Wang L, Abel R, Friesner RA, Berne BJ. Thermodynamic properties of liquid water: an application of a nonparametric approach to computing the entropy of a neat fluid. J Chem Theory Comput. 2009;5(6):1462–73.
102. Lazaridis T, Karplus M. Orientational correlations and entropy in liquid water. J Chem Phys. 1996;105:4294–316.
103. Lazaridis T. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory J Phys Chem B. 1998;102:3531–41.
104. Lazaridis T. Inhomogeneous fluid approach to solvation thermodynamics. 2. Applications to simple fluids. J Phys Chem B. 1998;102: 3542–50.
105. Nguyen CN, Young TK, Gilson MK. Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. J Chem Phys. 2012;137:044101.
106. Nguyen CN, Cruz A, Gilson MK, Kurtzman T. Thermodynamics of water in an enzyme active site: grid-based hydration analysis of coagulation factor Xa. J Chem Theory Comput. 2014;10:2769–80.
107. Ramsey S, Nguyen C, Salomon-Ferrer R, Walker RC, Gilson MK, Kurtzman T. Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST. J Comput Chem. 2016;37:2029–37.
108. Haider K, Cruz A, Ramsey S, Gilson MK, Kurtzman T. Solvation structure and thermodynamic mapping (SSTMap): an open-source, flexible package for the analysis of water in molecular dynamics trajectories. J Chem Theory Comput. 2018;14:418–25.
109. Kraml J, Kamenik AS, Waibl F, Schauperl M, Liedl KR. Solvation free energy as a measure of hydrophobicity: application to serine protease binding interfaces. J Chem Theory Comput. 2019;15(11):5872–82.
110. Waibl F, Kraml J, Fernández-Quintero ML, Loeffler JR, Liedl KR. Explicit solvation thermodynamics in ionic solution: extending grid inhomogeneous solvation theory to solvation free energy of salt–water mixtures. J Comput Aid Mol Des. 2022;156:1–16.
111. Waibl F, Kraml J, Hoerschinger VJ, Hofer F, Kamenik AS, Fernández-Quintero ML, et al. Grid inhomogeneous solvation theory for cross-solvation in rigid solvents. J Chem Phys. 2022;156(20):204101.
112. Chen L, Cruz A, Roe DR, Simmonett AC, Wickstrom L, Deng N, et al. Thermodynamic decomposition of solvation free energies with particle mesh Ewald and long-range Lennard-Jones interactions in grid inhomogeneous solvation theory. J Chem Theory Comput. 2021;17(5):2714–24.
113. Heinz LP, Grubmüller H. Computing spatially resolved rotational hydration entropies from atomistic simulations. J Chem Theory Comput. 2020;16:108–18.
114. Roe DR, Cheatham TE III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. J Chem Theory Comput. 2013;9(7):3084–95.
115. Nguyen CN, Kurtzman T, Gilson MK. Spatial decomposition of translational water–water correlation entropy in binding pockets. J Chem Theory Comput. 2016;12:414–29.
116. Huggins DJ. Quantifying the entropy of binding for water molecules in protein cavities by computing correlations. Biophys J. 2015;108: 928–36.
117. Sasikala WD, Mukherjee A. Single water entropy: hydrophobic crossover and application to drug binding. J Phys Chem B. 2014;118: 10553–64.
118. Reinhard F, Grubmüller H. Estimation of absolute solvent and solvation shell entropies via permutation reduction. J Chem Phys. 2007; 126:014102.
119. Reinhard F, Lange OF, Hub JS, Haas J, Grubmüller H. g_permute: permutation-reduced phase space density compaction. Comput Phys Commun. 2009;180:455–8.
120. Fogolari F, Esposito G. Optimal relabeling of water molecules and single-molecule entropy estimation. Biophysica. 2021;1:279–96.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fogolari F, Borelli R, Dovier A, Esposito G. The *k*th nearest neighbor method for estimation of entropy changes from molecular ensembles. WIREs Comput Mol Sci. 2023. e1691. https://doi.org/10.1002/wcms.1691