ABSTRACT

Title of dissertation: CONFORMATIONAL SAMPLING AND CALCULATION OF MOLECULAR FREE ENERGIES USING SUPERPOSITION APPROXIMATIONS

Sandeep Somani, Doctor of Philosophy, 2011

Dissertation directed by: Professor Michael K Gilson
University of Maryland College Park

The superposition approximations (SAs), first proposed in the distribution function theories of liquids, are a family of approximations to a multivariate probability distribution function (pdf) in terms of its lower order marginal pdfs. In this talk, we first present the relationship between various forms of SA, the measurement of correlation via mutual information, and approximations to the entropy of the full pdf via truncations of the Mutual Information Expansion.

Next, based on the SAs, a novel framework to construct computationally tractable approximations to the N-dimensional Boltzmann conformational distribution of molecule in terms of its low order marginal pdfs is presented. The marginal pdfs are obtained as normalized histograms of internal coordinates of a set of Boltzmann distributed conformations obtained by molecular dynamics (MD) simulation. We evaluate the accuracy of these approximate distributions constructed from marginal pdfs of order $l \leq 3$ for small molecules (≤ 52 atoms) by using a novel conformational sampling algorithm to sample from them and comparing the samples with the original MD conformations used to populate the pdfs. We find that the triplet ($l = 3$) level approximation has high conformational overlap with the physical Boltzmann distribution, and significantly better than that for

the singlet ($l = 1$) or doublet ($l = 2$) level approximations. The results shed light on the relative importance of correlations of different orders.

The singlet ($l = 1$) and doublet ($l = 2$) level approximate distributions are then used to define reference systems with known free energies, and then to compute the physical free energy of molecules using the reference system approach. Free energies are computed for small peptides as test molecules, and it is found that the convergence of the free energy estimate using a doublet reference is dramatically faster than with the singlet reference, consistent with greater overlap of the doublet reference system with the physical system. Potential further developments and practical applications are discussed.

CONFORMATIONAL SAMPLING AND CALCULATION OF MOLECULAR FREE ENERGY USING SUPERPOSITION APPROXIMATIONS

by

Sandeep Somani

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Michael K. Gilson, Chair/Advisor
Professor John D. Weeks
Professor Christopher Jarzynski
Professor Gerhard Hummer
Professor Sergei Sukharev

Dedicated to the memory of

S. Mukherjee

(Nov 4, 1931 – Dec 10, 2006)

who led an enlightened life in a mundane world

Acknowledgments

I owe my deepest gratitude to my advisor, Dr Mike Gilson for his patient tutoring and mentoring, and for suggesting an intellectually stimulating research project. He embodies the scientific spirit with a curious and a critical mind. I continue to learn from him and look forward to many more lessons in the future.

My department, faculty and peers at College Park have been a tremendous source of support and inspiration. I thank my committee members for their interest in my research. I am specially grateful to Dr Christopher Jarzynski for his guidance on the free energy calculation aspects of this work. Dr Prakash Narayan helped me with the information theory aspects. Dr Michael Coplan's continuous encouragement and keen interest in my academic progress over the last five years are also sincerely appreciated.

I thank my classmates, Suriyanarayanan Vaikuntanathan and Andy Ballard, for many discussions and critical comments on my research; and colleagues at CARB, Ravi Aduri and Swarna Pidugu, for sharing an experimentalist's perspective and their help with proof reading this thesis.

I am grateful for the financial assistance from the Graduate school in the form of Ann G. Wylie Dissertation Fellowship and Jacob K. Goldhaber Travel Award. I would also like to thank Debbie Jenkins for promptly taking care of all administrative formalities.

I am indebted to my prior supervisors at Bioinformatics Institute, Singapore – Dr Pawan Dhar and Dr Chandra Verma – for their mentoring and for initiating me into biology, and particularly for introducing me to the exciting field of modeling and simulation of biological systems. Research under their supervision had helped me crystallize my research interests.

I thank Tarun Pruthi and Sharmistha Chakraborty for their friendship, for being

great hosts on numerous occasions and for helping me to adjust to life in the US. Friends from the DESI student group at College Park provided the occasional escape from work.

I am grateful for the wishes of my parents-in-law and elders. I owe my modest accomplishments to the support and blessings of my parents.

Finally, my decision to move to the US for doing a PhD was a disruptive one, for not just me, but also my wife, Anu. She traded a comfortable life and a successful career in Singapore for the uncertainties and anxieties that inevitably accompanies a change of this magnitude – but, all with a smile! Without her, this journey would have hardly been possible.

Though challenging at times, the past five years in graduate school have been most rewarding and enriching, an experience that I will cherish forever.

Table of Contents

List of Tables

List of Figures

List of Definitions of Selected Symbols

Chapter 1

$\boldsymbol{\xi}$	Vector of internal coordinates in continuous conformational space
U_P	Physical potential energy function
p_N^B	Boltzmann conformational distribution of a molecule with N internal coordinates
$J(\boldsymbol{\xi})$	Jacobian
k_B	Boltzmann constant
T	Temperature
β	Inverse temperature ($(k_BT)^{-1}$)

Chapter 2

$X, Y, Z, X_1, X_2, ..$	Discrete valued random variables
$x, y, z, x_1, x_2, ..$	Specific values of random variables
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Sets of discrete random variables
Ω_X	State-space of X
$p(X)$	Probability distribution function (pdf) of X
$p(X\|y)$	Conditional pdf of X given $Y = y$
$\perp$	Independent of
$S(X)$	Shannon entropy of X
$S(X\|y)$	Conditional entropy of X given $Y = y$
$S(X\|Y)$	Average of $S(X\|y)$ for all possible Y
$D(p\|\|q)$	Kullback-Leibler divergence or distance between pdfs p and q
$I(X;Y), I_2$	Pairwise or second order mutual information between X and Y
$I(X;Y\|z)$	Conditional mutual information between X and Y given $Z = z$
$I(X;Y\|Z)$	Average of $I(X;Y\|z)$ for all possible Z
$I(X;Y;Z), I_3$	Third order mutual information
p_N	N-dimensional pdf
I_N	Mutual information among N random variables
S_N	Entropy of an N-dimensional pdf
$S_N^{(l)}$	l-level approximation to S_N
$S_N^{(m)}$	Approximation to S_N with mutual informations

	of different orders; hence "m" for "mixed"
$p_3^{(2)}$	Kirkwood superposition approximation
$p_4^{(3)}$	Fisher-Kopeliovich superposition approximation
$p_N^{(N-1)}$	Generalized Kirkwood superposition approximation
$p_N^{(l)}$	l-level superposition approximation
$p_N^{(m)}$	Mixed superposition approximation
$\mathscr{P}_{(N,j)}$	Product of all j-order marginal pdfs of an N-dimensional pdf
$a(j;N,l)$	Exponent of $\mathscr{P}_{(N,j)}$ in the SA-l approximation
	of an N-dimensional pdf
B	Number of discrete states of a random variable;

Chapter 3

$\tilde{p}_N^{(l)}$	Distribution sampled by the l-level sampling algorithm
$\mathbf{X}$	Vector of internal coordinates in discretized conformational space
M	Number of atoms in a molecule
N	Number of internal coordinates of a M-atom molecule ($= 3M - 6$)
$\xi_{i,\min}$	Minimum value of coordinate ξ_i observed in a MD simulation
$\xi_{i,\max}$	Maximum value of coordinate ξ_i observed in a MD simulation
Δ_i	Range of coordinate ξ_i observed in a MD simulation
δ_i	Bin-width used for discretizing ξ_i
Ω	Set of all conformations in the discretized conformational space
Ω_P	Set of conformations sampled from p_N or p_N^B
	used to populate the reference distributions
$\Omega^{(l)}$	Set of conformations accessible to the l-level sampling algorithm
B	Number of bins used to discretize a continuous coordinate

Chapter 4

N_R	Number of samples from reference distribution
N_P	Number of samples from physical (Boltzmann) distribution
$\bar{U}_P$	Effective potential energy in the discrete space

$U_R^{(l)}$	Reference potential energy function of the l-level reference
$\Delta U^{(l)}$	Energy difference $\bar{U}_P - U_R^{(l)}$
Γ_P	Continuous conformational space
Z_P	Configurational integral over Γ_P
$\bar{Z}_P$	Discrete space approximation to Z_P
$\bar{Z}_P^{(l)}$	Asymptotic limit of perturbation estimate using the l-level reference
$\hat{Z}_P^{(l)}$	Numerical estimate of $\bar{Z}_P^{(l)}$
$F_P, \bar{F}_P, \bar{F}_P^{(l)}, \hat{F}_P^{(l)}$	Free energies corresponding to $Z_P, \bar{Z}_P, \bar{Z}_P^{(l)}, \hat{Z}_P^{(l)}$, respectively
$Z_R^{(l)}$	Partition function of the l-level reference system
$\langle f \rangle$	Boltzmann average of a function of internal coordinates
$\langle \bar{f} \rangle$	Discrete space approximation to $\langle f \rangle$

Chapter 5

F_P^L, F_P^R, F_P^{LR}	Absolute free energy of ligand, receptor and complex
ΔF_{bind}	Binding free energy
$\tilde{S}_P^{(l)}$	Cross entropy of $\tilde{p}_N^{(l)}$ with respect to the Boltzmann distribution

List of Abbreviations

pdf	probability distribution function
BBGYK	Bogoliubov-Born-Green-Yvon-Kirkwood hierarchy
M2	Mining Minima
MD	Molecular Dynamics
MC	Monte Carlo
KL	Kullback-Leibler distance or divergence
MIE	Mutual Information Expansion
SA	Superposition Approximation
KSA	Kirkwood SA
FKSA	Fisher-Kopeliovich SA
GKSA	Generalized Kirkwood SA
SA-l	l-level SA
XYZ	Anchored cartesian internal coordinate system
BAT	Bond-Angle-Torsion internal coordinate system

Chapter 1

Introduction

1.1 Background

Atomistic modeling and simulation of molecules represents a powerful tool for understanding chemical and biological systems at the microscopic level. Computational approaches are indispensable for studying many complex real world systems where analytic and experimental analysis are difficult or impossible. Indeed, computational tools are widely used in industrial applications for applications ranging from material design to pharmaceutical drug design [1].

Some microscopic properties of molecules, such as covalent bonded structure and quantum energy levels, can be obtained from first principles using quantum mechanical calculations [2, 3]. In biological systems, quantum chemistry calculations have been widely used to study various phenomena involving changes in the electronic properties of the molecules, such as photosynthesis and the reaction mechanisms of enzymes [4, 5, 6]. Of particular practical importance is the application of quantum chemistry methods in computer aided drug design to compute properties of flexible biomolecules in a thermal solvent environment [7, 8]. However such calculations are currently challenging, due

to the combined cost of modeling the details of electronic structure and of accounting for conformational flexibility; i.e., nuclear motions [7]. In situations where the details of the electronic structure are not modified (e.g. no covalent bond modifications), a full quantum mechanical energy treatment may be replaced with a less expensive approximation via an empirical force-field, and nuclear motions may be treated classically, in accordance with the Born-Oppenheimer approximation. An empirical force-field, also termed a molecular mechanics force-field, specifies the potential energy of the molecule as a function of its configuration and is typically parameterized against quantum chemistry calculations and experimental data. The development of accurate and efficient force-fields [9, 10, 11, 12, 13, 14, 15], and hybrid QM/MM methods where only a small subset of atoms of the molecule are treated quantum mechanically, are active fields of research [4, 6].

Statistical thermodynamic quantities can be computed using molecular dynamics (MD) or Monte Carlo (MC) simulations with empirical force-fields. Such calculations were first applied to simple liquids [16], and soon after, to proteins [17]. With advances in simulation techniques and computing technology, the field of biomolecular modeling and simulations has since progressed tremendously and has shed light on diverse phenomena such as protein folding [18], non-covalent binding [19], molecular recognition [20] and allostery [21, 22]. Central challenges in biomolecular simulations are sampling the physically relevant regions of the high dimensional conformational space, and calculating free energies. This thesis develops a novel conformational sampling method based on approximation to the full Boltzmann distribution, and describes how this sampling method can be used to compute free energies. Potential applications and enhancements are also described.

1.2 Motivation

In a thermal environment, three-dimensional structure, or conformation, of a molecule fluctuates continuously. These fluctuations are key determinants of thermodynamic properties and are intimately connected with biomolecular functions, such as binding . The equilibrium fluctuations of a molecule with M atoms in contact with a heat bath at temperature T are captured by the Boltzmann probability distribution function (pdf) over its $N = 3M - 6$ internal degrees of freedom, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$:

$$p_N^B(\boldsymbol{\xi}) = \frac{1}{Z_P} \exp\left(-\beta\, U_{\mathrm{P}}(\boldsymbol{\xi})\right) J(\boldsymbol{\xi}) \tag{1.1}$$

where

$$Z_P = \int \exp\left(-\beta\, U_P(\boldsymbol{\xi})\right) J(\boldsymbol{\xi})\, d\boldsymbol{\xi}\ , \tag{1.2}$$

$\beta = 1/(k_B\, T)$, k_B being the Boltzmann constant; Z_P is termed the configurational integral; $J(\boldsymbol{\xi})$ is the Jacobian of the transformation from internal coordinates to Cartesian coordinates; and $U_P(\boldsymbol{\xi})$ is the energy as a function of conformation. (This expression omits a prefactor of the configuration integral that results from integration over the momentum degrees of freedom and which cancels upon taking a free energy difference [23].) For a molecule in solution, the energy function comprises the molecule's potential energy and a contribution from the solvent [24]. The Boltzmann distribution links conformational fluctuations to thermodynamic observables, such as free energy and entropy, but its high dimensionality and the potentially complex multi-particle energetic couplings make it computationally unworkable. It is therefore of interest to construct computationally tractable approximations to the Boltzmann distribution, perhaps by limiting the complexity of the correlations among internal coordinates that are accounted for in the approximate pdf.

It is not clear *a priori* how well a pdf with a reduced accounting of correlations could approximate a full Boltzmann pdf. However, some indication of what might be possible can be derived from studies of the Gibbs entropy, because this quantity, broadly speaking, depends upon the degree of correlation. This general statement can be made more specific by the aid of concepts drawn from the fields of probability and information theory, which have useful mathematical connections with the statistical description of microscopic system in terms of ensembles and probability distributions of microstates. Indeed, the mathematical form of the Gibbs entropy is identical to the Shannon information entropy of a general multivariate stochastic system. Concepts and tools from information theory can, therefore, be usefully applied in the statistical thermodynamical treatment of physical systems. One such concept, which is of particular interest here, is that of mutual information as a quantitative measure of correlations among multiple stochastic variables. Further, using the mutual information expansion (MIE), the entropy of a multivariate distribution can be written in terms of mutual information contributions corresponding to correlations of increasing orders. The MIE thus provides a tool to dissect the contributions of different orders of correlations to the fluctuations of the full system.

Similar ideas have been developed in statistical mechanics theories of liquids based on distribution functions of liquid particles. In particular, the MIE is closely related to the entropy expansion of a liquid in terms of distribution functions of increasing number of liquid particles. Furthermore, the mathematical form of the MIE can be obtained from superposition approximations (SA) like those which have been used to approximate the distribution functions of liquids. However, the MIE can be applied not just to liquids, but to any pdf, including one not explicitly connected to any physical system.

This thesis project was motivated by the prior observation that the entropy

associated with the conformational fluctuations, or the configurational entropy, of a molecule computed using the MIE was dominated by the low-order mutual informations among the internal coordinates of the molecule. This observation led to our initial hypothesis that the conformational fluctuations of a molecule might be dominated by low-order correlations, and that the SA family of approximations could be useful in developing computationally tractable approximations to the Boltzmann distribution. Exploration of this concept led to a novel approximation of the Boltzmann pdf, which we then applied to the calculation of molecular free energies.

1.3 Overview of thesis and contributions

In Chapter 2 we present an information theory based view of correlations and the MIE of the entropy of a multivariate pdf. We show that approximation of the entropy with low-order terms in the MIE is directly related to a superposition approximation of the distribution in terms of its marginal pdfs.

Chapters 3 and 4 present the main contributions of this work. In Chapter 3, the SA framework is used to develop novel conformational sampling algorithms to sample molecular conformations from the high dimensional conformational space using pdfs of up to order three. We show that the overlap between the distribution thus sampled, and the physical Boltzmann distribution improves on incorporating more correlations. The sampling distributions represent a computationally tractable and a normalized approximation to the Boltzmann distribution.

Due to the normalization property, reference systems with known free energy can be set up in terms of the sampling distributions and the free energy of a physical system of interest can then be obtained as a free energy difference. This approach is used in Chapter

4 to compute the absolute free energy, or the configurational integral, of a molecule using samples drawn from the aforementioned conformational sampling algorithms. We show that the convergence of the estimated free energy dramatically improves upon using a reference which includes pairwise correlations among the internal coordinates via the SA framework.

Chapter 5 presents potential extensions and applications of the ideas and methods developed in the previous chapters.

Chapter 2

Superposition Approximations

2.1 Introduction

Superposition approximations (SAs) are a family of approximations which express a multivariate probability distribution in terms of its marginal distributions of subsets of the variables. They provide a tool to model correlations in a system with many degrees of freedom. Superposition approximations have a long history in statistical mechanical theory of liquids and in information theory of communication.

Superposition approximations were first proposed in the distribution function theories of liquid by Kirkwood and Boggs [25] where the Kirkwood superposition approximation (KSA) was used as a closure equation for truncating the Bogoliubov-Born-Green-Yvon-Kirkwood (BBGYK) [26] hierarchy at the doublet, or two-particle, level enabling calculation of the pair correlation function, or the radial distribution function, $g(r)$. The KSA was derived by approximating the three-particle potential of mean force as the sum of the three two-particle potential of mean forces, which is equivalent to approximating the three-particle joint distribution function as the product of the three two-particle distributions. In practice, based on empirical comparison with other theoretical methods and experiments, the KSA closure is accurate only at low densities, for which

the three-particle and higher order correlations are weak. Fisher and Kopeliovich [27] improved upon the KSA by approximating the four-particle distribution in terms of lower-order distribution functions and closing the BBGYK hierarchy at the triplet or three particle level.

The SAs in liquid theory can be derived based on considerations of permutation symmetry of particle labels and correct asymptotic limits. Using a variational principle for the free energy of a liquid, Reiss [28] showed that, assuming a functional form as the product of the marginal distributions, the KSA and FKSA were the optimal closures at the doublet and triplet level, respectively, and also extended it to higher levels. Bugaenko *et. al.*, [29] and more recently, Singer [30], derived the superposition approximation closures for correlation functions in liquids based on the maximum entropy principle. The generalization of the KSA to express an N-dimensional distribution function in terms of its marginal pdfs of up to order $N-1$ is called the generalized Kirkwood superposition approximation (GKSA) [31, 32, 30]. Based on the GKSA, a series expansion to the entropy of a liquid can be derived, where the kth term is a function of the joint distribution function of k particles [33, 34]. The entropy expansion has been used to assess the contributions of higher order correlations to the entropy of the liquid [35].

Remarkably similar concepts and mathematical forms arise in the field of information theory. The Shannon information entropy, which is a measure of uncertainty in a stochastic system, is closely related to the Gibbs entropy in statistical thermodynamics, which measures disorder in a many-particle system. The connection between information and correlation is central to communication theory, which deals with problems of estimating the amount of information transferred given correlations among multiple stochastic inputs and outputs. The earliest attempt to address such a problem using information theory

is apparently due to McGill, who defined "transmission information" as a measure of correlation [36]. Fano subsequently identified the transmission information with mutual information and presented the generalized mutual information as a measure of the correlations among N (> 2) variables [37].

An important result of the information theory is the Mutual Information Expansion (MIE) of the entropy of an N-dimensional probability distribution function in terms of mutual informations among different subsets of variables [37, 38]. The MIE is analogous to the entropy expansions in liquid theory, except that they are applicable to general discrete valued distributions of a heterogeneous system. The MIE of entropy has been applied to problems beyond communication theory, such as signal processing [31], configurational entropy calculation of molecules [39] and general complex systems such as frustrated spin systems [32].

The main goal of this chapter is to motivate mutual information as a measure of correlations and establish its relationship with the superposition approximation. This chapter begins by presenting the basics of discrete probability theory. Attention is restricted to discrete distributions, since we are interested in modeling conformational distributions of molecules in a discretized internal coordinate space. Also information theoretic concepts are easier to present in terms of probability distributions of discrete-valued variables. Next, relevant concepts from information theory are presented and the MIE is discussed. A series of approximations to the entropy of a multivariate probability distribution obtained by truncating the MIE at different levels is discussed. The connections between various superposition approximations, mutual information and entropy expansions are discussed.

2.2 Basic concepts of discrete probability

Consider a discrete-valued random variable X, which assumes values from a finite set $\Omega_X = \{x_1, x_2, ..., x_{nx}\}$ with nx elements. We imagine multiple measurements or observations of X, each of which returns a value x_i from the state-space Ω_X. With each x_i, we assign a real number $p(x_i) \in [0, 1]$ as the frequency of observing x_i in the limit of infinite observation. A probability distribution function, $p(X)$, is a vector of probability values corresponding to the elements of the state space Ω_X:

$$\begin{aligned} \Omega_X &= \{x_1, x_2, ..., x_{nx}\} \\ p(X) &= \{p(x_1), p(x_2), ..., p(x_{nx})\} \ . \end{aligned} \tag{2.1}$$

In this work, we subscribe to the above *frequentist* interpretation of probability, though the concepts presented here should be compatible with the interpretation where probability indicates the *degree of belief* in a particular outcome. (For a discussion of the frequentist versus Bayesian views of probability see Jaynes [40].) Since the probability values measure the fractional occurrence of all possible outcomes, the sum of the probability values over all possible outcomes is one,

$$\sum_{x \in \Omega_X} p(x) = 1 \tag{2.2}$$

and the pdf is said to be normalized. Observation of a random variable is sometimes referred to as the drawing of a sample from its pdf. Any set of non-negative real numbers can be thought of as a pdf with its normalization given by the sum of all numbers.

Before proceeding further, a comment on the notation is in order. We use upper-case letters for labels of random variables, and lower-case to denote specific values from the state-space. Also, for brevity, $p(X)$ is used for the pdf of X, instead of the more accurate notation of $p_X(X)$. Thus, $p(X)$ is a different pdf than, say, the pdf $p(Y)$ of variable Y

defined over a state-space Ω_Y with ny states.

Next, considering simultaneous measurement of the two random variables, X and Y, the joint probability, $p(x, y)$, is the frequency of observations with $X = x$ and $Y = y$. The joint probability distribution, $p(X, Y)$, is the set of probability values for all possible combinations of the two variables, that is, $\Omega_X \times \Omega_Y$, and can be viewed as a matrix

$$p(X,Y) = \begin{Bmatrix} p(x_1,y_1) & \cdots & p(x_1,y_{ny}) \\ \vdots & \ddots & \vdots \\ p(x_{nx},y_1) & \cdots & p(x_{nx},y_{ny}) \end{Bmatrix} \tag{2.3}$$

with each row denoting a specific value of X and column denoting a specific value of Y. The joint distribution function is also normalized, so that,

$$\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y) = 1. \tag{2.4}$$

Given the joint distribution, summing over one of the variables gives the marginal probability distribution function, or the marginal, of the other variable. Thus, the marginals of X and Y are, respectively,

$$\begin{aligned} p(X) &= \sum_{y\in\Omega_Y} p(x,y) \\ \text{and, } p(Y) &= \sum_{x\in\Omega_X} p(x,y)\,. \end{aligned} \tag{2.5}$$

In effect, the marginal for X is obtained by summing the rows in Eq. 2.3, and for Y by summing the columns.

2.2.1 Conditional probability distributions and independence

The probability distribution of X conditional on knowledge that $Y = y$ is given by the product rule [40]

$$p(X|y) = \frac{p(X,y)}{p(y)}\,,\ p(y) \neq 0 \tag{2.6}$$

and, similarly, the conditional distribution of Y given $X = x$ is

$$p(Y|x) \;=\; \frac{p(x,Y)}{p(x)}\,. \tag{2.7}$$

In terms of the matrix representation of the joint distribution, $p(X,y)$ corresponds to the $Y = y$ column, and $p(x,Y)$ corresponds to the $X = x$ row. The normalization of the conditional pdfs can be seen as follows

$$\begin{aligned}\sum_{y\in\Omega_Y} p(y|x) &= \sum_{y\in\Omega_Y} \frac{p(x,y)}{p(x)} \\ &= \frac{1}{p(x)} \sum_{y\in\Omega_Y} p(x,y) \\ &= \frac{1}{p(x)} p(x) \\ &= 1 \end{aligned} \tag{2.8}$$

and similarly for $p(X,y)$. Rearranging the product rule yields the chain rule of conditional distributions for two variables:

$$p(x,y) = p(x)p(y|x) = p(y)p(x|y)\,. \tag{2.9}$$

Notice that $p(x)p(y|x)$ is the probability of picking the $X = x$ row of $p(X,Y)$, times the probability of picking $Y = y$ column in that row.

Variables X and Y are said to be independent if their joint distribution equals the product of their marginal distributions; otherwise they must be correlated. Denoting independence by "$\perp$", we have

$$X \perp Y \Rightarrow p(x,y) = p(x)p(y) \tag{2.10}$$

and from Eqs. 2.6 and 2.7

$$\begin{aligned} X \perp Y \Rightarrow p(X|y) &= p(X) \quad \text{, and} \\ p(Y|x) &= p(Y)\,. \end{aligned} \tag{2.11}$$

Note that, if X and Y are perfectly correlated, for example, if they are related as $x = y$, then the joint distribution is a diagonal matrix so that $p(X|y)$ will be unity for one of X and zero for the others; and similarly for $p(Y|x)$.

2.2.2 Generalization to N random variables

The above definitions can be generalized in a straightforward manner to N discrete-valued random variables , each of which assumes a finite number of distinct values. The probability that variables $X_1, \ldots, X_N$ take values $x_1, \ldots, x_N$, respectively, is denoted by

$$p_N(x_1, x_2, \ldots, x_N) \equiv p_N(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) \tag{2.12}$$

where p_N denotes the joint pdf of all variables. Marginal pdf, $p_k(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$, of a subset of variables where $i_1, \ldots, i_k \in \{1, \ldots, N\}$, is obtained by summing over the other $N - k$ variables:

$$p_k(X_{i_1}, \ldots, X_{i_k}) = \sum_{X_{i_{k+1}}} \cdots \sum_{X_{i_N}} p_N(x_{i_1}, \ldots, x_{i_N}) \tag{2.13}$$

where $\sum_{X_{i_j}}$ denotes the sum over all possible values of X_{i_j}. The subscript in p_k, which denotes the dimensionality, or the order, of the pdf, will sometimes be dropped if the arguments of the pdf are explicitly listed. At order k, there are C_k^N different marginal pdfs of p_N, each corresponding to a unique combination of k variables. The marginal pdfs at orders $k = 1, 2$ and 3 are termed singlet, doublet and triplet pdfs, respectively. All marginal distributions and p_N are non-negative and normalized.

As in the case of two variables, independence of two sets of variables $\mathbf{X} = \{X_1, \ldots, X_{N_x}\}$ and $\mathbf{Y} = \{Y_1, \ldots, Y_{N_y}\}$ implies that their joint distribution is given by the product distribution

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \ . \tag{2.14}$$

The conditional distribution of $\mathbf{X}$ given $\mathbf{y}$ is

$$p(\mathbf{X}|\mathbf{y}) = \frac{p(\mathbf{X},\mathbf{y})}{p(\mathbf{y})} \tag{2.15}$$

and the chain rule for the joint distribution between, say, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ is

$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x})\ p(\mathbf{y}|\mathbf{x})\ p(\mathbf{z}|\mathbf{x},\mathbf{y})\,. \tag{2.16}$$

Also, in the case of more than two variables, we have the notion of conditional independence where two sets of variables, otherwise correlated, become independent given knowledge of a third set; each set may contain one or more variables. If $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on knowledge of $\mathbf{Z}$, then we have the following relations:

$$\begin{aligned} \mathbf{X}\perp\mathbf{Y}\mid\mathbf{Z}\ \Rightarrow\quad p(\mathbf{x},\mathbf{y}|\mathbf{z}) &= p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}) \\ p(\mathbf{x},\mathbf{y},\mathbf{z}) &= p(\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}) \\ p(\mathbf{x}|\mathbf{y},\mathbf{z}) &= p(\mathbf{x}|\mathbf{z})\,. \end{aligned} \tag{2.17}$$

Note that, since $\mathbf{X}$ and $\mathbf{Y}$ are correlated, $p(\mathbf{x}|\mathbf{y}) \neq p(\mathbf{x})$.

2.3 Information theory view of correlations and entropy

The central quantity in information theory is the Shannon entropy [41], which measures the uncertainty in a random variable. The Shannon entropy, or simply entropy, $S(X)$, of a discrete-valued random variable X with pdf $p(X)$ is given by

$$S(X) \equiv -\sum_{x\in\Omega_X} p(x)\ \ln p(x) \tag{2.18}$$

where $p(X)$ is the pdf of X, and the units are set by the base of the logarithm, and $0\ln 0 \equiv 0$. The entropy of X gives the amount of information required to specify its value

given its probability distribution. Consider the following simple examples that illustrate this property and the sense in which entropy is a measure of uncertainty.

Let the logarithm in Eq. 2.18 be in base 2 thereby setting the units to bits. Consider the case where X is deterministic, in the sense that X takes only a single value x_i, so that $p(X = x_i) = 1$ and $p(X \neq x_i) = 0$. For this case, the entropy is zero, consistent with the fact that, given $p(X)$, no further information is required to know the value of X. Next, suppose X can take two values, x_1 and x_2, with equal probabilities: $p(x_1) = p(x_2) = 1/2$, and the probability is zero for other values. In this case, the entropy is 1 bit, as would be expected, since a single bit of information suffices to specify whether or not $X = x_1$, simultaneously specifying whether or not $X = x_2$. The increased uncertainty in X on adding another possible value was reflected in the increase of entropy from 0 to 1 bit. The uncertainty in X is maximum when all values of X are equally likely, that is, $p(x) = 1/nx$, where nx is the size of the state-space of X. In this case, the entropy takes its maximum value of $\ln nx$. The bounds on entropy are

$$0 \leq S(X) \leq \ln nx \tag{2.19}$$

where the lower bound corresponds to the deterministic limit, and the upper bound to the maximum uncertainty limit. For a multivariate system the entropy is given by

$$S(\mathbf{X}) = -\sum_{\mathbf{X}} p(\mathbf{x}) \ln p(\mathbf{x}) \tag{2.20}$$

where $p(\mathbf{X})$ is the joint distribution of the system. Also, the entropy of two independent sets of variables is additive,

$$S(\mathbf{X}, \mathbf{Y}) = S(\mathbf{X}) + S(\mathbf{Y}) \tag{2.21}$$

since $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ for independent variables.

2.3.1 Conditional entropy

Suppose we have another random variable Y, in addition to X. We ask the question: what is the uncertainty in X given that $Y = y$? This is given by the entropy of the conditional distribution $p(X|y)$:

$$S(X|y) = -\sum_{x\in\Omega_X} p(x|y)\ln p(x|y). \tag{2.22}$$

We define conditional entropy, $S(X|Y)$, as the *average* of the uncertainty in X given y, for all possible values of Y. The conditional entropy is obtained by taking an average of $S(X|y)$ with respect to the probability distribution of Y:

$$S(X|Y) \equiv \sum_{y\in\Omega_Y} p(y)\, S(X|y)\,. \tag{2.23}$$

Using Eq. 2.22 we get

$$\begin{aligned}
S(X|Y) &= \sum_{y\in\Omega_Y} p(y)\left(-\sum_{x\in\Omega_X} p(x|y)\ln p(x|y)\right)\\
&= -\sum_{y\in\Omega_Y}\sum_{x\in\Omega_X} p(y)\,p(x|y)\ln p(x|y)\\
&= -\sum_{y\in\Omega_Y}\sum_{x\in\Omega_X} p(x,y)\ln\frac{p(x,y)}{p(y)}\\
&= -\sum_{y\in\Omega_Y}\sum_{x\in\Omega_X} p(x,y)\ln p(x,y) - \left(-\sum_{y\in\Omega_Y}\sum_{x\in\Omega_X} p(x,y)\ln p(y)\right)\\
&= S(X,Y) - S(Y)
\end{aligned} \tag{2.24}$$

where, in the last step, we summed over X to get $S(Y)$. Thus, the average uncertainty in X conditioned on knowledge of Y is expressed as the following conditional entropy:

$$S(X|Y) = S(X,Y) - S(Y)\,. \tag{2.25}$$

Notice that, if X and Y are independent then, we have

$$\begin{aligned}
S(X|Y) &= S(X)\\
\text{and } S(Y|X) &= S(Y)
\end{aligned} \tag{2.26}$$

which is consistent with the intuitive expectation that if the two variables are independent, then knowledge of one does not give any information on the other. In the other limit, if X is perfectly correlated with Y so that specifying Y determines X, we expect the conditional entropy to be zero. This can be seen as follows. In case of perfect correlation, the matrix of the joint distribution, $p(X,Y)$, is diagonal, so the conditional distribution $p(X|y)$ is zero for each value of y (*i.e.*, for each column) for all x except one. As a result, from Eq. 2.22, $S(X|y) = 0$ giving $S(X|Y) = 0$. Thus, the bounds on conditional entropy are

$$\begin{aligned} 0 \leq S(X|Y) \leq S(X) \\ 0 \leq S(Y|X) \leq S(Y) \end{aligned} \tag{2.27}$$

where the lower bound corresponds to perfect correlation and upper bound to independence of the two variables. Note that conditioning can only reduce the entropy of or the uncertainty in a variable.

Based on the conditional independence relations from Eq. 2.17, and additivity of entropy, if X and Y are conditionally independent on knowledge of Z, we have

$$S(X,Y|Z) = S(X|Z) + S(Y|Z)\,. \tag{2.28}$$

An important relation based on the conditional entropy is the chain rule for the entropy of a multivariate system

$$S(X_1,...,X_N) = S(X_1) + S(X_2|X_1) + S(X_3|X_1,X_2) + \ldots + S(X_N|X_1,...,X_{N-1})\,. \tag{2.29}$$

Using the fact that conditioning reduces entropy, *i.e.* $S(X_i|X_1,\ldots,X_{i-1}) \leq S(X_i)$, we get the following inequality for the entropy of the full system

$$S(X_1,...,X_N) \leq S(X_1) + S(X_2) + \ldots + S(X_N) \tag{2.30}$$

and the equality is obtained when all variables are independent. In other words, correlations reduce the entropy of a system. The above relations can be extended to sets of multiple variables, e.g. $S(\mathbf{X}|\mathbf{Y}) = S(\mathbf{X},\mathbf{Y}) - S(\mathbf{Y})$, etc.

2.3.2 Kullback-Leibler distance

The Kullback-Leibler (KL) distance, also known as relative entropy or KL divergence, between two normalized distributions, $p(\mathbf{X})$ and $q(\mathbf{X})$, that are defined on the same state space, is given by [42]

$$D(p||q) \equiv \sum_{\mathbf{X}} p(\mathbf{X}) \ln \frac{p(\mathbf{X})}{q(\mathbf{X})} \ . \tag{2.31}$$

$D(p||q)$ is a measure of the deviation between the two distributions. The KL distance is zero iff the two distributions are equal, and is not symmetric, that is, $D(p||q) \neq D(q||p)$. An important property of KL distance is that it is non-negative [43], so that,

$$\begin{aligned} D(p||q) &\geq 0 \\ \Rightarrow \sum_{\mathbf{X}} p(\mathbf{X}) \ln p(\mathbf{X}) &\geq \sum_{\mathbf{X}} p(\mathbf{X}) \ln q(\mathbf{X}) \\ \Rightarrow S(\mathbf{X}) &\leq -\sum_{\mathbf{X}} p(\mathbf{X}) \ln q(\mathbf{X}) \ . \end{aligned} \tag{2.32}$$

The expression on the left-hand side is the entropy of p, and the expression on the right-hand side is termed the cross-entropy of q with respect to p. The inequality implies that the cross-entropy of any approximation to p will be greater than the entropy of p. Note that the above inequality for normalized distributions is a special case of a more general statement. By using the log-sum inequality [43], it can be shown that, the inequality holds as long as normalization factor of p is greater than that of q.

2.3.3 Mutual information between two variables

Using the concept of conditional entropy discussed in Section 2.2.1, we can answer the question: How does the uncertainty in X change when one is provided with knowledge of Y? We have seen that the uncertainty in X is given by the entropy $S(X)$, while the average uncertainty in X given knowledge of Y is the conditional entropy $S(X|Y)$. We thus define the reduction in uncertainty of X due to learning the value of Y as

$$I(X;Y) \equiv S(X) - S(X|Y)\,. \tag{2.33}$$

Since the reduction in uncertainty is equivalent to the gain in information, $I(X;Y)$ can also be viewed as the gain in information about X due to knowledge of Y. Using Eq. 2.25,

$$I(X;Y) = S(X) + S(Y) - S(X,Y) \tag{2.34}$$

Similarly, the gain of information about Y due to knowledge of X is

$$\begin{aligned} I(Y;X) &\equiv S(X) - S(X|Y) \\ &= S(X) + S(Y) - S(X,Y) \end{aligned} \tag{2.35}$$

where Eq. 2.25 is used in the second step. $I(X;Y)$ and $I(Y;X)$ are termed the mutual information between X and Y. It is evident that the mutual information is symmetric in the two variables, implying that the gain in information about X due to knowledge of Y is same as the gain in information about Y due to knowledge of X. In terms of probability distributions, we have

$$I(X;Y) = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\ \ln\frac{p(x,y)}{p(x)p(y)} \tag{2.36}$$

which is same as the KL distance between the joint and the product distributions. Therefore, the mutual information is zero iff the two variables are independent and is

positive otherwise. If the two variables are perfectly correlated, we get $I(X;Y) = S(X)$, implying that the uncertainty in X is completely removed on knowledge of Y, as would be expected. Note that if X and Y are perfectly correlated then $S(X) = S(Y)$. The bounds on mutual information are

$$0 \leq I(X;Y) \leq \min\{S(X), S(Y)\} \tag{2.37}$$

where the lower bound corresponds to independence and upper bound to perfect correlation. Based on these properties, the mutual information has been used as a measure of correlation between the two variables. It is always non-negative, and greater mutual information indicates higher correlation. Note that mutual information can measure non-linear correlations as well, unlike linear measures, such as the Pearson's correlation coefficient. As a side note, since $S(X|X) = 0$, from Eq. 2.34, $I(X,X) = S(X)$; based on this property, entropy is sometimes also referred to as the self-information. Extending the above relations to two sets of multiple variables, we have,

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= S(\mathbf{X}) + S(\mathbf{Y}) - S(\mathbf{X},\mathbf{Y}) \\ &= \sum_{\mathbf{X}} \sum_{\mathbf{Y}} p(\mathbf{x},\mathbf{y}) \ \ln \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \,. \end{aligned} \tag{2.38}$$

2.3.4 Mutual information among three variables

We have seen that the gain in information for X on learning Y is the mutual information $I(X;Y)$, which is also a measure of correlation between X and Y. We next consider how the correlation between variables X and Y may change when the value of a third variable, Z, becomes known. By analogy with conditional entropy, discussed above, we define a mutual information between X and Y conditioned on knowledge that the value

of Z is z:

$$\begin{aligned} I(X;Y|z) &\equiv S(X|z) - S(X|Y,z) \\ &= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y|z)\ \ln\frac{p(x,y|z)}{p(x|z)p(y|z)}. \end{aligned} \tag{2.39}$$

Note that the expression in terms of distributions is similar to that for the unconditioned mutual information in Eq. 2.36, except that all distributions are now conditioned on z. Also, since it is the KL distance between two distributions, it is non-negative. Averaging $I(X;Y|z)$ over all possible values of Z gives the expression for the conditional mutual information among the three variables:

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in\Omega_Z} p(z) I(X;Y|z) \\ &= \sum_{z\in\Omega_Z} p(z) \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y|z)\ \ln\frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \sum_{z\in\Omega_Z}\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y,z)\ \ln\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)} \\ &= -S(X,Y,Z) - S(Z) + S(X,Z) + S(Y,Z)\,. \end{aligned} \tag{2.40}$$

In terms of conditional entropies, using Eq. 2.25, we have

$$I(X;Y|Z) = S(X|Z) + S(Y|Z) - S(X,Y|Z) \tag{2.41}$$

which is same as Eq. 2.34 for pairwise mutual informations with all entropies now conditioned on Z. Conditional mutual information is non-negative and it is zero iff X and Y are conditionally independent given Z, as can be seen from Eqs. 2.28 and 2.41. In another limit, if Z is independent of both X and Y, then from Eq. 2.26, the conditional mutual information reduces to the mutual information between X and Y, as expected, since Z does not provide any additional information on X or Y. Therefore, we have,

$$\begin{aligned} X \perp Y|Z &\Rightarrow I(X;Y|Z) = 0 \\ X \perp Z \,\text{and}\, Y \perp Z &\Rightarrow I(X;Y|Z) = I(X;Y)\,. \end{aligned} \tag{2.42}$$

Using the definition for the conditional mutual information, we now write the change in correlation between X and Y due to knowledge of Z as

$$I(X;Y;Z) \equiv I(X;Y) - I(X;Y|Z) \tag{2.43}$$

This expression may be usefully reformatted by using Eq. 2.34 and substituting Eq. 2.40 into Eq. 2.43 to yield

$$I(X;Y;Z) = S(X) + S(Y) + S(Z) - (S(X,Y) + S(X,Z) + S(Y,Z)) + S(X,Y,Z) \tag{2.44}$$

involving entropies of all marginals and the full three-dimensional distribution. Remarkably $I(X;Y;Z)$ is symmetric in the three variables, similar to the expression for the pairwise mutual information. Therefore,

$$\begin{aligned} I(X;Y;Z) &= I(X;Y) - I(X;Y|Z) \\ &= I(Y;Z) - I(Y;Z|X) \\ &= I(X;Z) - I(X;Z|Y) \end{aligned} \tag{2.45}$$

so that the $I(X;Y;Z)$ gives the change in correlation between any pair of variables due to the third variables. The quantity $I(X;Y;Z)$ is termed the mutual information at third order and measures the information shared by the three variables. It can be considered as a measure of correlation existing among the three variables above and beyond that captured by the pairwise mutual information. In particular, conditional independence of X and Y on knowledge of Z, cannot be inferred from the pairwise mutual informations. Also, there may be correlations at the triplet level, even if there are no correlations at the pairwise level. In contrast to the pairwise mutual information, the third order mutual information may be either positive or negative. In other words, knowledge of a third interacting variable

may increase or reduce the correlation between the other two variables. To illustrate this, perhaps unexpected result, in Section 2.3.4.1 we take a digression from the main text to present examples for a simple system with three binary random variable.

In physical systems, one can imagine, for example, that correlations in motions of two distant side chains in a protein may become independent on treating the protein backbone as rigid. In other words, side chain motions may become independent conditional on the backbone coordinates, so the third order mutual information among the coordinates of the two side chains and the backbone will be positive. As another example, Matsuda used analytically solvable model systems of 3-6 interacting spins to relate the phenomenon of frustration to the sign of the third and higher order mutual informations [32].

We list below the main relations of this section and a few additional useful identities:

$$\begin{aligned} I(X;Y|Z) &= I(X;Y,Z) \;-\; I(X;Z) \\ &= S(X|\,Z) - S(X\,|\,Y,\,Z) \\ &= S(X\,|\,Z) + \;\; S(Y\,|\,Z) - \; S(X,Y\,|\,Z) \\ I(X\,;Y\,;Z) &= I(\,X;Y\,) \;+\; I(\,Y;Z) \;-\; I(Y\,;\,X\,,\,Z) \end{aligned} \tag{2.46}$$

where $I(X;Y,Z)$ denotes the mutual information between distributions $p(X)$ and $p(Y,Z)$, $I(X;Y,Z) = S(X)+S(Y,Z)-S(X,Y,Z)$ and is different from $I(X;Y;Z)$. These relations are valid for three sets of multiple variables as well. Finally, using the fact that both terms in the definition of $I(X;Y;Z)$ (Eq. 2.45) are non-negative, the following bounds can be obtained on the third order mutual information corresponding to conditions where one of the terms is zero and the other takes an extreme value [44, 32]:

$$\begin{aligned} -\min\{S(X),S(Y),S(Z)\} \le I(X;Y;Z) \;&\le\; \min\{I(X;Y),I(Y;Z),I(X;Z)\} \\ &\le\; \min\{S(X),S(Y),S(Z)\}\,. \end{aligned} \tag{2.47}$$

2.3.4.1 Sign of the third order mutual information

The sign of the third order mutual information depends on the relative magnitude of the two terms in Eq. 2.45. This section presents scenarios where the third order mutual information takes a particular sign.

Considering the negative case first, $I(X;Y;Z)$ will be negative if, X and Y are independent in the absence of knowledge of Z, implying $I(X;Y) = 0$, but the two become correlated conditional on Z, implying $I(X;Y|Z) > 0$. This can be seen in terms of distributions as follows. Given $X \perp Y$, we can write the joint distribution of the three variables as

$$p(x, y, z) = p(x)p(y)p(z|x, y) \tag{2.48}$$

which gives

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(z)p(y)p(z|x, y)}{p(z)}\,. \tag{2.49}$$

In general, the above distribution does not factorize as the product $p(x|z)p(y|z)$, and, therefore, from Eq. 2.42, $I(X;Y|Z)$ can be positive. To see this more concretely with an example, suppose X, Y and Z are binary variables, and X and Y are independent and their joint distribution is:

$$p(X, Y) = \begin{pmatrix} 1/4 & 1/4 \\ \\ 1/4 & 1/4 \end{pmatrix}\,. \tag{2.50}$$

Now, suppose $Z = (X + Y)$ mod2 so that the value of Z is zero if either X or Y is zero, and one if both are one [45]. Using chain rule, and the above information we can compute the joint probability of, for example, $p(X = 0, Y = 0, Z = 0)$, as

$$\begin{aligned} p(X = 0, Y = 0, Z = 0) &= p(X = 0, Y = 0)p(Z = 0|X = 0, Y = 0) \\ &= 1/4 \times 1 \\ &= 1/4\,. \end{aligned} \tag{2.51}$$

Table 2.1: Calculation of $p(X,Y,Z)$ for $X \perp Y, Z = (X+Y) \mathrm{mod}\, 2$

X	Y	Z	$p(X,Y)$	$p(Z\|X,Y)$	$p(X,Y,Z)$
0	0	0	1/4	1	1/4
0	1	0	1/4	0	0
1	0	0	1/4	0	0
1	1	0	1/4	1	1/4
0	0	1	1/4	0	0
0	1	1	1/4	1	1/4
1	0	1	1/4	1	1/4
1	1	1	1/4	0	0

Above computation for all combinations of the three variable are listed in Table 2.1.

Using Table 2.1, the marginals of $p(X,Y,Z)$ are obtained as

$$p(X) = p(Y) = p(Z) = \begin{pmatrix} 1/2 \\ \\ 1/2 \end{pmatrix} \tag{2.52}$$

$$p(X,Y) = p(X,Z) = p(Y,Z) = \begin{pmatrix} 1/4 & 1/4 \\ \\ 1/4 & 1/4 \end{pmatrix} \tag{2.53}$$

and the corresponding entropies using Eq. 2.18 are

$$\begin{aligned} S(X) = S(Y) = S(Z) &= 1 \\ S(X,Y) = S(Y,Z) = S(X,Z) &= 2 \end{aligned} \tag{2.54}$$

where base 2 is used for the logarithm. The entropy of the full distribution, computed using Table 2.1, is

$$S(X,Y,Z) = 2\,. \tag{2.55}$$

Using above entropy values and Eqs. 2.34 and 2.44, the mutual informations are obtained as

$$\begin{aligned} I(X;Y) = I(X;Y) = I(X;Y) &= 1+1-2 = 0 \\ I(X;Y;Z) &= (1+1+1) - (2+2+2) + 2 = -1\,. \end{aligned} \tag{2.56}$$

Thus, all pairwise mutual informations are zero, indicating no correlations at the pairwise level, but the third order mutual information is non-zero is negative. Thus, introduction of a third variable given by a deterministic function of the other two variables introduced correlations among all three variables.

We can similarly imagine scenarios where the third order mutual information is positive. For instance, if X and Y are correlated in the absence of information on Z, but become independent if Z is known, then $I(X;Y;Z) = I(X;Y) > 0$. The joint distribution in this case is given by

$$\begin{aligned} p(x,y,z) &= p(z)p(x|z)p(y|z) \\ &= p(z)\frac{p(x,z)}{p(z)}\frac{p(y,z)}{p(z)} \\ &= \frac{p(x,z)p(y,z)}{p(z)} \end{aligned} \tag{2.57}$$

and, in general, the marginal $p(X,Y)$ from the above distribution does not factorize as $p(X)p(Y)$ giving $I(X;Y) > 0$. Following is an illustration for the case of binary variables. Using Eq. 2.57, starting with the following distributions

$$p(Z) = \begin{pmatrix} 1/2 \\ \\ 1/2 \end{pmatrix} \quad ; \quad p(X,Z) = \begin{pmatrix} 1/2 & 0 \\ \\ 0 & 1/2 \end{pmatrix} \quad ; \quad p(Y,Z) = \begin{pmatrix} 0 & 1/2 \\ \\ 1/2 & 0 \end{pmatrix}, \tag{2.58}$$

we can construct the full distribution and verify that the entropies are

$$S(X) = S(Y) = S(Z) = S(X,Y) = S(Y,Z) = S(X,Z) = S(X,Y,Z) = 1 \tag{2.59}$$

giving the mutual informations

$$\begin{aligned} & I(X;Y) = I(Y;Z) = I(X;Z) = 1 \\ \text{and} \qquad & I(X;Y;Z) = 1 \end{aligned} \tag{2.60}$$

which shows that $I(X;Y;Z)$ is positive. Note that the conditional mutual information $I(X;Y|Z)$ using Eq. 2.40 is indeed zero.

2.3.5 Mutual information expansion (MIE) of entropy

The ideas described in the previous sections, can be generalized to define the mutual information among any number of variables. By analogy to the definition of third order mutual information (Eq 2.43), the mutual information among N variables captures the change in mutual information (or correlation) among any subset of $N-1$ variables due to the knowledge of the last variable:

$$I(X_1;\ldots;X_N) \equiv I(X_1;\ldots;X_{N-1}) - I(X_1;\ldots;X_{N-1}|X_N)\,. \tag{2.61}$$

Expressions for the N-order mutual information in terms of the entropies of all marginals of the joint distribution of the N-variables can be derived by using the following recursion formula [32]

$$\begin{aligned} I_N(X_1;\ldots;X_N) &= I_{N-1}(X_1;\ldots;X_{N-2};X_{N-1}) + I_{N-1}(X_1;\ldots;X_{N-2};X_N) \\ &- I_{N-1}(X_1;\ldots;X_{N-2};X_{N-1},X_N) \end{aligned} \tag{2.62}$$

where I_k denotes a k-order mutual information. Note that the comma in the third term indicates a mutual information involving the joint pdf of variables X_{N-1} and X_N. For illustration, this recursion relation is now used to obtain an expression for the fourth-order mutual information which is defined as

$$I(X_1;X_2;X_3;X_4) \equiv I(X_1;X_2;X_3) - I(X_1;X_2;X_3|X_4)\,. \tag{2.63}$$

Using Eq. 2.62 and a shorthand notation $X_i \equiv i$, we have

$$\begin{aligned}
I_4(1;2;3;4) &= I_3(1;2;3) + I_3(1;2;4) - I_3(1;2;3,4) \\
&= S(1) + S(2) + S(3) - (S(1,2) + S(1,3) + S(2,3)) + S(1,2,3) \\
&\quad + S(1) + S(2) + S(4) - (S(1,2) + S(1,4) + S(2,4)) + S(1,2,4) \\
&\quad - (S(1) + S(2) + S(3,4) - (S(1,2) + S(1,3,4) + S(2,3,4)) + S(1,2,3,4)) \\
&= S(1) + S(2) + S(3) + S(4) \\
&\quad - (S(1,2) + S(1,3) + S(1,4) + S(2,3) + S(2,4) + S(3,4)) \\
&\quad + S(1,2,3) + S(1,2,4) + S(1,3,4) + S(2,3,4) \\
&\quad - S(1,2,3,4)
\end{aligned} \tag{2.64}$$

where the second step used Eq. 2.44. To illustrate the pattern, we list the expressions of mutual informations that have been derived so far:

$$\begin{aligned}
I_2(1;2) &= S(1) + S(2) - S(1,2) \\
I_3(1;2;3) &= S(1) + S(2) + S(3) - (\, S(1,2) + S(1,3) + S(2,3)\,) + S(1,2,3) \\
I_4(1;2;3;4) &= S(1) + S(2) + S(3) + S(4) \\
&\quad - \quad (S(1,2) + S(1,3) + S(1,4) + S(2,3) + S(2,4) + S(3,4)) \\
&\quad + \quad S(1,2,3) + S(1,2,4) + S(1,3,4) + S(2,3,4) \\
&\quad - \quad S(1,2,3,4)\,.
\end{aligned} \tag{2.65}$$

In general, for any $N > 2$, the N-th order mutual information is given by [37, 43]:

$$I_N(X_1;\ldots;X_N) \equiv \sum_{j=1}^{N} (-1)^{j+1} \sum_{C_j^N} S_j(X_{i_1},\ldots,X_{i_j}) \tag{2.66}$$

where $i_1, i_2, .. \in 1, .., N$ and $\sum_{C_j^N}$ denotes a summation over all unique combinations of j variables out of the full N variables. The N-order mutual information, similar to pairwise

and third order mutual informations, is symmetric in all variables. I_N for $N > 2$ can be of either sign and, due to this property, bounds analogous to Eqs 2.37 and 2.47 cannot be placed on I_N for $N > 3$.

Note that the last term of Eq. 2.66, is the entropy of the joint distribution of all variables. The equations for I_N can be inverted to obtain an expansion for S_N, the entropy of the full N-dimensional distribution, in terms of mutual informations of increasing orders. For $N = 2, 3$ and 4, we have

$$\begin{aligned}
S(1,2) &= S(1) + S(2) - I_2(1;2) \\
S(1,2,3) &= S(1) + S(2) + S(3) \; - (I_2(1;2) + I_2(1;3) + I_2(2;3)) + I_3(1;2;3) \\
S(1,2,3,4) &= S(1) + S(2) + S(3) + S(4) \\
&- (I_2(1;2) + I_2(1;3) + I_2(1;4) + I_2(2;3) + I_2(2;4) + I_2(3;4)) \\
&+ I_3(1;2;3) + I_3(2;3;4) + I_3(1;3;4) + I_3(1;2;4) \\
&- I_4(1;2;3;4)\,. \qquad (2.67)
\end{aligned}$$

which can be verified using expressions for mutual information from Eq. 2.65. Generalizing for any $N > 2$, gives the Mutual Information Expansion (MIE) for the entropy of an N-dimensional system [31, 32, 30] :

$$\begin{aligned}
S_N &\equiv S(X_1, \ldots, X_N) = \sum_N S(X_{i_1}) - \sum_{C_2^N} I_2(X_{i_1}, X_{i_2}) \\
&+ \sum_{C_3^N} I_3(X_{i_1}, X_{i_2}, X_{i_3}) + \ldots + (-1)^{N+1} I_N(X_{i_1}, \ldots, X_{i_N})\,. \qquad (2.68)
\end{aligned}$$

This expansion allows one to compute the entropy while systematically including the influence of correlations at successively higher orders as captured by the corresponding mutual informations. The MIE is exact if mutual informations at all orders, including at the highest order I_N are included. However, since I_N requires S_N itself (Eq. 2.66), the

MIE is not particularly useful for computing the entropy if there are strong correlations at all orders up to N. However, the expansion becomes useful if correlations beyond a certain order l ($< N$) are absent or weak enough to justify dropping terms of order greater than l, to provide a l-level approximation, $S_N^{(l)}$, to the full entropy, S_N:

$$\begin{aligned} S_N &\approx S_N^{(1)} = \sum_N S_1 \\ &\approx S_N^{(2)} = \sum_N S_1 - \sum_{C_2^N} I_2 \\ &\approx S_N^{(3)} = \sum_N S_1 - \sum_{C_2^N} I_2 + \sum_{C_3^N} I_3 \\ &\vdots \\ &\approx S_N^{(l)} = \sum_N S_1 - \sum_{C_2^N} I_2 + \ldots + (-1)^{l+1} \sum_{C_l^N} I_l \end{aligned} \tag{2.69}$$

where $\sum_N S_1$ denotes the sum of entropy of all singlet marginals. $S_N^{(1)}$, $S_N^{(2)}$ and $S_N^{(3)}$ are termed the singlet, doublet and triplet level approximations of the entropy, respectively. From Eq. 2.30, the singlet approximation to the entropy places an upper bound on the true entropy, i.e. $S_N \leq S_N^{(1)}$. Thus, for a correlated system, the mutual information terms in the MIE collectively reduce the entropy of the system relative to an uncorrelated system. Also, since the higher-order mutual informations, except pairwise, can be of either sign the entropy approximations at successive orders may not decrease monotonically as higher order mutual informations are included. The l-level entropy approximation do not place any bounds on the true entropy, S_N, a possible explanation for which is given in the next section.

2.4 Mutual information and MIE in terms of superposition approximations

In the previous section we described I_N, the mutual information at order N, as a measure of correlation existing among N variables, and presented expressions for it in terms of the entropies of the marginal pdfs of the N-dimensional distribution, p_N. We also presented a series of approximations (Eq. 2.69) to the entropy of an N-dimensional distribution in terms of sums of mutual informations of different orders. Here, we first review the GKSA distribution which expresses an N-dimensional pdf in terms of all its marginals up to order $N-1$ and show that the N-order mutual information can be written in terms of the GKSA. We then derive the SA-l distribution which allows one to express an N-dimensional distribution in terms of marginal pdfs of up to a given order $l < N$ and show that the l-level entropy approximation can be written in terms of the SA-l. The overall scheme can be summarized as

$$p_N(X_1,\ldots,X_N)\,;\, S_N \quad \rightarrow \quad \{\text{marginals of } p_N\} \quad \rightarrow \quad p_N^{(l)}(X_1,\ldots,X_N)\,;\, S_N^{(l)}\,. \tag{2.70}$$

2.4.1 Generalized Kirkwood superposition approximation

We begin by noting that, for an $N = 2$ variable system, Eq. 2.36 relates the pairwise mutual information to the joint distribution and the product distribution of the marginal pdfs as

$$I_2(X_1;X_2) = \sum\sum p(x_1,x_2)\ln\left(\frac{p(x_1,x_2)}{p(x_1)p(x_2)}\right). \tag{2.71}$$

The mutual information at any order N, can be written in a similar form in terms of the GKSA, $p_N^{(N-1)}$ as [37, 32]:

$$I_N(X_1;\ldots;X_N) = (-1)^N \sum_{X_1,\ldots,X_N} p_N(\mathbf{x})\ln\frac{p_N(\mathbf{x})}{p_N^{(N-1)}(\mathbf{x})}\,. \tag{2.72}$$

The superscript $(N-1)$ indicates that the GKSA is a function of all marginals of p_N up to order $N-1$. The first proposed and the simplest SA, corresponding to $N=3$, is the KSA [25]

$$p_3^{(2)}(1,2,3) = \frac{p(1,2)p(1,3)p(2,3)}{p(1)p(2)p(3)} \tag{2.73}$$

which is a function of only the one- and two-dimensional marginals of the full distribution $p_3(1,2,3)$. This can be verified by substituting the KSA in Eq. 2.72 and comparing the resulting expression in terms of entropy of the marginals of p_3 with I_3 in Eq. 2.65. For $N=4$, the $p_4^{(3)}$ distribution is the FKSA [27]

$$p_4^{(3)}(1,2,3,4) = \frac{p(1,2,3)p(1,2,4)p(1,3,4)p(2,3,4)}{\frac{p(1,2)p(1,3)p(1,4)p(2,3)p(2,4)p(3,4)}{p(1)p(2)p(3)p(4)}} \tag{2.74}$$

which includes all marginal pdfs of up to order $N-1=3$. Denoting the product of all C_j^N j-order marginal pdfs of the N-dimensional distribution by

$$\mathscr{P}_{(N,j)} \equiv \prod_{1\le i_1<i_2<\ldots<i_j\le N} p_j(i_1,\ldots,i_j) \tag{2.75}$$

the GKSA for any $N \geq 2$, is given by [37, 32, 30]:

$$\begin{aligned} p_N^{(N-1)}(X_1,\ldots,X_N) &= \mathscr{P}_{(N,N-1)}^{+1}\mathscr{P}_{(N,N-2)}^{-1}\times\ldots\times\mathscr{P}_{(N,1)}^{(-1)^{N-2}} \\ &= \prod_{j=N-1}^{1}\mathscr{P}_{(N,j)}^{(-1)^{N-1-j}}. \end{aligned} \tag{2.76}$$

The exponents of the marginals of successive orders alternate between $+1$ and -1, with the highest order $(=N-1)$ having an exponent of $+1$. Substituting the GKSA in Eq. 2.72, gives the expression for I_N in terms of entropies of all marginal distributions (Eq. 2.66).

Note that, if $p_N = p_N^{(N-1)}$, then $I_N = 0$, implying that in the absence of correlations at the highest order, the GKSA is exact. Based on this property, which is analogous to the definition of independence of two variables, GKSA has been used as a definition of "semi-independence" of more than two variables [38]. Also note that, if there are correlations

at order N, the GKSA distribution is not normalized, and, in general, the normalization can be either greater or less than one [46, 26, 30]. As a result despite the resemblance of Eq. 2.72 to the KL distance expression (Eq. 2.31), the non-negativity of KL distance does not apply here. This is consistent with the fact that I_N, for $N > 2$, can be of either sign.

The GKSA is related to the $l = N - 1$ level entropy approximation, Eq. 2.69, as

$$S_N^{(N-1)} = -\sum_{\mathbf{X}} p_N(\mathbf{X}) \ln p_N^{(N-1)}(\mathbf{X}) \,. \tag{2.77}$$

This can be seen by using the logarithm to convert the product of pdfs in the GKSA to a summation and, for each term, marginalizing the variables not present in the marginal pdf under the logarithm. In other words, the cross-entropy of the GKSA distribution with respect to p_N gives the $(N-1)$-level entropy approximation. Since, GKSA is not normalized if there are correlations at the highest order $S^{(N-1)}$ does not give a bound on S_N. In general, the l-level entropy approximation can be written as

$$S_N^{(l)} = -\sum_{\mathbf{X}} p_N(\mathbf{X}) \ln p_N^{(l)}(\mathbf{X}) \tag{2.78}$$

where $p^{(l)}$ denotes the l-level SA, or SA-l, approximation to p_N which is derived next.

2.4.2 The Superposition Approximation at level l (SA-l)

The SA-l distributions include marginal pdfs of p_N of up to order l and have the general form

$$\begin{aligned} p_N^{(l)} &= \mathscr{P}_{(N,l)}^{a(l;N,l)} \mathscr{P}_{(N,l-1)}^{a(l-1;N,l)} \times \ldots \times \mathscr{P}_{(N,l)}^{a(1;N,l)} \\ &= \prod_{j=l}^{1} \mathscr{P}_{(N,j)}^{a(j;N,l)} \end{aligned} \tag{2.79}$$

where $a(j; N, l)$ is the exponent of the product of j-order marginal pdfs; it depends on the level of approximation, l, and the dimensionality, N. The exponents are derived by

recursively applying the GKSA on the marginals of the highest order until marginals of only order l and lower remain. The procedure is best illustrated through simple examples. In the next subsection, we derive the doublet (SA-2) and triplet (SA-3) level SAs for an $N = 5$ dimensional distribution. Generalization to any N and l is discussed in the subsequent subsection.

2.4.2.1 Examples: SA-2 and SA-3 for a 5-dimensional distribution

From Eq.(2.76) the GKSA expression for $N = 5$ is

$$\begin{aligned} & p_5^{(4)}(1,2,3,4,5) = \\ & \frac{p(1,2,3,4)p(1,3,4,5)p(1,2,3,5)p(1,2,4,5)p(2,3,4,5)}{p(1,2,3)p(1,2,4)p(1,2,5)p(1,3,4)p(1,3,5)p(1,4,5)p(2,3,4)p(2,3,5)p(2,4,5)p(3,4,5)} \\ \times & \frac{p(1,2)\,p(1,3)\,p(1,4)\,p(1,5)\,p(2,3)p(\,p(2,4)\,p(2,5)\,p(3,4)\,p(3,5)\,p(4,5)}{p(1)p(2)p(3)p(4)p(5)} \\ = & \mathscr{P}^{+1}_{(5,4)}\mathscr{P}^{-1}_{(5,3)}\mathscr{P}^{+1}_{(5,2)}\mathscr{P}^{-1}_{(5,1)}\,. \end{aligned} \tag{2.80}$$

By applying the FKSA from Eq. 2.74 to each of the five 4-D pdfs we can express the product of the 4-D pdfs in terms of the 3-, 2- and 1-D pdfs as

$$\begin{aligned} \mathscr{P}^{+1}_{(5,4)} &= \frac{(p(1,2,3)p(1,2,4)p(1,2,5)p(1,3,4)p(1,3,5)p(1,4,5)p(2,3,4)p(2,3,5)p(2,4,5)p(3,4,5))^2}{\frac{(p(1,2)\,p(1,3)\,p(1,4)\,p(1,5)\,p(2,3)p(\,p(2,4)\,p(2,5)\,p(3,4)\,p(3,5)\,p(4,5))^3}{(p(1)p(2)p(3)p(4)p(5))^4}} \\ &= \mathscr{P}^{+2}_{(5,3)}\mathscr{P}^{-3}_{(5,2)}\mathscr{P}^{+4}_{(5,1)}\,. \end{aligned} \tag{2.81}$$

The triplet level SA, $p_5^{(3)}$, is then obtained by substituting Eq. 2.81 in Eq. 2.80

$$\begin{aligned} p_5^{(3)} &= \frac{(p(1,2,3)p(1,2,4)p(1,2,5)p(1,3,4)p(1,3,5)p(1,4,5)p(2,3,4)p(2,3,5)p(2,4,5)p(3,4,5))^1}{\frac{(p(1,2)\,p(1,3)\,p(1,4)\,p(1,5)\,p(2,3)p(2,4)\,p(2,5)\,p(3,4)\,p(3,5)\,p(4,5))^2}{(p(1)p(2)p(3)p(4)p(5))^3}} \\ &= \mathscr{P}^{+1}_{(5,3)}\mathscr{P}^{-2}_{(5,2)}\mathscr{P}^{+3}_{(5,1)}. \end{aligned} \tag{2.82}$$

Next, applying the KSA to express the product of the triplet marginals in terms of doublet and singlet marginals gives

$$\mathscr{P}^{+1}_{(5,3)} = \mathscr{P}^{+3}_{(5,2)}\mathscr{P}^{-6}_{(5,1)} \tag{2.83}$$

and the doublet level SA, $p_5^{(2)}$, is obtained by substituting Eq. 2.83 in Eq. 2.82:

$$\begin{aligned} p_5^{(2)} &= = \frac{p(1,2)\,p(1,3)\,p(1,4)\,p(1,5)\,p(2,3)p(\,p(2,4)\,p(2,5)\,p(3,4)\,p(3,5)\,p(4,5)}{(p(1)p(2)p(3)p(4)p(5))^3} \\ &= \mathscr{P}_{(5,2)}^{+1}\mathscr{P}_{(5,1)}^{-3}\,. \end{aligned} \tag{2.84}$$

Thus the exponents for SA-2 and SA-3 for $N = 5$, are

$$\begin{aligned} &SA-2\ : a(2;5,2) = +1 \quad a(1;5,2) = -3 \\ &SA-3\ : a(3;5,3) = +1 \quad a(2;5,3) = -2 \qquad\qquad a(1;5,3) = +3\ . \end{aligned} \tag{2.85}$$

2.4.2.2 SA-l for any N, l

Generalizing the above procedure, we obtain the following expression for the exponents in Eq. 2.79

$$a(j;N,l) = (-1)^{l-j}\prod_{i=1}^{l-j}\frac{N-l+i-1}{i} \tag{2.86}$$

where $j = 1,\ldots,l\,; l < N$, as derived in the Appendix A. To illustrate the pattern, we list the first five SA-l:

$$\begin{aligned} p_N^{(1)} &= \mathscr{P}_{(N,1)}^{+1} \\ p_N^{(2)} &= \mathscr{P}_{(N,2)}^{+1}\mathscr{P}_{(N,1)}^{-(N-2)} \\ p_N^{(3)} &= \mathscr{P}_{(N,3)}^{+1}\mathscr{P}_{(N,2)}^{-(N-3)}\mathscr{P}_{(N,1)}^{+\frac{(N-3)(N-2)}{2!}} \\ p_N^{(4)} &= \mathscr{P}_{(N,4)}^{+1}\mathscr{P}_{(N,3)}^{-(N-4)}\mathscr{P}_{(N,2)}^{+\frac{(N-4)(N-3)}{2!}}\mathscr{P}_{(N,1)}^{-\frac{(N-4)(N-3)(N-2)}{3!}} \\ p_N^{(5)} &= \mathscr{P}_{(N,5)}^{+1}\mathscr{P}_{(N,4)}^{-(N-5)}\mathscr{P}_{(N,3)}^{+\frac{(N-5)(N-4)}{2!}}\mathscr{P}_{(N,2)}^{-\frac{(N-5)(N-4)(N-3)}{3!}}\mathscr{P}_{(N,1)}^{+\frac{(N-5)(N-4)(N-3)(N-2)}{4!}}\,. \end{aligned} \tag{2.87}$$

One can furthermore verify that, for $l = N-1$, the SA-l becomes the GKSA, as it should, since it was the starting point for deriving SA-l. Also, if all variables are independent, the

SA-l ($l > 1$) reduces to the product of the 1-D pdfs, as expected. In this work, only the singlet, doublet and triplet level approximations are used.

2.4.3 Mixed superposition approximations

The SA-l distributions (Eq 2.79) correspond to the l-level entropy approximations, which include all mutual informations of up to order l. However, one can imagine a scenario where only select mutual informations of various orders contribute significantly to the overall entropy; i.e., only certain variables of the full distribution are highly correlated. A mixed-order truncation of the mutual information expansion, where select mutual information terms are retained, can be written as the cross-entropy a SA approximation of mixed order with respect to the true distribution. The mixed-SA distributions can be derived by starting from the SA-l corresponding to the highest order of included mutual informations, and then approximating the marginal pdfs corresponding to the dropped mutual informations using appropriate GKSAs, as illustrated below.

Suppose, for an $N = 5$ dimensional system, we wish to derive the mixed-SA distribution, $p_5^{(m)}$, corresponding to the entropy expansion

$$\begin{aligned} S^{(m)} = S(1) \quad &+ \quad S(2) + S(3) + S(4) + S(5) \\ &- \quad (I_2(1;2) + I_2(1;3) + I_2(2;3) + I_2(2;5) + I_2(3;4) + I_2(4;5)) \\ &+ \quad I_3(1;2;3)\,. \end{aligned} \tag{2.88}$$

The above expansion effectively assumes that the remaining mutual informations – $I(1;4), I(1;5), I(2;4)$, $I(3;5)$ among doublets, all triplets except $I(1,2,3)$ and all fourth-order and the fifth order mutual informations – in the full MIE are zero. In order to derive $p_5^{(m)}$, since in Eq. 2.88 the highest order among the mutual informations is order three, we

start with the SA-3 approximation from Eq. 2.82

$$p_5^{(3)} = \frac{p(1,2,3)p(1,2,4)p(1,2,5)p(1,3,4)p(1,3,5)p(1,4,5)p(2,3,4)p(2,3,5)p(2,4,5)p(3,4,5)}{\frac{(p(1,2)\,p(1,3)\,p(1,4)\,p(1,5)\,p(2,3)p(2,4)\,p(2,5)\,p(3,4)\,p(3,5)\,p(4,5))^2}{(p(1)p(2)p(3)p(4)p(5))^3}} . \tag{2.89}$$

Expanding all triplet pdfs, except $p(1,2,3)$ which corresponding to $I_3(1,2,3)$, using the KSA and furthermore the doublets for $p(1,4), p(1,5), p(2,4)$ and $p(3,5)$ corresponding to the zero pairwise mutual informations, gives the final mixed-SA:

$$p_5^{(m)} = \frac{p(1,2,3)p(2,5)p(3,4)p(4,5)}{p(2)p(3)p(4)p(5)} . \tag{2.90}$$

As a consistency check, note that $p_5^{(m)}$ reduces to the product of the singlets if all variables are independent. It can be verified that the cross-entropy of $p_5^{(m)}$ matches with Eq. 2.88.

2.5 Conclusions

This chapter used concepts from information theory to motivate mutual information as a measure of correlations among variables of a multivariate stochastic system. We also presented the MIE of entropy, an exact expansion of the entropy of the system in terms of mutual information sums of increasing order, and developed approximations to the entropy by truncating the MIE at various levels under the assumption that higher order correlations are weak. We showed that any given entropy approximation can be written as a cross-entropy of a superposition approximation to the full distribution of the system. Below we summarize some properties of SA-based distributions and make a few additional points.

(i) **Normalization**

If the neglected correlations are not absent, the SA-based distributions are not normalized, and consequently the cross-entropy does not provide a bound on the

true entropy, S_N. In Chapter 3 we develop a normalized distribution, $\tilde{p}^{(l)}$, closely related to the SA-l, whose cross-entropy places an upper bounds on S_N (Chapter 5).

(ii) **Probabilistic interpretation of GKSA**

Based on probabilistic conditions of independence, Singer showed that the GKSA is exact if any variable out of the total N variables is independent of all other variables [30]. This is equivalent to the N-order mutual information vanishing if any one of the N variables is independent of the other variables (Section 2.3.4). In terms of N-particle distribution functions of liquids, this condition is equivalent to the assumption that one of the particles is not interacting with the other particles.

(iii) **SA-based distributions do not preserve marginals**

The SA-based distributions can be considered probabilistic closure equations, since they approximate a higher dimensional distribution in terms of its marginals. It is worth mentioning that the SA distributions, in general, are not marginal-preserving closures; i.e., a marginal of the SA-l need not equal the corresponding marginal of the full distribution that the SA-l approximates.

(iv) **SA-l is a fully coupled distribution**

The SA-l distributions, except for the trivial case of $l = 1$, are fully coupled "all-to-all" distributions, which means that they cannot be factorized by grouping the marginals into factors with non-overlapping subsets of the variables. This property has implications for numerical calculations of marginals or the normalization of the SA-l distributions. For instance, calculation of the normalization, which requires summing over the N-dimensional state space will be an $O(B^N)$ calculation, where B is the number of discrete states for each variable, and hence computationally

infeasible for large N.

Finally, we note that the SA-based approximations of a multivariate distribution and the MIE-based approximations to the entropy of the distribution are based on assumptions regarding the strength of correlations of various orders among the variables of the system. However, for a complex and high-dimensional system, such as the molecular systems considered in this work, it is difficult to determine the strength of various correlations *a priori*, necessitating empirical tests of the approximations. One such test is developed in the next chapter.

Chapter 3

Superposition Approximation Based Conformational Sampling

3.1 Introduction

The thermal fluctuations of the internal degrees of freedom of a flexible molecule are correlated, due to the bonded and non-bonded interactions among its atoms. The configurational entropy is a measure of a molecule's thermal fluctuations. Recent calculations of the configurational entropy for small (<50 atoms) molecules, with two independent approaches [47, 48], have provided preliminary insights into the contributions of correlations of various orders to the conformational fluctuations.

In one approach, the MIE of entropy was used to approximate the full entropy using mutual information terms of orders up to only l=1, 2 or 3, effectively assuming that correlations of higher orders are absent [48]. The retained mutual information terms were computed from marginals of the Boltzmann distribution; these, in turn, were computed as normalized histograms of coordinate values in a Boltzmann distributed ensemble of conformations generated by a MD simulation. The MIE-based low-order approximations of the entropy were furthermore compared with numerical results from the Mining Minima

(M2) method [47]. In M2, the configurational integral of a molecule is computed as a sum over local energy minima on a molecular mechanics force-field energy surface, where the energy minima are enumerated using an aggressive search algorithm. M2 provides the free energy and average energy of a molecule, and the configurational entropy can be computed from the difference between these two quantities. Since the M2 method uses the full configurational integral (subject to the approximation of summing over local energy wells), it implicitly accounts for all physically relevant correlations at a given temperature. The observation of relevance here is that, for small molecules, molecular entropies estimated with the MIE at the doublet and triplet levels – i.e., neglecting correlations above second and third order, respectively – agreed well with independent M2 calculations [48]. Because M2 implicitly includes all correlations, this observation suggests that most of the physically relevant fluctuations of these molecules involved only low-order correlations. This leads to the hypothesis that conformational probability distributions may be well described with a tractable set of low-order distribution functions. In this chapter, a novel approach to test this hypothesis is presented.

More particularly, we pose the question: how accurately can the conformational fluctuations of a molecule in a thermal environment be described without accounting for high-order correlations? This is addressed by developing a conformational sampling algorithm that allows one to sample conformations in the full N-dimensional space using only the low-order marginal distributions of the full Boltzmann distribution, the same marginals that were used for the MIE entropy calculations mentioned above. The sampling algorithm at level l uses pdfs of highest order l. The ensemble of conformations sampled at different levels is evaluated by comparing with the MD ensemble used to populate the pdfs. Based on tests on multiple small molecules with different bonded topologies,

we find, in brief, that the ensemble of sampled conformations better resembles the MD ensemble as successively higher-order correlations are included. For molecules with linear chain-like topologies, conformations sampled at the doublet level match MD conformations rather well, while the triplet level sampling generated high quality conformations for all molecules. These results suggest that the low-order correlations suffice to describe most of the conformational fluctuations of molecules in a thermal environment. The rest of this chapter is divided into two main sections. The first develops the sampling algorithm and discusses its mathematical and computational properties, and the second describes the application of the sampling algorithm to molecular system. A practical application of the sampling algorithms for calculation of absolute free energy of molecules is presented in Chapter 4. This chapter is based on Ref [49].

3.2 SA-l based ancestral sampling algorithms

Following notation from Chapter 2, we consider an N-dimensional pdf $p_N(X_1, X_2, \ldots, X_N)$ of discrete valued variables $X_1, X_2, \ldots, X_N$ each of which can take B different values, so that the N-dimensional discrete space consists of B^N points. The distribution p_N will typically be a high dimensional distribution but its low-order marginals can be numerically approximated using samples drawn from p_N. The goal in this section is to sample points the N-dimensional space using marginal pdfs of p_N of up to order $l < N$. The sampling algorithm is based on the SA-l distributions described in Section 2.4.2, and is done using a variant of the ancestral sampling algorithm [50].

3.2.1 Ancestral sampling

The ancestral algorithm allows exact sampling from an N-dimensional distribution and is based upon the chain rule:

$$p_N(X_1, ..., X_N) = p(X_1)\, p(X_2|X_1)\, p(X_3|X_1, X_2)\ \times ... \times p(X_N|X_1, ..., X_{N-1})\,. \qquad (3.1)$$

This product can be represented graphically by the directed, acyclic graph shown in Figure 3.1, where each node represents a variable and the incoming arrows originate from the "parent" nodes, that is, the nodes of the conditioning variables. The following pseudo-code explains how the ancestral algorithm samples a point $\mathbf{x} = (x_1, ..., x_N)$ from p_N:

Algorithm 1: Ancestral sampling

Step 1: $x_1 \sim p(X_1)$

Step 2: $x_2 \sim p(X_2|pa(X_2)) = p(X_2|x_1)$

Step 3: $x_3 \sim p(X_3|pa(X_3)) = p(X_3|x_1, x_2)$

Step 4: FOR $k = 4$ to N

$x_k \sim p(X_k|pa(X_k)) = p(X_k|x_1, ..., x_{k-1})$

ENDFOR

---

where $pa(X_k)$ are the parent variables for X_k and "$\sim$" means "sampled from". Thus, the first variable is sampled from its one-dimensional singlet distribution, and each subsequent variable is sampled from its one-dimensional distribution conditioned upon the values of all the variables that have been sampled so far. This algorithm is exact since, in the limit of infinite sampling, the sampled points are distributed as p_N. It is also worth noting that variables can be sampled in any order, so long as one has access to the required conditionals; and that, in contrast with Monte Carlo or molecular dynamics sampling, successive samples are uncorrelated.

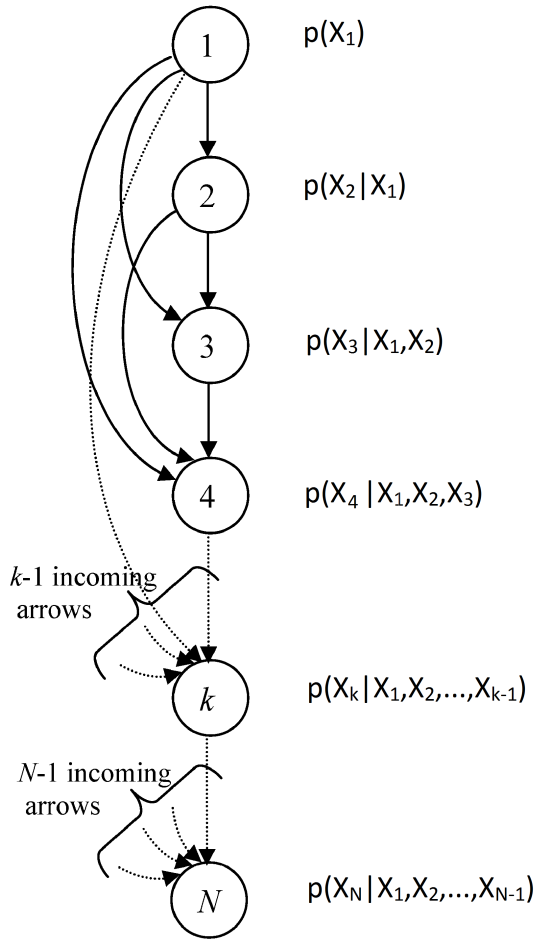


Figure 3.1: Representation of an N-dimensional distribution function $p(X_1, ..., X_N)$ as a directed graph. Variables are represented by circles containing the label of the respective variables. Solid arrows indicate the conditional dependencies of each variable as per the one dimensional conditional distribution indicated on the right of each node. Dashed lines schematize elided portions of the graph. The ancestral sampling uses exact conditional distributions for each variable. The l-level sampling algorithms presented in Section 3.2 use approximations to conditional pdfs for $(l+1)$-th and following variables. Doublet level algorithm uses Eq. 3.10 while triplet level algorithm uses Eq. 3.11.

3.2.2 Superposition approximation based conditional distribution

In ancestral sampling, the k-th variable is sampled from its conditional distribution given values of the previous $k-1$ sampled variables. Using the product rule for conditional pdfs, we have

$$p(X_k|x_1,...,x_{k-1}) = \frac{p_k(x_1,...,x_{k-1},X_k)}{p_{k-1}(x_1,...,x_{k-1})} \tag{3.2}$$

which requires marginal pdfs of order k and $k-1$. Since the marginal pdfs of orders $k > l$ are not available, we approximate them using the SA-l

$$\begin{aligned} p(X_k|x_1,.,x_{k-1}) &\approx p_k^{(l)}(X_{i_k}|x_1,.,x_{k-1}) \\ &= p_k^{(l)}(x_1,.,x_{k-1},X_k)\frac{1}{N_k(x_1,.,x_{k-1})} \end{aligned} \tag{3.3}$$

where the pdf of the already sampled variables in the denominator of Eq. 3.3 is absorbed into the normalization constant N_k, as elaborated below. The doublet level approximations ($l = 2$) to $p(X_3|x_1,x_2)$ and $p(X_4|x_1,x_2,x_3)$ are now derived as illustrations.

Considering $p(X_3|x_1,x_2)$ first, we apply the SA-2 (Eq. 2.73) approximation to a 3-D pdf to obtain

$$\begin{aligned} p(X_3|x_1,x_2) &= \frac{p(x_2,x_1,X_3)}{p(x_1,x_2)} \\ &\approx \frac{p(x_2,X_3)p(x_1,X_3)p(x_1,x_2)}{p(x_1)p(x_2)p(X_3)}\frac{1}{p(x_1,x_2)} \\ &= \frac{p(x_1,X_3)p(x_2,X_3)}{p(X_3)}\left[\frac{1}{p(x_1)p(x_2)}\right] \end{aligned} \tag{3.4}$$

where pdfs that do not depend on X_3, the variable to be sampled, are collected in the square bracket. Eq. 3.4 is a one-dimensional distribution in X_3 and is not normalized because the SA-l distributions are not normalized. However, normalization can be imposed in the standard manner by dividing by the sum of the distribution for all values of X_3. Thus,

dividing Eq. 3.4 by

$$\left[\frac{1}{p(x_1)p(x_2)}\right]\sum_{X_3}\frac{p(x_1,x_3)p(x_2,x_3)}{p(x_3)} \tag{3.5}$$

cancels the pdfs in the square brackets giving the required doublet level approximation of the third order conditional as

$$\begin{aligned} p(X_3|x_1,x_2) &\approx p^{(2)}(X_3|x_1,x_2) \\ &= \frac{p(x_1,X_3)p(x_2,X_3)}{p(X_3)}\frac{1}{N_3(x_1,x_2)}\,. \end{aligned} \tag{3.6}$$

The normalization factor $N_3(x_1,x_2)$ is

$$N_3(x_1,x_2) = \sum_{X_3}\frac{p(x_1,x_3)p(x_2,x_3)}{p(x_3)} \tag{3.7}$$

which is a sum over only the pdfs that contain X_3. Following similar steps, we can write the doublet level approximation of the conditional distribution of X_4 given x_1, x_2 and x_3, as

$$\begin{aligned} p(X_4|x_1,x_2,x_3) &\approx p^{(2)}(X_4|x_1,x_2,x_3) \\ &= \frac{p_4^{(2)}(x_1,x_2,x_3,X_4)}{p_3^{(2)}(x_1,x_2,x_3)} \\ &= \frac{p(x_1,x_2)\,p(x_1,x_3)\,p(x_1,X_4)\,p(x_2,x_3)\,p(x_2,X_4)\,p(x_3,X_4)}{\left(p(x_1)\,p(x_2)\,p(x_3)\,p(X_4)\right)^2}\frac{1}{p_3^{(2)}(x_1,x_2,x_3)} \\ &= \frac{p(x_1,X_4)\,p(x_2,X_4)\,p(x_3,X_4)}{\left(p(X_4)\right)^2}\frac{1}{N_4(x_1,x_2,x_3)} \end{aligned} \tag{3.8}$$

where the SA-2 for $N = 4$ (Eq. 2.79 with $N = 4$ and $l = 2$) is used in the second step, and all pdfs independent of X_4 cancel due to normalization. Note that pdfs that do not contain the variable to be sampled essentially generate constants multiplying the 1-D conditional distribution of the variable, and therefore are eliminated by normalization. The normalization factor in Eq. 3.8 is given by

$$N_4(x_1,x_2,x_3) = \sum_{X_4}\frac{p(x_1,x_4)\,p(x_2,x_4)\,p(x_3,x_4)}{\left(p(x_4)\right)^2}\,. \tag{3.9}$$

More generally, at the doublet level, the normalized conditional probability distribution of variable X_k, $k > 2$, given values of the other $k-1$ variables, takes the form

$$\begin{aligned} p(X_k|x_1,\ldots,x_{k-1}) &\approx p^{(2)}(X_k|x_1,\ldots,x_{k-1}) \\ &= \frac{\prod\limits_{1\le i\le k-1} p(x_i,X_k)}{(p(X_k))^{k-2}} \frac{1}{N_k\left(x_1,\ldots,x_{k-1}\right)}\,. \end{aligned} \tag{3.10}$$

Similarly, the triplet level approximation of the conditional distribution, obtained by using SA-3 in Eq. 3.3, is:

$$\begin{aligned} p(X_k|x_1,\ldots,x_{k-1}) &\approx p^{(3)}(X_k|x_1,\ldots,x_{k-1}) \\ &= \frac{(p(X_k))^{\frac{(k-3)(k-2)}{2}} \prod\limits_{1\le i<j\le k-1} p(x_i,x_j,X_k)}{\left(\prod\limits_{1\le i\le k-1} p(x_i,X_k)\right)^{(k-3)}} \frac{1}{N_k\left(x_1,\ldots,x_{k-1}\right)} \end{aligned} \tag{3.11}$$

where $k > 3$. In this work, only the doublet and triplet level approximations to the conditional pdfs are used, but expressions for approximate conditional pdfs using SA-l at higher levels and mixed SA can be derived similarly. Note that the normalization factors can be computed efficiently on the fly as they are summations over a single variable; and that the normalization factor for the conditional pdf of X_k depends only on the previously sampled variables $x_1,\ldots,x_{k-1}$.

3.2.3 Sampling based on low-order marginal pdfs

The SA-l based conditional probability distributions are now inserted into the ancestral sampling algorithm to enable ancestral-style sampling from an approximation of the targeted N-dimensional pdf, based upon only its singlet, doublet and triplet marginal pdfs. Sampling at level l will refer to sampling using marginal pdfs of order l and lower. Thus, the singlet level algorithm uses only the N singlet pdfs, doublet level uses the

$N(N-1)/2$ doublet pdfs as well, and the triplet-level algorithm incorporates, in addition, the $N(N-1)(N-2)/6$ triplet pdfs. We now describe the sampling algorithms, starting, for completeness, with the singlet level sampling algorithm, although it does not require construction of SA-l based conditional pdfs.

For singlet level sampling, all variables are assumed to be independent, so one simply samples each variable from its singlet distribution without reference to the other variables. The pseudocode for the singlet level sampling algorithm thus is:

Algorithm 2: Singlet level sampling

Step 1: $x_1 \sim p(X_1)$

Step 2: $x_2 \sim p(X_2)$

:

Step k: $x_k \sim p(X_k)$

:

Step N: $x_N \sim p(X_N)$

This algorithm effectively samples from the N-dimensional distribution

$$\tilde{p}_N^{(1)} = p(X_1) \times p(X_2) \times \ldots \times p(X_N)\,. \tag{3.12}$$

Similar to ancestral sampling, singlet level sampling is independent of the sampling order. Also, since all singlet pdfs are normalized, $\tilde{p}_N^{(1)}$ is normalized.

For sampling at higher levels, the conditional distributions for the first l variables are computed using Eq. 3.2, with the available marginal pdfs, just as in the regular ancestral algorithm; but the SA-l based approximations are used for the subsequent variables. Thus, for the doublet level sampling algorithm, approximate conditional pdfs computed using the singlet and doublet marginal pdfs are used to sample the third and subsequent variables.

The pseudocode for doublet level sampling is:

Algorithm 3: Doublet level ($l = 2$) sampling

Step 1: $x_1 \sim p(X_1)$

Step 2: $x_2 \sim p(X_2|x_1)$

Step 3: FOR k = 3 to N

$x_k \sim p^{(2)}(X_k|x_1, ..., x_{k-1})$ from Eq. 3.10

ENDFOR

Similarly, triplet-level sampling uses the approximate conditional pdfs to sample variable X_4 onwards:

Algorithm 4: Triplet level ($l = 3$) sampling

Step 1: $x_1 \sim p(X_1)$

Step 2: $x_2 \sim p(X_2|x_1)$

Step 3: $x_3 \sim p(X_3|x_1, x_2)$

Step 3: FOR k = 4 to N

$x_k \sim p^{(3)}(X_k|x_1, ..., x_{k-1})$ from Eq. 3.11

ENDFOR

The doublet and triplet level sampling algorithms generalize to higher levels of approximation – i.e., to larger values of l – and the standard ancestral algorithm is recovered when $l = N$.

3.2.4 Properties of SA-l based ancestral sampling algorithms

As detailed above, the sampling algorithm at level l samples points in the N-dimensional space using estimates of the marginal pdfs of the full-dimensional target distribution, p_N, that are of orders l and lower, where $l < N$. These estimated marginals, termed the reference marginals, are constructed numerically, as normalized histograms,

from a finite set of samples from p_N. Given a set of N_P samples, $\Omega_P = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_P}\}$ where $\mathbf{x}_i \sim p_N$, the probability associated with a reference histogram bin is the fractional occupancy of the bin. For example, if the number of samples in Ω_P with $X_1 = x_1, X_2 = x_2, X_3 = x_3$ is n, then the corresponding entry in the 3-D reference pdf has the value $p_3(x_1, x_2, x_3) = n/N_P$, and similarly for all reference pdfs. The reference pdfs are normalized since they contain fractional occupancy of each bin. An important consequence of using a finite set of samples to populate the reference pdfs is that if a certain combination of variable values is absent from Ω_P then it is assigned a zero probability in the reference distribution, even though the corresponding entry in the exact marginal of p_N might be non-zero. The consequences of zeros, or "holes", in the reference distributions on the output of the present sampling algorithms are discussed below (item 5).

Mathematical and computational properties of the SA-l based ancestral sampling algorithms are discussed next. In view of application to conformational sampling later, we sometimes refer to the N-dimensional discrete space as the conformational space, and to a point in the space as a conformation.

(i) **Sampling distribution**

The distribution sampled by the l-level sampling algorithm, denoted by $\tilde{p}_N^{(l)}$, is obtained by substituting the conditional pdf (Eqs. 3.10 and 3.11), from which each variable is sampled, into the chain rule of Eq. 3.1. Thus, the doublet level algorithm samples from the distribution

$$\tilde{p}_N^{(2)} = p(X_1, X_2)p^{(2)}(X_3|X_1, X_2) \times \ldots \times p^{(2)}(X_N|X_1, \ldots, X_{N-1}) \tag{3.13}$$

and the triplet-level algorithm from

$$\tilde{p}_N^{(3)} = p(X_1, X_2, X_3)p^{(3)}(X_4|X_1, X_2, X_3) \times \ldots \times p^{(3)}(X_N|X_1, \ldots, X_{N-1})\,. \tag{3.14}$$

Although the closed form expression of the $\tilde{p}_N^{(l)}$ distribution is complicated, its value can be readily be obtained numerically for each sampled conformation. This will be important when it comes to using the present algorithm for free energy calculations (Chapter 4).

(ii) **The sampling distribution, $\tilde{p}_N^{(l)}$, is normalized**

The sampling distribution is automatically normalized since all the conditional pdfs used in the chain rule are normalized. This can be seen as follows:

$$\begin{aligned}\sum_{X_1}\cdots\sum_{X_N}\tilde{p}_N^{(l)} &= \sum_{X_1}\cdots\sum_{X_N}\left(\begin{array}{c} p(x_1,...,x_l)\;\; p^{(l)}(x_{l+1}|x_1,...,x_{l-1})\times\ldots\times \\ p^{(l)}(x_{N-1}|x_1,\ldots,x_{N-2})p^{(l)}(x_N|x_1,\ldots,x_{N-1})\end{array}\right) \\ &= \sum_{X_1}\cdots\sum_{X_{N-1}}\left(\begin{array}{c} p(x_1,...,x_l)\;\; p^{(l)}(x_{l+1}|x_1,...,x_{l-1})\times\ldots\times \\ \underbrace{\left(\sum_{X_N}p^{(l)}(x_N|x_1,\ldots,x_{N-1})\right)}_{=1}\end{array}\right) \end{aligned} \tag{3.15}$$

where the innermost summation over the conditional pdf of X_N is one due to normalization. Similarly, successive conditional pdfs are summed out, finally giving the summation over the highest order reference pdf which is one by construction:

$$\begin{aligned}\sum_{X_1}\cdots\sum_{X_N}\tilde{p}_N^{(l)} &= \sum_{X_1}\cdots\sum_{X_l}p(x_1,...,x_l) \\ &= 1\,. \end{aligned} \tag{3.16}$$

Therefore, the sampling distributions are normalized.

(iii) **Relationship between the l-level sampling distribution and the SA-l**

The sampling distribution at level l, $\tilde{p}_N^{(l)}$, is different from the N-dimensional SA-l distribution, $p_N^{(l)}$. One important difference is that the sampling distribution is normalized, whereas the SA-l, in general, is not. To further appreciate the relationship between these two distributions, it is instructive to compare their

analytic expressions for the simple case of a three dimensional distribution at the doublet level. The SA-2 for $N = 3$ is

$$p_3^{(2)}(X_1, X_2, X_3) = \frac{p(X_1, X_2)p(X_1, X_3)p(X_2, X_3)}{p(X_1)p(X_2)p(X_3)} \tag{3.17}$$

and the corresponding sampling distribution is

$$\tilde{p}_3^{(2)}(X_1, X_2, X_3) = \frac{p(X_1, X_2)p(X_1, X_3)p(X_2, X_3)}{p(X_3)} \frac{1}{\sum_{\bar{X}_3} \frac{p(X_1, \bar{X}_3)p(x_2, \bar{X}_3)}{p(\bar{X}_3)}} \tag{3.18}$$

where $\bar{X}_3$ is a dummy variable for X_3. From the above expressions it appears that the sampling distribution does not include all reference marginal pdfs. However, comparing Eq. 3.18 with Eq. 3.17, the pdfs absent from $\tilde{p}_3^{(2)}$, $p(X_1)$ and $p(X_2)$, are marginals of the included 2-D pdfs, $p(X_1, X_2)$, $p(X_1, X_3)$ and $p(X_2, X_3)$. Therefore, if the 2-D pdfs are non-zero, $p(X_1)$ and $p(X_2)$ will also be non-zero, and for any conformation with a non-zero sampling probability the corresponding SA-l probability will also be non-zero. An important implication of this overlap between the sampling distribution and the corresponding SA-l distribution is that, due to the product form of the SA-l, any conformation with non-zero $\tilde{p}_N^{(l)}$ necessarily falls in the non-zero bins of each reference pdf.

Comparison of Eq. 3.17 with Eq. 3.18 brings out another distinction between the sampling distribution and the SA-l distribution – $p_3^{(2)}$ is symmetric in all three variables but $\tilde{p}_3^{(2)}$ in Eq. 3.18 is symmetric in only X_1 and X_2. In general, due to the use of the approximate conditional distributions, $\tilde{p}_N^{(l)}$ is symmetric in only the first l sampled variables; the remaining asymmetry leads to a dependency of $\tilde{p}_N^{(l)}$ on the order in which the variables are sampled, in contrast to the SA-l which is symmetric in all variables.

(iv) **Zeros in reference pdfs may lead to null samples**

If the reference pdfs have holes then the doublet and higher level sampling algorithms can fail to yield a complete set of variables in a sampling iteration. Consider the following SA-3 based conditional distribution for X_4 given values of the first three variable:

$$p^{(3)}(X_4|x_1,x_2,x_3) = \frac{p(X_4)p(x_1,x_2,X_4)p(x_1,x_3,X_4)p(x_2,x_3,X_4)}{p(x_1,X_4)p(x_2,X_4)p(x_3,X_4)} \frac{1}{N_4\left(x_1,x_2,x_3\right)} . \tag{3.19}$$

In order to be able to sample X_4, the conditional probability for at least one out of the B possible values of X_4 must be non-zero. This may not hold if, for each value of X_4, one or more of the pdfs in Eq. 3.19 are zero. In practice, a sampling iteration is terminated if this null condition is encountered, and a new iteration is begun. The possibility of generating a null sample is likely to decreases as more data is used to populate the reference marginals.

(v) **Reference pdfs restrict the accessible conformational space**

As discussed in the previous two items, any conformation sampled from the l-level sampling algorithm necessarily falls in the non-zero bins of *every* reference pdf used, thereby restricting the conformational space accessible to the sampling algorithms. The accessible region of conformational space can shrink as higher-order reference pdfs are incorporated, since these may include zero-probability bins not present in the lower-order pdfs. Indeed, this will often be the case, because higher-order pdfs have more bins and are therefore require more data to be adequately populated. Therefore, the conformational region accessible to singlet level sampling will typically be larger than that accessible to doublet level sampling, which in turn will be larger than at the triplet-level, and so on. Denoting the set of conformations accessible to l-level

sampling by $\Omega^{(l)}$, we have

$$\Omega \supseteq \Omega^{(1)} \supseteq \Omega^{(2)} \supseteq \Omega^{(3)} \supseteq \ldots \supseteq \Omega_P \tag{3.20}$$

where Ω is the set of all possible B^N conformations in the discrete state-space. Note that the points used to populate the reference distributions, Ω_P, by construction are associated with bins having non-zero probability in all reference distributions and, therefore, will fall in regions accessible to sampling at all orders. In other words, although the accessible region may be restricted, it will always be at least as large as the region represented by the original samples from the true distribution p_N. Indeed, due to the omission of higher level marginals, the sampling algorithms can generate conformations which are distinct from those used to populate the reference pdfs. Note that since conformations in Ω_P are used to populate the reference pdfs, the accessible region $\Omega^{(l)}$, in turn, is influenced by Ω_P.

(vi) **Computational cost and memory requirement**

The computational cost of the l-level sampling algorithm for generating a single sample in N-dimensions scales as $O(N^l)$. At the singlet level ($l = 1$), all variables are sampled independently from their singlet pdfs, so the cost is simply N times the cost of sampling a single variable, giving a linear scaling with N. Sampling at higher levels has the additional cost of computing the conditional pdfs, which, for the last variable sampled, involves a product of all the reference pdfs. At the doublet level, since there are $O(N)$ singlet pdfs and $O(N^2)$ doublet pdfs, the number of multiplication operations is $O(N^2)$, which implies quadratic scaling of the computational cost. Similarly, the computational cost of triplet-level sampling scales as $O(N^3)$. All algorithms scale linearly with respect to B, the number of discrete

values for each variables.

The present MATLAB [51] implementation took 0.0017, 0.04 and 0.6 seconds on a 3.8 GHz Pentium 4 PC for an $N = 32, B = 30$ test case at the singlet, doublet and triplet levels, respectively. We anticipate that the computer time could be substantially reduced by reimplementation of the algorithm in a lower level language, such as C or Fortran, and by straightforward code optimizations. Furthermore, the wall-clock time required to generate a given number of samples would be reduced substantially by distributing the computation on multiple parallel nodes, especially since the sampling iterations are independent, so that no internode communication would be required. In other words, the sampling can be done in an embarrassingly parallel fashion.

The reference pdfs of order l are stored as l-dimensional matrices with B^l elements. Since there are C_l^N pdfs at order l, the total numbers of elements in all doublet and triplet reference distributions are, respectively, $(N(N-1)/2)B^2$ and $[(N(N-1)(N-2))/6]B^3$. In general, the storage required for l-order pdfs is $O(N^l B^l)$, and these numbers can become large for molecular systems. For example, in this work, the largest molecular system considered at the doublet level (tetra-alanine in Chapter 4) has $N = 150$ and $B = 30$ and required storage of $\sim 10^7$ numbers for the doublet distributions; and the largest system at the triplet level (host-guest complex in this Chapter) has $N = 32, B = 30$ which corresponds to $\sim 10^9$ bins requiring about 1 GB of memory at double precision (8 byte floats). However, the storage requirement was reduced by as much as six-fold by using a sparse matrix representation of the pdfs, where only the non-zero entries are stored.

(vii) **Numerical implementation of conditional pdfs**

The conditional probability distribution constructed at each step of the SA-l based sampling algorithms (Eqs. 3.10 and 3.11) requires the product of the probability values from the reference distributions. The number of factors in this product becomes large as the number of variables increase. Direct multiplication of these factors can lead to underflow errors, because the probabilities that are multiplied are less than one. This problem is solved by taking the logarithms of the factors and adding them, instead of multiplying the factors themselves.

3.3 Application of SA-l based sampling to molecular systems

In this section, the SA-l based sampling algorithms developed above are applied to the problem of conformational sampling for molecules. The dimensionality of the conformational space, N, is the number of internal coordinates of the molecule; the true distribution, p_N, is the Boltzmann distribution corresponding to a molecular mechanics force-field (Eq. 1.1); and the reference distributions are populated using MD simulation data. The continuous conformational space sampled by the MD simulations is discretized to compute the reference distributions as detailed below. We wish to determine whether the molecular conformations sampled using the low-order reference distributions are distributed similarly to the original MD conformations used to populate the reference pdfs.

Following is an overview of the overall computational protocol followed here:

(i) Run a constant temperature MD simulation of the molecule of interest (Section 3.3.2), saving N_P conformations.

(ii) For each saved snapshot of the MD trajectory, extract the bond-angle-torsion (BAT) and Anchored Cartesian (XYZ) internal coordinates (Section 3.3.1); discretize each coordinate, ξ_i, into equally spaced bins of width, $\delta_i = \Delta_i/B$, where $\Delta_i = \xi_{i,\max} - \xi_{i,\min}$ is the difference between the maximum and minimum observed value.

(iii) Map the continuous space BAT and XYZ conformations to the discrete space where a conformation, $\mathbf{X} = (X_1, \ldots, X_N)$, is specified by N integers in $\{1, .., B\}$ denoting the bin number for each coordinate.

(iv) Construct normalized histograms of the discrete space MD coordinates to obtain the first-, second- and third-order reference pdfs for both coordinate systems (Section 3.3.3).

(v) Use these reference pdfs in the SA-l based sampling algorithms (Section 3.2) to generate N_R samples, each representing a conformation in the discretized conformational space.

(vi) For each reference sample, map the sampled bin numbers to real values of the internal coordinates by assigning the center-of-bin values to each coordinate as

$$\xi_i(X_i) = \xi_{i,\min} + (X_i - 1/2)\delta_i \tag{3.21}$$

and reconstruct the three-dimensional conformation of the molecule.

(vii) Compare the distributions of internal coordinates, conformations and force-field energies obtained by SA-l sampling to those from the original MD run (Section 3.3.4).

3.3.1 Internal coordinate systems for molecules with branched topologies

The MD conformations are mapped to an internal coordinate system to remove three translational and three rotational degrees of freedom that are not involved in the Hamiltonian for the field-free systems considered here. The correlations among the internal coordinates, as captured by the reference pdfs, depend on the specific internal coordinate system used. To assess the impact of the choice of coordinate system, we examine both the bond-angle-torsion (BAT) and anchored Cartesian (XYZ) coordinate systems [52, 53].

The XYZ system is defined in terms of three root atoms, and the molecule is oriented such that atom 1 is at the origin, atom 2, which is bonded to atom 1, is on the positive x-axis, and atom 3 is in the $x-y$ plane, thereby fixing six Cartesian coordinates to zero. In BAT coordinates, the conformation of an M-atom molecule is given by $M-1$ bond-lengths, $M-2$ bond-angles, and $M-3$ torsions. Note that both the XYZ and BAT coordinate systems are non-unique, in the sense that the coordinates depend upon the choices of root atoms and, for BAT coordinates, the choices of bond-angles and torsions selected as internal coordinates and the treatment of so-called phase angles [54]; the coordinate setups used for the molecules studied here are described in the next section. The correlation among the coordinates is expected to be less in the BAT system because the bonded energy terms in the force-field used for MD simulation are defined in terms of the BAT coordinates, and the natural circular motions of atoms associated with torsional fluctuations are naturally handled in BAT coordinates. The XYZ system, on the other hand, is more convenient for computing the force-field energy of molecules since most software implementations of the force-fields require the input to be in Cartesian coordinates. It is straightforward to map back and forth between XYZ and BAT coordinates.

3.3.2 Molecular test systems

Three molecular systems are studied: nonane, cyclohexane, and a small host-guest complex [55, 56]. Figure 3.2 diagrams the chemical structures of these molecules and the subsets of atoms used for testing the sampling algorithms. For each system, 5 million MD snapshots spanning 50 ns of a single MD simulation were processed to generate the reference marginals. The MD trajectories were those used in a previous study of configurational entropy from our group [48]. The MD simulations use an all-hydrogen CHARMM force-field [57] and approximate the effects of solvent with a simple distance-dependent dielectric model [58], $D_{ij} = 4r_{ij}$, where r_{ij} is the distance in angstroms between atoms i and j.

In order to reduce the computational cost, internal coordinates with very narrow distributions, such as dihedrals within a phenyl ring, which are not expected to significantly influence the overall conformation of the molecule are not sampled. Instead, these coordinates are held fixed at their equilibrium values established by the force-field. The coordinates for which we compute reference marginals, so that they contribute to the SA-l based sampling, are termed "active". For all three molecules, atoms 1, 2 and 3 in Figure 3.2 are the root atoms for both the XYZ and BAT coordinate systems. Because we use united-atom representations of the molecules, there are $N = 12$ internal coordinates for cyclohexane, (for the BAT coordinates, these comprise 5 bond-lengths, 4 bond-angles and 3 torsions) and $N = 21$ internal coordinates for nonane (for BAT coordinates, these comprise 8 bond-lengths, 7 bond-angles and 6 torsions), all of which were treated as active. The labeling of the BAT coordinates for nonane and cyclohexane is as follows: i-th bond is between atoms i and $i + 1$; the i-th angle is between atoms i,$i + 1$ and $i + 2$; and the i-th torsion is the angle between the plane of atoms $(i, i + 1, i + 2)$ and the plane of atoms $(i + 1, i + 2, i + 3)$. (See also Figure 4.1 which illustrates the two coordinate systems for

propane.)

Nonane and cyclohexane conformations were sampled in both BAT and Cartesian coordinates at all three levels: singlet, doublet and triplet. The sampling sequence in XYZ coordinates follows the atom numbering of Figure 3.2 where, for each atom, the x-coordinate is sampled first, followed by the y- and z-coordinates. In BAT coordinates, for nonane and cyclohexane, the bond-length coordinates of all atoms in their indexed order were sampled first, followed by angles and torsions in the same order.

The dimensionality of the host-guest system was reduced by limiting attention to a skeleton of 23 atoms (numbered atoms in Figure 3.2) out of the simulated 56-atom complex. The retained atoms correspond to 63 ($= 3 \times 23 - 6$) internal degrees of freedom. Out of the 63 BAT coordinates, 32 coordinates were treated as active: all bond-angle and torsional degrees of freedom except the ones in the rings, along with one pseudo-bond, two pseudo-angles and three pseudo-dihedrals [47] that together specify the position and orientation of the guest with respect to the host. Table 3.1 lists the active degrees of freedom as well as the equilibrium values of the inactive ones. Three torsion angles (indicated in Table 3.1) that might be viewed as flexible are treated as phase angles [54] of flexible torsions that share the same rotatable bond; these phase angles have narrow distributions and are therefore treated as inactive. For the host-guest system, conformations were sampled only in the BAT system at the singlet, doublet and triplet levels, following the order listed in Table 3.1.

3.3.3 Calculation of reference marginal pdfs

For each active coordinate, the continuously varying coordinate values from the MD simulations were discretized by setting up $B = 30$ equally spaced bins between the minimum and maximum value observed in the MD data, except in the case of the torsional

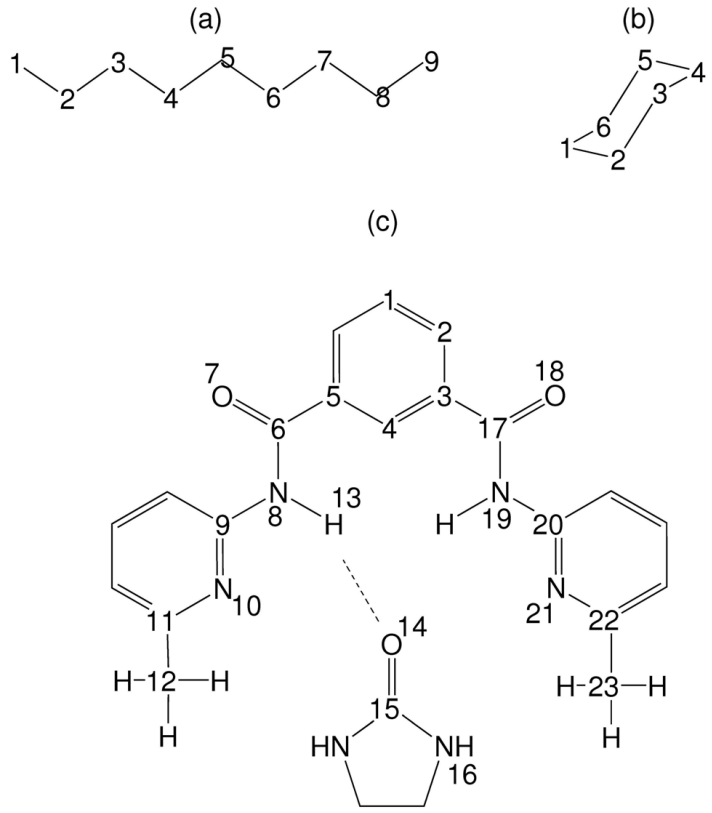


Figure 3.2: Molecular systems used for testing the SA-l based sampling algorithms: (a) Nonane, (b) Cyclohexane and (c) Host-guest complex. Atoms included in the sampled structure are numbered. The internal coordinate system is set up such that atom 1 is on the origin, the bond between atom 1 and 2 is along the x-axis and atom 3 is in the x-y plane. For nonane and cyclohexane, only the carbon chain is sampled, although the MD simulations included all hydrogens. For the host-guest complex, a dotted line represents the pseudo-bond used in defining the position and orientation of the guest relative to the host.

coordinates of cyclohexane. To effectively capture the bimodal distributions of the torsional coordinates of cyclohexane, 15 equally spaced bins were used in the two intervals – $[0, 220.3]$ and $[497.5, 720]$ degrees – for a total of 30 bins. In the discrete space a continuous BAT or XYZ coordinate value maps to an integer denoting the number of the bin to which the value belongs. The reference singlet, doublet and triplet distributions are then constructed as normalized histograms as described in the computational protocol above.

Table 3.1: List of BAT degrees of freedom corresponding to the 23-atom skeleton of the host-guest complex. All 22 bonds, 21 angles and 20 torsions are listed using the atom numbers shown in Figure 3.2. Active variables are indicated by A. For the inactive variables, the equilibrium values based on the force-field are listed.

Bond	Bond Equilibrium Value (Å)	Angle	Angle Equilibrium Value (Rad)	Torsion	Torsion Equilibrium Value (Rad)
1-2	1.383	1-2-3	2.0944	1-2-3-4	0
2-3	1.383	2-3-4	2.0944	1-2-3-17	A
3-4	1.383	2-3-17	A	2-3-4-5	0
3-17	1.46	3-4-5	2.0944	2-3-17-18	Phase of 2-3-17-19
4-5	1.383	3-17-19	A	2-3-17-19	A
5-6	1.46	3-17-18	A	3-17-19-20	A
6-8	1.345	4-5-6	A	3-4-5-6	A
6-7	1.225	5-6-8	A	4-5-6-7	Phase of 4-5-6-8
8-9	1.355	5-6-7	A	4-5-6-8	A
8-13	1.0	6-8-9	A	5-6-8-13	A
9-10	1.327	6-8-13	A	5-6-8-9	Phase of 5-6-8-13
10-11	1.327	8-9-10	A	6-8-9-10	A
11-12	1.5	8-13-14	A	6-8-13-14	A
13-14	A	9-10-11	2.0159	8-9-10-11	A
14-15	1.225	10-11-12	A	8-13-14-15	A
15-16	1.345	13-14-15	A	9-10-11-12	A
17-19	1.345	14-15-16	A	13-14-15-16	A
17-18	1.225	17-19-20	A	17-19-20-21	A
19-20	1.355	19-20-21	A	19-20-21-22	A
20-21	1.327	20-21-22	2.0159	20-21-22-23	A
21-22	1.327	21-22-23	A	-	-
2-23	1.5	-	-	-	-

3.3.4 Evaluation of sampled conformations

In order to assess the contributions of successively higher-order correlations, distributions of conformations generated via sampling at the singlet, doublet and triplet levels were compared to the distributions of the MD conformations. Three types of comparisons were done.

First, the samples from the SA-l based sampling were used to compute one-, two- and three-dimensional pdfs, just as done for the MD conformations. These pdfs, which represent the marginals of the sampling distribution, were compared with the reference marginal pdfs obtained directly from MD. The difference between a sampled distribution and the corresponding reference distribution from MD is reported as the root mean square deviation (RMSD) across the bins of the pdfs.

Second, the sampled conformations were compared with the MD conformations by comparing the distributions of energies and key intramolecular distances. Doing this requires reconstructing the molecular conformation associated with a sample in the discrete space of bin numbers for each active coordinate. Because only the active internal coordinates are sampled, it was necessary to modify the MD snapshots, in which all coordinates fluctuate, so that they would be on a comparable footing. This was done by extracting the active internal coordinates from the MD snapshots, and substituting the equilibrium values for the inactive coordinates, precisely as done for the sampled conformations. Then the MD conformations were reconstructed with these idealized coordinates, and compared with the sampled conformations. For the host-guest complex, distributions were compared for multiple interatomic distances that characterize the conformations. For nonane, the distance between the terminal carbons was examined. For cyclohexane, the distance between carbons 1 and 6 was examined; this distance is

not part of the BAT coordinate system and depends upon the values of the internal coordinates in the same way that nonane's end-to-end distance depends upon its internal coordinates. Note that, in the XYZ coordinate system, the end-to-end distances for both nonane and cyclohexane are, in principle, functions of only the Cartesian coordinates of the last carbon. Nonetheless, since these coordinates are the last three variables to be sampled, the distributions of end-to-end distances in the sampled conformations depend on all sampled XYZ coordinates.

Third, the distributions of energy for the sampled conformations, computed with the same CHARMM [57] force-field model, were compared with those for the reconstructed (see above) MD conformations. Comparisons were made for both the total molecular energy and the separate terms provided by the force-field. The separate terms provide additional physical insight; for example, if the sampled conformations were to yield more conformations with high Lennard-Jones energies, this would imply steric clashes.

3.4 Results

For nonane and cyclohexane, 500,000 conformations were sampled, while for the host-guest complex, 200,000 conformations were sampled. Convergence was established based on the medians of energy and the sampled interatomic distances; a representative convergence plot of median total energy for the host-guest system is shown in Figure 3.3. Similar convergence is obtained for nonane and cyclohexane.

The first three subsections here assess the accuracy of conformations sampled at the singlet, doublet and triplet levels by studying the geometries and energies of the sampled conformations from the superposition approximations. The fourth subsection compares marginal distributions of the superposition approximations with the

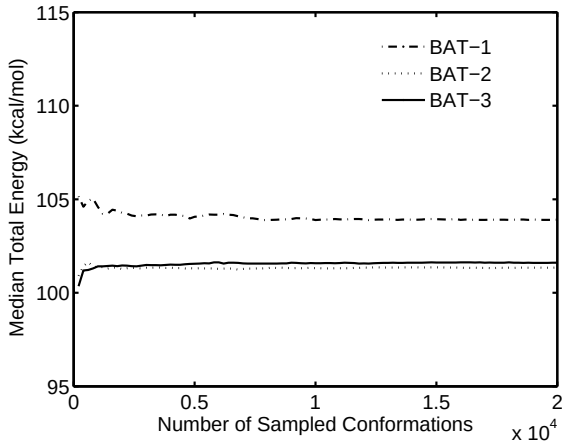


Figure 3.3: Convergence of median total energy as a function of the number of conformations sampled in the BAT coordinate system at the singlet (BAT-1), doublet (BAT-2) and triplet (BAT-3) levels, for the host-guest complex. The means and standard deviations of the energy converge similarly, and similar results are obtained for the other molecular systems in this study.

corresponding reference marginals.

Here BAT-1, BAT-2 and BAT-3 refer to conformational sampling in BAT coordinates at the singlet, doublet and triplet levels respectively, while XYZ-1, XYZ-2 and XYZ-3 refer to sampling in Cartesian coordinates at the corresponding levels of approximation. As noted in Section 3.2.4, the distributions of the samples from the doublet and triplet level algorithms depend upon the sequence in which the variables are sampled. Changing the sequence of sampling, such as by sampling torsions first versus sampling bond-lengths first, when BAT coordinates are used, was found to alter the results in detail, but had little effect on the overall accuracy of the sampled distributions. As discussed in Section 3.2.4, the doublet and triplet level sampling may generate null samples. For the three molecular systems studied here, such null samples never occurred for the doublet level algorithm. At the triplet level, they occurred for $< 0.01\%$ of the iterations for nonane and the complex, and never occurred for cyclohexane.

A note on the presentation of results for the three molecules – all tables and figures corresponding to a molecule are collected following the subsection of the molecule and the tables describing the statistics of a distribution immediately follows the figure with the distribution.

3.4.1 Nonane

Reference distributions were computed in both XYZ and BAT coordinates and used to sample at all three levels in both coordinate systems. Thus, conformations are sampled based on a total of 6 sampling algorithms. The probability distributions of end-to-end distances (carbon 1 to carbon 9) from MD and from the six sampled sets are compared numerically in Table 3.2, and Figure 3.4 graphs the corresponding distributions.

For BAT coordinates, both the doublet and triplet-level superposition approximations provide excellent agreement with the MD results, and are markedly more accurate than singlet level sampling. In particular, Figure 3.4 shows that doublet and triplet level sampling produces a substantially smaller fraction of conformations with excessively short end-to-end distances than does singlet level sampling. This is also evident from the shorter minimum distance for singlet sampling (0.09 Å) as compared to doublet (0.24 Å) and triplet sampling (0.67 Å) (Table 3.2). The triplet level is slightly more accurate than the doublet, but the difference is less striking than that between doublet and singlet.

For Cartesian coordinates, all three sampled cases yield distance distributions that deviate markedly from the MD distributions: although the numerical statistics in Table 3.2 look reasonable, Figure 3.4 shows that the shapes of the distributions are poor. As with sampling in BAT coordinates, the population in the short-distance end of the distribution is notably higher for the singlet than for the doublet or triplet level samples. Interestingly,

in Cartesian coordinates the triplet-level approximation does not appear to yield greater accuracy than the doublet level distribution, at least by the present measure.

The BAT coordinates samples of nonane yield distributions of total energy that agree very well with MD overall at the doublet and triplet levels, as shown in Figure 3.5a. Interestingly, the total energy distribution at the singlet level is only slightly inferior to those at higher levels, indicating low correlations among various BAT internal coordinates. However, the energy distributions from superposition approximations in Cartesian coordinates disagree strongly with the reference MD distribution: they are wide and shifted to much higher energies, as shown in Figure 3.5b. The singlet level results are particularly poor, as the minimum total energy among all conformations is 307.1 kcal/mol (see Table 3.3), which is outside the range of energies in Figure 3.5b. The fraction of such high energy conformations is lower for doublet and triplet level samples. It is not surprising that the energy distributions from Cartesian sampling are poor, given the similarly poor end-to-end distance distributions described above.

The distributions of the total energy of nonane for the BAT samples show excellent agreement with the MD distributions especially at the doublet and triplet levels (in Figure 3.5a,b). This is consistent with the good agreement between the median total energy of the sampled conformations with the MD conformations (Table 3.3). The larger deviations for the mean energy and other statistics result from the small fraction of high energy conformations among the sampled conformations, as evident from Figure 3.5b and Table 3.4. Closer examination of Table 3.3 indicates that the bond-stretch and bond-angle energies are well-behaved, but van der Waals energies are sometimes much too large. This indicates that the high-energy conformations among the BAT samples result from excessively close atom-atom contacts, consistent with the small values of the minimum

end-to-end distances listed in Table 3.2. In contrast, the data in Table 3.3 show that the large errors in the energy distributions based upon XYZ coordinates result not only from van der Waals overlaps, but also from severe errors in bond-lengths and bond-angles. The superior performance of the BAT coordinates, relative to Cartesian coordinates, is traceable to the fact that the BAT coordinates do an excellent job of capturing the marginal pdfs of those bond-lengths and angles which are included in the coordinate set. Thus, the energetics of these stiff degrees of freedom are well reproduced.

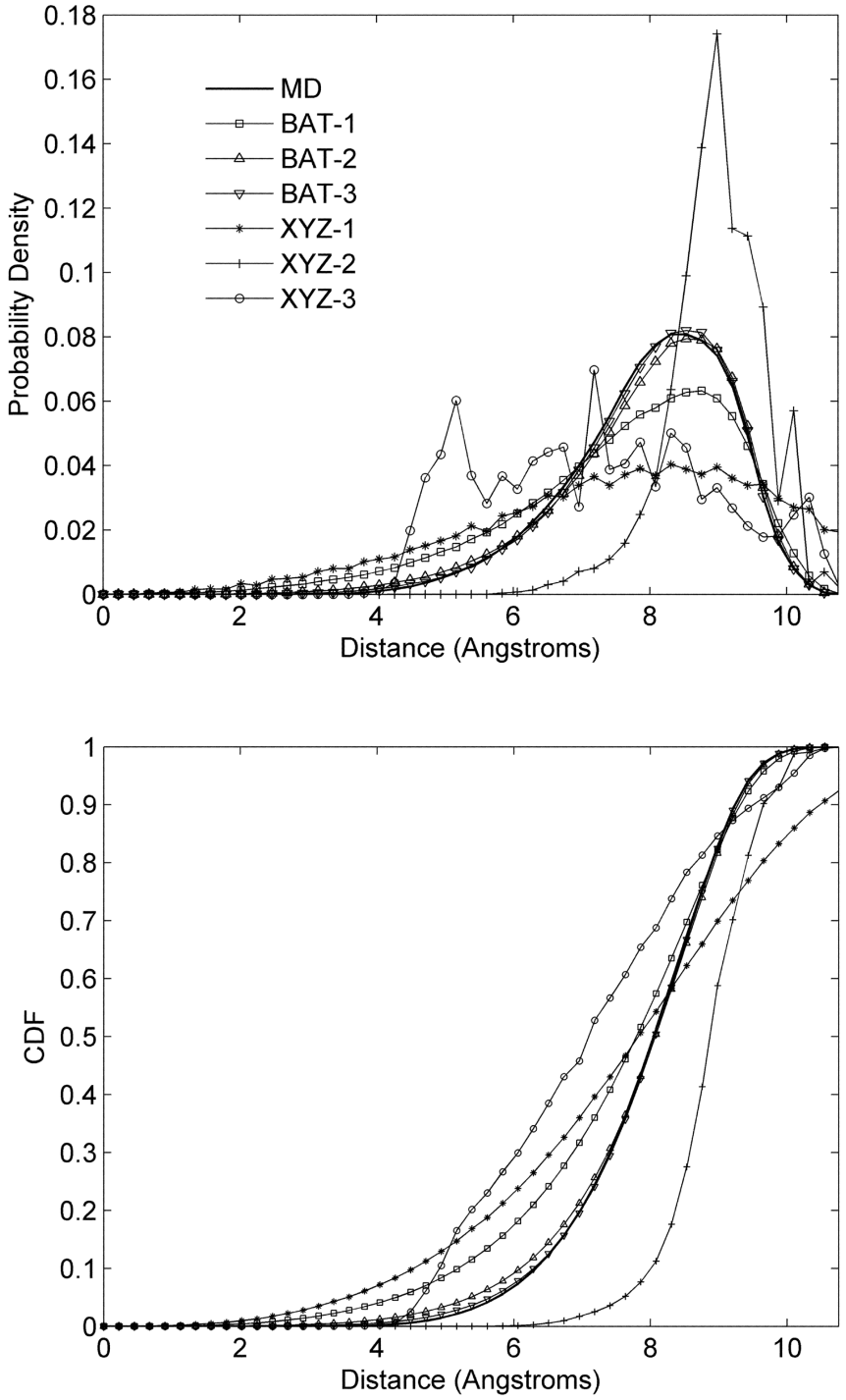


Figure 3.4: Probability distributions (top) and corresponding cumulative distributions (bottom) of end-to-end distances for nonane, from MD conformations and sampling algorithms at singlet, doublet and triplet levels in BAT and XYZ coordinates.

Table 3.2: Statistics of end-to-end distances (Å) for nonane from MD and sampled conformations.

	Median	Mean	Standard Deviation	Minimum	Maximum
MD	8.16	8.01	1.14	2.99	10.98
BAT-1	7.91	7.60	1.64	0.09	10.89
BAT-2	8.18	7.96	1.29	0.24	10.80
BAT-3	8.18	8.01	1.18	0.67	10.85
XYZ-1	7.93	7.76	2.28	0.31	14.84
XYZ-2	8.98	8.96	0.69	3.87	11.08
XYZ-3	7.23	7.24	1.66	3.05	10.97

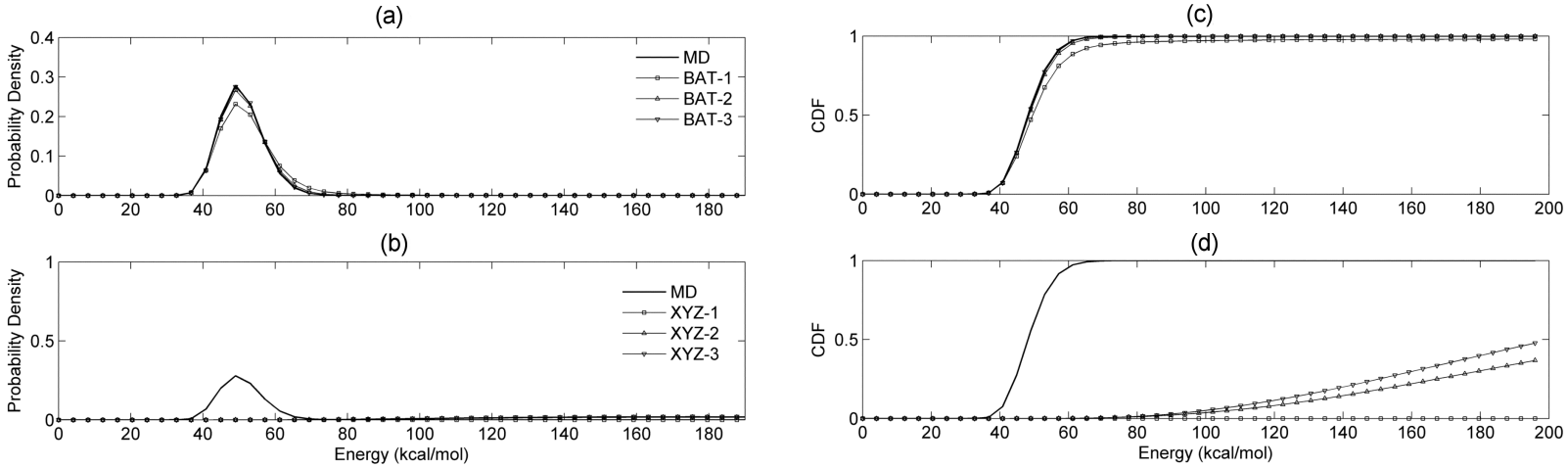


Figure 3.5: Probability distributions and corresponding cumulative distributions (right) of total energy of nonane based upon MD conformations and conformations sampled in (a,c) BAT and (b,d) XYZ coordinate systems at singlet, doublet and triplet levels.

Table 3.3: Statistics of energy distributions (kcal/mol) from MD and sampled conformations for nonane.

	Bond	Angle	Torsion	Coulomb	VDW	Total
				Median		
MD	7.7	4.4	3.23	34.5	-0.63	50.2
BAT-1	7.8	4.5	3.21	34.9	-0.57	51.5
BAT-2	7.8	4.5	3.22	34.4	-0.62	50.6
BAT-3	7.8	4.5	3.22	34.4	-0.63	50.4
XYZ-1	1.8×10^4	318.4	4.91	36.0	15.33	2.2×10^4
XYZ-2	141.2	48.6	2.78	31.7	-0.52	232.3
XYZ-3	114.4	41.2	3.85	34.7	-0.61	202.8
				Mean		
MD	8.34	4.9	3.22	34.8	-0.6	50.7
BAT-1	8.49	5.0	3.23	36.3	3.6×10^3	3.6×10^3
BAT-2	8.48	5.0	3.23	35.0	182.6	234.3
BAT-3	8.50	5.0	3.22	34.8	11.5	63.0
XYZ-1	2.1×10^4	325.8	4.90	37.7	9.7×10^4	1.2×10^5
XYZ-2	171.93	55.6	2.86	32.1	179.7	442.2
XYZ-3	136.75	47.7	3.84	35.1	252.2	475.6
				Standard Deviation		
MD	4.15	2.6	1.00	2.46	0.3	5.9
BAT-1	4.22	2.7	1.11	5.59	1.2×10^5	1.2×10^5
BAT-2	4.22	2.7	1.03	3.20	2.6×10^4	2.6×10^4
BAT-3	4.26	2.7	1.00	2.66	4.2×10^3	4.2×10^3
XYZ-1	1.21×10^4	120.0	1.14	7.94	6.5×10^5	6.5×10^5
XYZ-2	120.9	26.2	0.92	1.85	2.4×10^4	2.4×10^4
XYZ-3	93.4	31.9	1.10	3.22	3.3×10^4	3.3×10^4
				Minimum		
MD	0.04	0.02	0.01	28.1	-1.49	31.7
BAT-1	0.16	0.02	0.00	28.2	-1.57	32.2
BAT-2	0.17	0.02	0.00	28.4	-1.51	32.0
BAT-3	0.31	0.10	0.00	28.2	-1.53	32.8
XYZ-1	1.9×10^2	0.69	0.48	22.6	-1.35	307.1
XYZ-2	0.88	0.47	0.06	27.6	-1.01	43.1
XYZ-3	2.17	0.23	0.00	27.6	-1.42	44.0
				Maximum		
MD	45.6	34.5	7.64	49.0	5.06	100.2
BAT-1	49.1	28.1	7.88	254.4	9.9×10^6	9.9×10^6
BAT-2	45.7	27.4	7.87	130.7	9.0×10^6	9.0×10^6
BAT-3	42.0	23.7	7.22	80.5	2.5×10^6	2.5×10^6
XYZ-1	1.1×10^5	955.5	9.56	352.4	9.9×10^6	$\times10^7$
XYZ-2	1.5×10^3	452.4	7.44	84.7	9.9×10^6	9.9×10^6
XYZ-3	1.2×10^3	386.5	8.52	80.6	8.9×10^6	8.9×10^6

Table 3.4: Fraction of sampled conformations with energies greater than the maximum energy obtained in MD simulation of each test system.

	Number of conformations sampled	Number of high-energy conformations	Fraction of high-energy conformations
		Nonane	
BAT-1	500,000	14,500	2.9×10^{-2}
BAT-2	500,000	1,450	2.9×10^{-3}
BAT-3	500,000	170	3.4×10^{-4}
XYZ-1	500,000	500,000	1.00
XYZ-2	500,000	480,000	0.96
XYZ-3	500,000	475,000	0.95
		Cyclohexane	
BAT-1	500,000	401,539	0.80
BAT-2	500,000	192,204	0.38
BAT-3	500,000	48,680	0.09
XYZ-1	500,000	488,561	0.98
XYZ-2	500,000	112,324	0.22
XYZ-3	500,000	12,392	0.02
		Host Guest Complex	
BAT-1	200,000	22,000	0.11
BAT-2	200,000	2,000	0.01
BAT-3	200,000	1,000	0.005

3.4.2 Cyclohexane

As for nonane, we studied the singlet, doublet and triplet level samples in XYZ and BAT coordinates, for a total of six sets of samples. Results are assessed geometrically in terms of the distribution of distances between carbon 1 and carbon 6, the only two successive atoms in the ring whose bond-length is not part of the BAT coordinate system. As shown in Figure 3.6 and further detailed in Table 3.5, here the Cartesian coordinate system yields a more accurate distribution than does the BAT coordinate system, and the triplet-level distribution is more accurate than the doublet- level, which in turn is substantially better than singlet level distribution. The improvement upon including correlations in much more apparent in the distance distributions for samples in BAT coordinates. In BAT samples, the singlet level distributions deviate drastically from the MD results, but the doublet and triplet level distributions are similar in structure to those from MD, being unimodal and centered at roughly the same bond-length. Thus, for cyclohexane, higher order correlations are required to accurately capture the geometry, in both BAT and Cartesian coordinates. In absolute terms, the distribution of end-to-end distances from sampling in BAT coordinates for cyclohexane is more accurate than the distribution of end-to-end distances from sampling in Cartesian coordinates for nonane.

The MD trajectory used to compute the reference marginals includes chair, boat, and twist-boat conformations. In BAT coordinates, these conformations are established by the three internal torsions. Figure 3.7 compares representative conformations from MD and BAT sampling at the three levels. The BAT-3 results resemble MD closely, and the BAT-2 results are similar, though somewhat less accurate. However, many of the BAT-1 conformations are quite distorted. The reason for the poor BAT-1 conformations has to do with the fact that singlet level sampling completely ignores correlations, so that the

effective 2-D pdf linking each pair of torsions is just the product of their respective 1-D pdfs. Figure 3.8 (left) shows that the correct 2-D pdf of a pair of torsions has two maxima, corresponding to two different chair conformations.

The corresponding singlet distributions (Figure 3.8, middle) also have two maxima, so their product (Figure 3.8, right) has four, rather than two maxima. Two of the four maxima correspond to the correct maxima seen in the reference 2-D pdf (Figure 3.8, left); the other two are artifacts of singlet level sampling and produce distorted conformations. Table 3.6 summarizes the statistics of energy components computed from sampled conformations and reference MD conformations. On the whole, in both BAT and XYZ coordinate systems, triplet level samples are closer to MD values than doublet level samples. At the triplet level, the statistics for the bond energies from Cartesian sampling match the reference MD values better than those from BAT sampling. However, the statistics are similar for the other energy components, (angle, torsion and van der Waals). This is generally consistent with the observation that sampling in BAT coordinates leads to an inappropriately wide distribution of the atom 1- atom 6 bond length, as noted in the previous paragraph. Overall, the distribution of total energy shown in Figure 3.9 indicates that triplet-level samples yield the most accurate conformational distributions, Cartesian coordinates being slightly better than BAT, and that the results become progressively worse on going to doublet and then single-level sampling. The cumulative distribution functions in Figure 3.9, as well as the data in Table 3.4, further document marked reductions in the numbers of abnormally high energy conformations as more correlation is accounted for, in both BAT and XYZ coordinates. Thus, the energy distributions, like the distance distributions, highlight the importance of including higher order correlations for this constrained, yet flexible, chemical ring.

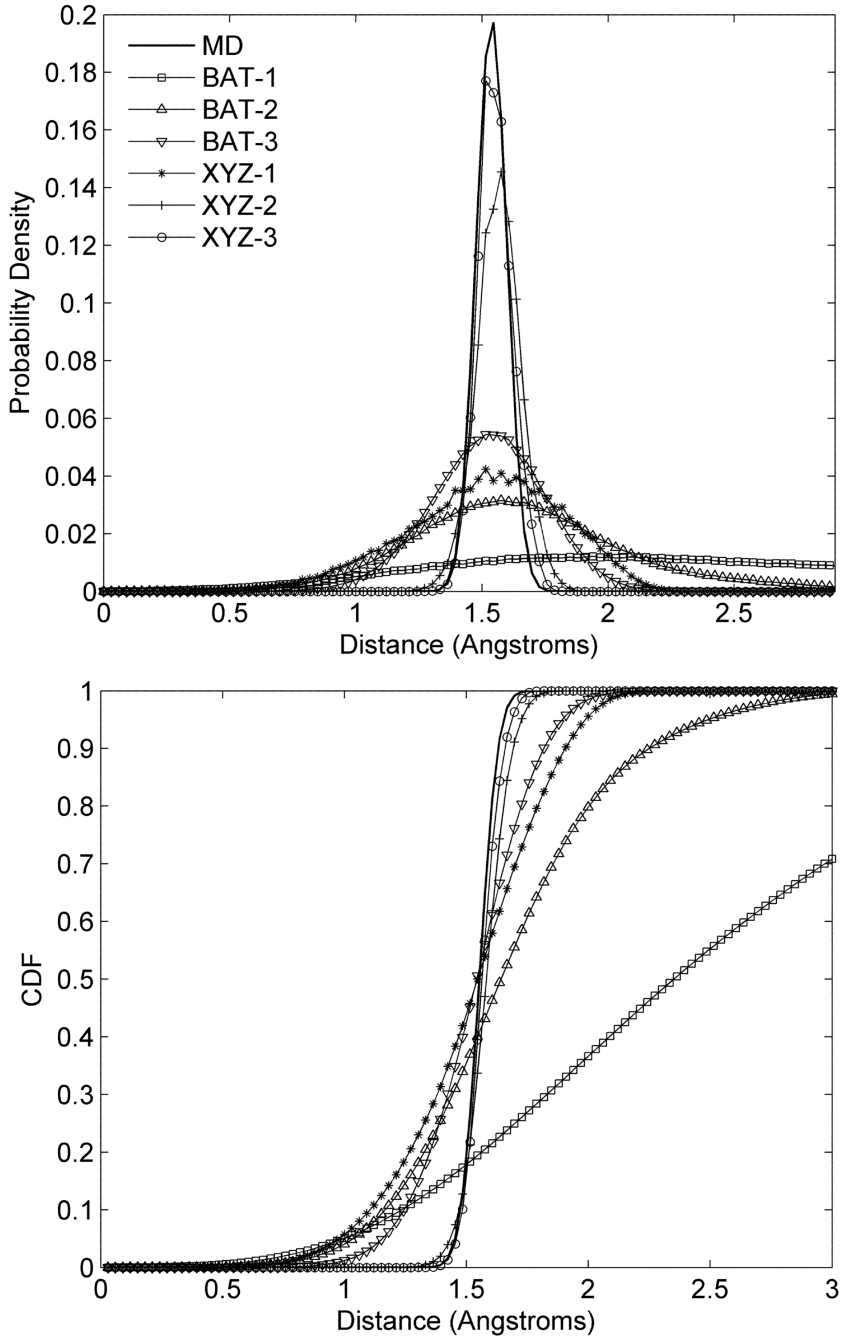


Figure 3.6: Probability distributions (top) and corresponding cumulative distributions (bottom) of cyclohexane end-to-end distances computed from MD conformations and sampled conformations.

Table 3.5: Statistics of end-to-end distance (Å) from MD and sampled conformations for cyclohexane.

	Median	Mean	Standard Deviation	Minimum	Maximum
MD	1.54	1.54	0.06	1.24	1.83
BAT-1	2.34	2.45	1.00	0.03	5.82
BAT-2	1.63	1.66	0.43	0.09	3.74
BAT-3	1.53	1.52	0.22	0.34	2.41
XYZ-1	1.53	1.51	0.31	0.25	2.40
XYZ-2	1.57	1.57	0.09	1.08	1.93
XYZ-3	1.55	1.55	0.07	1.27	1.84

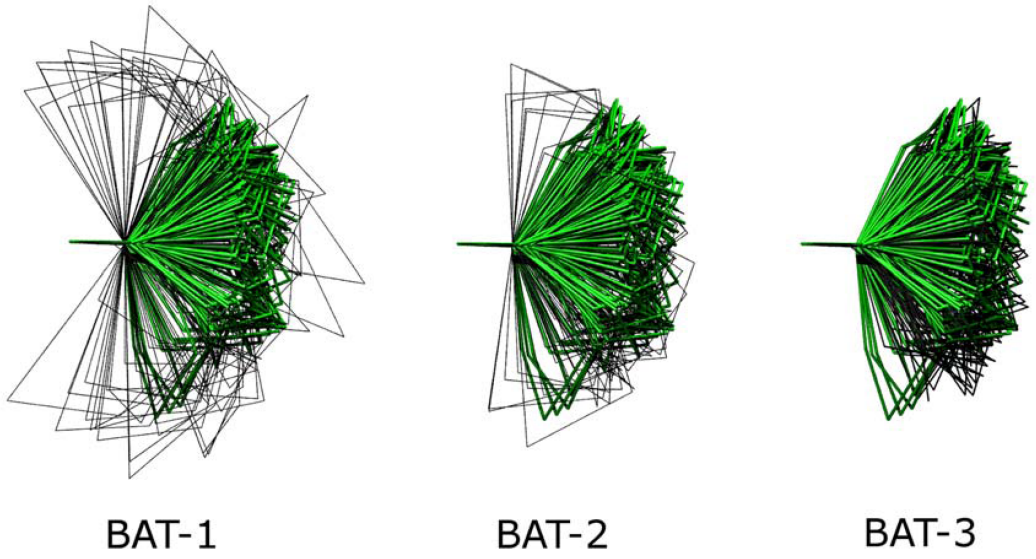


Figure 3.7: Representative cyclohexane conformations from MD (in green) and from sampling (in black) in BAT coordinates at different levels. Each conformation is oriented such that atom 1 is at origin, atom 2 is along the x-axis and atom 3 is in x-y plane. Each set contains 100 conformations.

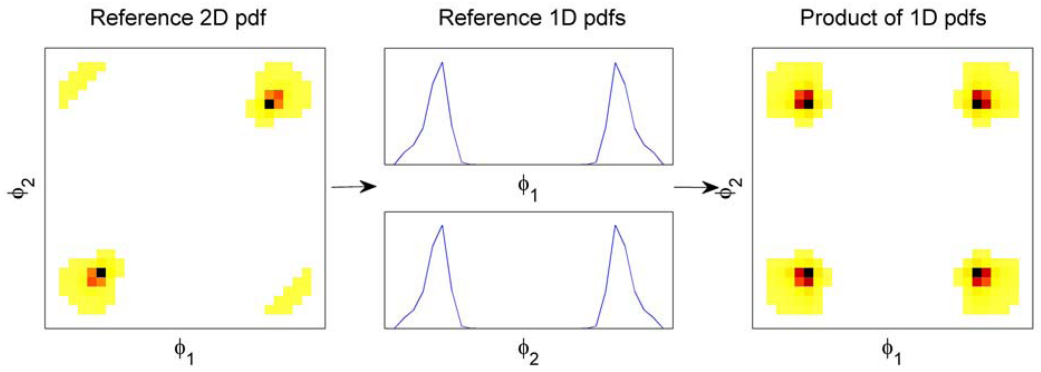


Figure 3.8: The 2-D pdf (right) obtained by multiplying 1-D pdfs (middle) of cyclohexane has non-zero probability in regions not present in the true 2-D pdf (left). Bins with zero probability in the 2-D pdfs are in colored white.

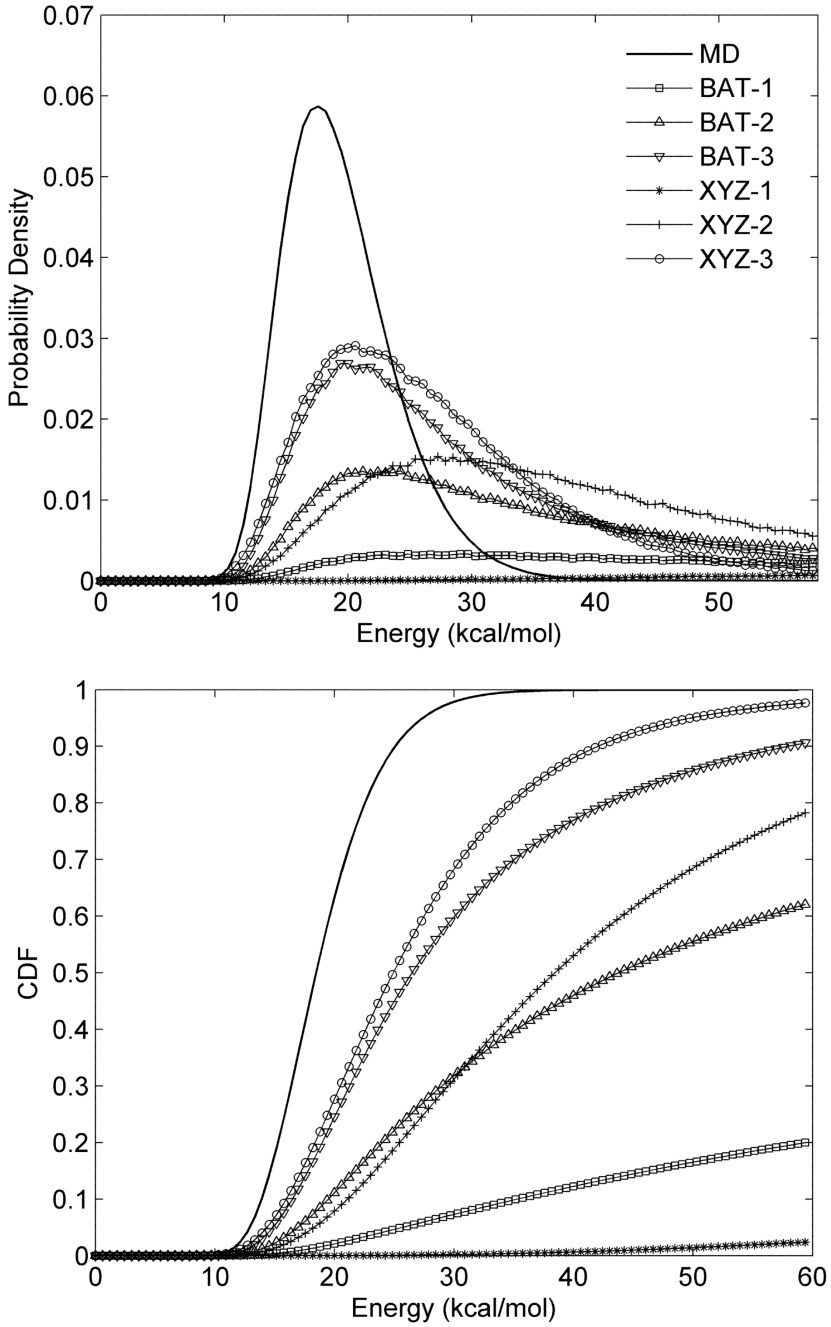


Figure 3.9: Probability distributions (top) and corresponding cumulative distributions (bottom) of the total energy of cyclohexane computed from MD conformations and sampled conformations using different sampling schemes.

Table 3.6: Statistics of energy (kcal/mol) for cyclohexane, computed from MD and sampled conformations. Coulombic term is not reported as it is always zero for this molecule.

	Bond	Angle	Torsion	VDW	Total
			Median		
MD	5.3	3.9	7.9	1.0	18.8
BAT-1	198.4	36.6	7.1	1.5	245.4
BAT-2	24.4	8.4	7.6	1.1	44.1
BAT-3	11.7	4.9	7.9	1.0	26.8
XYZ-1	288.8	56.8	6.9	1.8	363.4
XYZ-2	20.5	7.6	8.1	0.9	38.8
XYZ-3	10.0	5.0	7.8	1.1	25.2
			Mean		
MD	5.9	4.4	7.8	1.2	19.4
BAT-1	498.7	53.8	7.1	3.4	563.0
BAT-2	60.3	14.3	7.5	1.7	83.8
BAT-3	18.5	5.8	7.8	1.4	33.5
XYZ-1	448.8	67.6	7.0	8.1	531.5
XYZ-2	27.7	9.2	8.0	1.2	46.1
XYZ-3	12.7	6.1	7.8	1.3	27.9
			Mean		
MD	3.4	2.6	0.6	0.9	4.5
BAT-1	669.9	50.4	0.8	8.4	713.0
BAT-2	95.7	15.9	0.8	2.1	107.9
BAT-3	19.9	4.1	0.7	1.2	21.5
XYZ-1	444.0	48.4	0.9	90.2	485.9
XYZ-2	24.2	6.5	0.7	0.9	27.0
XYZ-3	10.2	4.3	0.6	0.9	12.1
			Minimum		
MD	0.02	0.02	5.28	-0.21	7.87
BAT-1	0.16	0.09	4.71	-0.30	9.09
BAT-2	0.03	0.05	5.17	-0.29	8.27
BAT-3	0.01	0.04	5.32	-0.26	8.26
XYZ-1	0.32	0.15	3.99	-0.30	12.44
XYZ-2	0.08	0.07	5.37	-0.16	8.63
XYZ-3	0.08	0.04	5.66	-0.12	8.47
			Maximum		
MD	39.9	36.1	10.0	20.5	58.9
BAT-1	4.9×10^3	292.4	9.9	1.6×10^3	5.2×10^3
BAT-2	1.3×10^3	162.5	9.9	152.7	1.4×10^3
BAT-3	379.3	79.8	9.9	28.4	412.8
XYZ-1	4.3×10^3	353.3	10.0	2.8×10^4	2.8×10^4
XYZ-2	369.1	108.0	10.0	30.4	415.8
XYZ-3	180.1	76.0	10.0	18.1	245.2

3.4.3 Host-guest complex

Conformations of the host-guest complex were sampled in BAT coordinates at the singlet, doublet and triplet levels. Table 3.7 and Figure 3.10 analyze the distances between eight atom pairs: 1-11, 1-22, 12-15, 15-23, 11-22, 12-23, 8-19 and 1-15 (see Figure 3.2c for atom numbers). Overall, the distance distributions from doublet- and triplet-level sampling agree well with the reference MD distributions, the triplet-level being somewhat more accurate than doublet. The singlet level samples give notably poorer distributions, especially for distances between host and guest atoms, as shown in Figure 3.10c,d and h.

For both BAT-2 and BAT-3 samples, the median values of all energy components are in excellent agreement with MD, as are the mean values of all energy components other than van der Waals (Table 3.8). Although the tabulated statistics of BAT-1 samples are comparable to those of BAT-2 and BAT-3 samples, Figure 3.11 shows that the distribution of total energies is substantially inferior with singlet level sampling. It is also evident that the mean van der Waals energy of the BAT-3 samples is substantially better than that of BAT-1 and BAT-2 samples, though all are skewed towards higher values due to the presence of a few conformations with bad contacts. These lead to small tails of high energy conformations for BAT-2 and BAT-3 sampling, as evident from the cumulative probability distributions of energy (Figure 3.11). Table 3.4 furthermore documents sharp reductions in the number of conformations with abnormally high energy as more correlation is accounted for in the sampling.

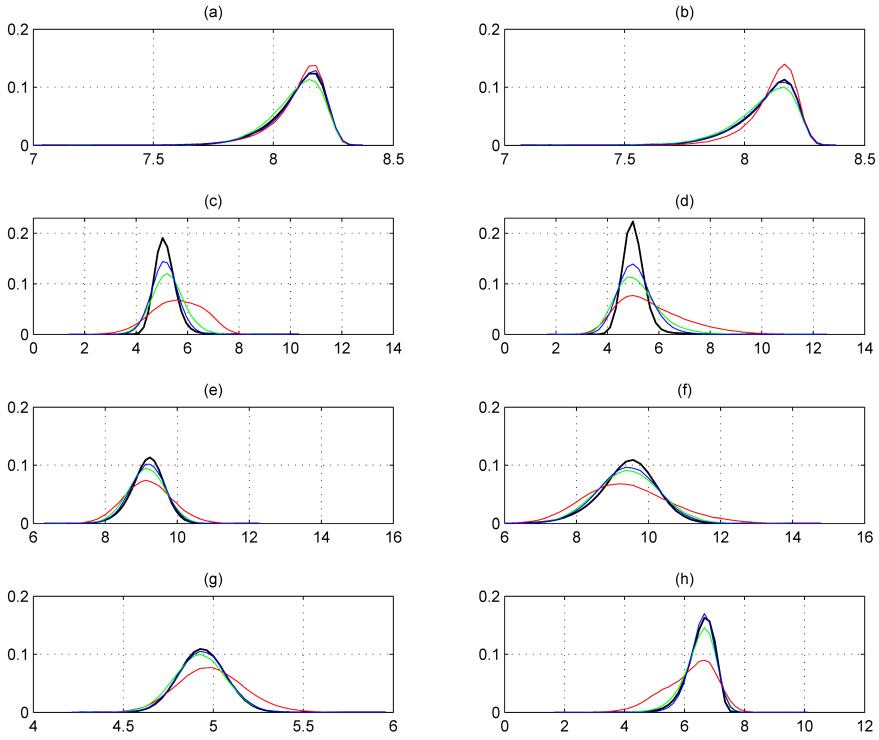


Figure 3.10: Probability distributions of interatomic distances (Å) in the host-guest complex, for atom pairs (a) 1-11, (b) 1-22, (c) 12-15, (d) 15-23, (e) 11-22 (f) 12-23 (g) 8-19 (h) 1-15. Color code: MD in black, BAT-1 in red, BAT-2 in green, BAT-3 in blue.

Table 3.7: Statistics of seven key distances (Å) of host-guest conformations reconstructed from active MD BAT coordinates and sampled BAT coordinates at doublet (BAT-2) and triplet (BAT-3) levels. Column headings give the atom pairs using atom numbers from Figure 3.2

	1-11	1-22	12-15	15-23	8-19	11-22	12-23	1-15
				Median				
MD	8.12	8.12	5.10	4.99	4.94	9.20	9.47	6.60
BAT-1	8.13	8.14	5.63	5.47	4.98	9.17	9.35	6.31
BAT-2	8.11	8.10	5.22	5.06	4.93	9.15	9.43	6.57
BAT-3	8.13	8.12	5.11	5.04	4.94	9.18	9.43	6.62
				Mean				
MD	8.10	8.09	5.14	5.03	4.94	9.17	9.43	6.54
BAT-1	8.11	8.12	5.61	5.71	4.99	9.18	9.43	6.18
BAT-2	8.09	8.08	5.24	5.17	4.93	9.15	9.43	6.50
BAT-3	8.11	8.09	5.12	5.09	4.94	9.17	9.42	6.58
				Standard Deviation				
MD	0.11	0.12	0.43	0.45	0.13	0.46	0.79	0.45
BAT-1	0.11	0.10	1.00	1.34	0.19	0.67	1.20	0.86
BAT-2	0.11	0.13	0.64	0.88	0.14	0.52	0.88	0.54
BAT-3	0.11	0.12	0.52	0.68	0.13	0.48	0.83	0.43
				Minimum				
MD	7.06	7.07	3.42	3.12	4.27	6.31	4.95	3.22
BAT-1	7.00	7.26	1.37	1.66	4.21	6.43	5.23	1.66
BAT-2	7.19	7.16	2.20	2.13	4.30	6.58	5.22	2.20
BAT-3	7.04	7.28	2.79	2.66	4.36	7.07	6.06	3.69
				Maximum				
MD	8.37	8.38	10.34	10.33	5.92	12.17	14.50	7.92
BAT-1	8.35	8.35	9.60	12.53	5.96	12.29	14.79	10.11
BAT-2	8.37	8.36	9.00	11.58	5.68	11.45	13.56	8.41
BAT-3	8.36	8.37	8.17	10.22	5.51	11.10	12.64	8.09

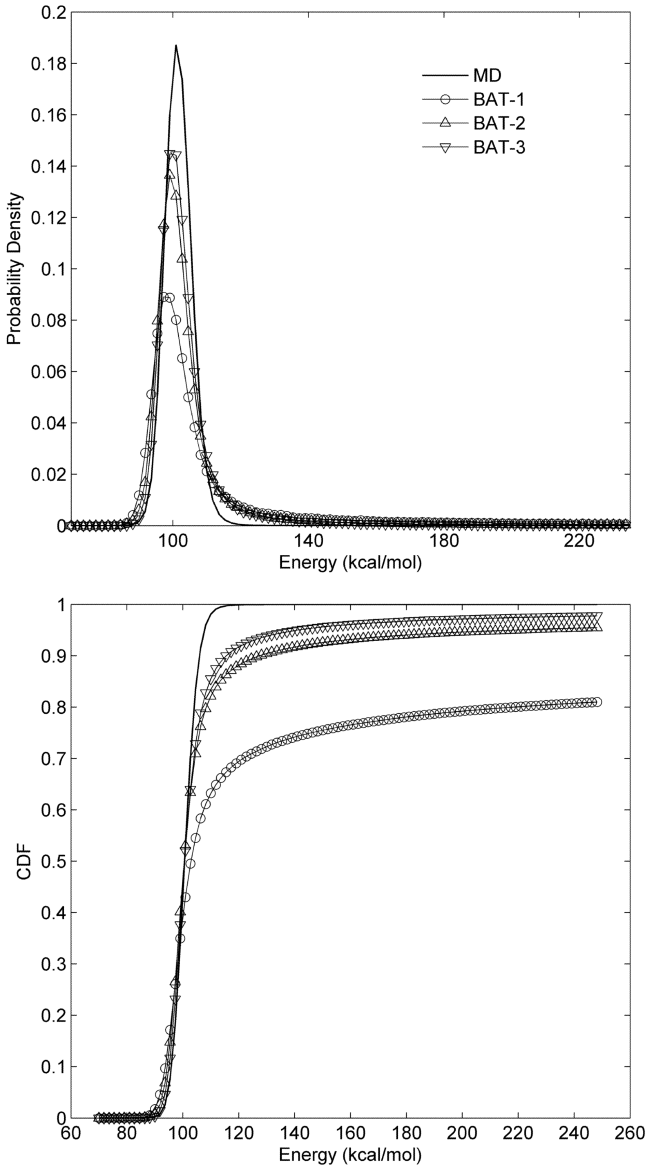


Figure 3.11: Probability distributions (top) and corresponding cumulative distributions (bottom) of total energy of the host-guest complex for MD and sampled conformations.

Table 3.8: Statistics of energy (kcal/mol) of host-guest complex for conformations reconstructed from active MD BAT coordinates and sampled BAT coordinates at doublet (BAT-2) and triplet (BAT-3) levels.

	Bond	Angle	Torsion	Improper	Coulomb	VDW	Total
				Median			
MD	0.0014	7.16	3.87	0.12	65.1	24.7	101.5
BAT-1	0.0015	7.69	3.98	0.16	62.4	26.9	104.0
BAT-2	0.0015	7.20	4.14	0.17	63.6	25.1	101.4
BAT-3	0.0015	7.19	3.85	0.16	64.4	25.0	101.6
				Mean			
MD	0.0015	7.35	4.07	0.27	65.1	24.9	101.7
BAT-1	0.0015	7.94	4.21	0.27	64.0	5.4×10^4	5.4×10^4
BAT-2	0.0015	7.38	4.35	0.28	64.4	3.4×10^3	3.5×10^3
BAT-3	0.0015	7.36	4.05	0.25	65.1	959.7	1.0×10^3
				Standard Deviation			
MD	0.0004	1.89	1.53	0.38	2.97	1.66	4.26
BAT-1	0.0004	2.19	1.65	0.39	11.3	4.8×10^5	4.8×10^5
BAT-2	0.0004	1.90	1.72	0.40	5.49	1.1×10^5	1.1×10^5
BAT-3	0.0004	1.87	1.50	0.34	4.75	6.2×10^4	6.2×10^5
				Minimum			
MD	0.0003	1.98	0.19	0.00	46.6	23.0	82.9
BAT-1	0.0003	2.37	0.48	0.00	-583.1	23.0	80.8
BAT-2	0.0003	2.11	0.49	0.00	-111.6	23.0	79.8
BAT-3	0.0004	2.29	0.61	0.00	40.6	23.2	84.8
				Maximum			
MD	0.0044	24.1	17.2	8.23	84.0	2.5×10^3	2.6×10^3
BAT-1	0.0042	23.1	16.4	7.48	1.2×10^3	9.9×10^6	9.9×10^6
BAT-2	0.0041	19.4	16.9	4.54	238.7	9.2×10^6	9.2×10^6
BAT-3	0.0040	19.0	12.5	2.97	262.9	8.9×10^6	8.9×10^6

3.4.4 Comparing sampled and reference marginal pdfs

Another way to assess the accuracy of the sampled distributions is to compare their marginals with those of the original MD simulations. For singlet level sampling, all 1-D sampled marginal pdfs converge trivially to the corresponding reference pdfs. For higher level sampling, only the marginals of the first l variables converge to the reference marginals (data not shown), as expected based upon the structure of the sampling algorithm. The marginals of the subsequent variables are sampled from approximate conditional distribution and therefore are expected to deviate from the reference marginals.

The deviations were quantified by computing the root mean square deviations (RMSD) between all 1-, 2- and 3-D sampled and reference marginal pdfs for each sampling case. Table 3.9 reports the mean RMSD for the three marginal pdfs in each sampling case. As expected, for a specific molecule and coordinate system, the mean RMSD of singlet marginal pdfs is lower for singlet level sampling than for higher-level sampling. However, the mean RMSDs of doublet and triplet marginal pdfs from doublet- or triplet-level sampling are, in general, lower than those from singlet level sampling, indicating presence of correlations similar to those in MD simulations. Sampling of nonane in XYZ coordinates does not follow this trend, indicating that triplet-level is not sufficient to capture the correlations in this coordinate system. This observation is consistent with the analysis of distance and energy distributions, above.

Marginal pdfs generated from sampled conformations are graphically compared with the corresponding reference marginals from the MD simulations in Figure 3.12 and Figure 3.13. Figure 3.12 shows the doublet marginal pdf from the XYZ study of nonane which yielded the highest RMSD values at the singlet, doublet and triplet levels; and Figure 3.13 similarly analyzes the doublet marginal which yielded the highest RMSD in

BAT-1 sampling. It is worth noting that the doublet marginal pdf from BAT-1 sampling (Figure 3.13) is much closer to its reference MD marginal than is the XYZ-1 result for nonane (Figure 3.12), confirming earlier results where BAT coordinate performed better than XYZ.

As discussed in Section 3.2.4, the sampling algorithms generate conformations that have non-zero probability in all the reference marginals. This property is apparent in Figure 3.12 and Figure 3.13, where the populated bins of a representative 2-D marginal from doublet- and triplet-level sampling (bottom panel) are a subset of those populated by the MD trajectories (top left), unlike the marginal from singlet level sampling. In this way, the reference marginals constrain the range of conformational space accessible to the sampling algorithm.

Table 3.9: Accuracy of singlet, doublet and triplet marginal distributions (columns) computed from conformations sampled at the singlet, doublet and triplet levels (rows), for the three molecular systems. Results are reported as mean RMSD across all

	Singlet	Doublet	Triplet
Nonane			
BAT-1	0.0013	0.0023	0.0020
BAT-2	0.0013	0.0014	0.0015
BAT-3	0.0020	0.0019	0.0019
XYZ-1	0.0012	0.0267	0.0157
XYZ-2	0.0915	0.0620	0.0336
XYZ-3	0.0775	0.0530	0.0290
Cyclohexane			
BAT-1	0.0012	0.0062	0.0045
BAT-2	0.0046	0.0037	0.0025
BAT-3	0.0028	0.0022	0.0017
XYZ-1	0.0012	0.0207	0.0129
XYZ-2	0.0238	0.0180	0.0093
XYZ-3	0.0195	0.0148	0.0083
Host-Guest Complex			
BAT-1	0.0021	0.0060	0.0050
BAT-2	0.0098	0.0085	0.0055
BAT-3	0.0063	0.0052	0.0039

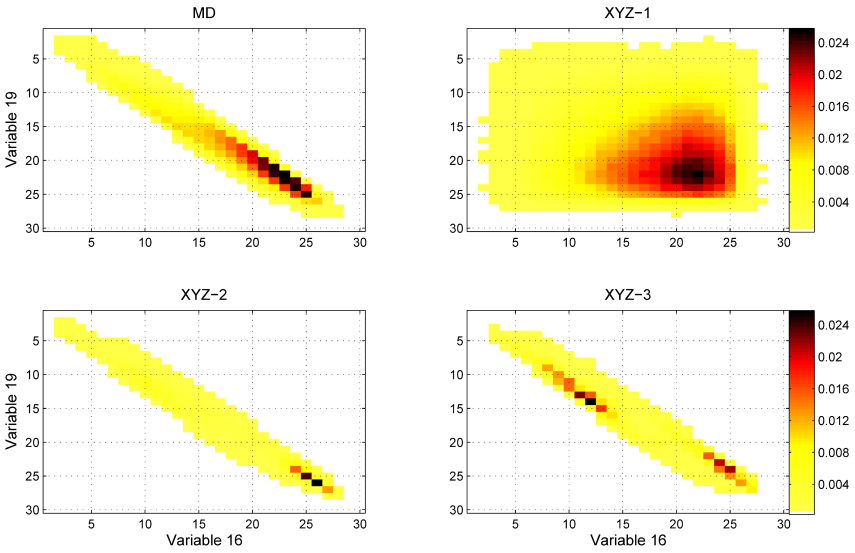


Figure 3.12: Comparison of doublet marginals from MD with XYZ-1, XYZ-2 and XYZ-3 sampling of variables 16 and 19 of nonane, which correspond to the x-coordinates of atoms 8 and 9 (Figure 3.2). The RMSDs of the sampled doublet marginal pdfs are 0.0504, 0.0955 and 0.0692 for XYZ-1, XYZ-2 and XYZ-3, respectively. Cells are colored on a linear scale of probability, except that cells with identically zero probability are colored white.

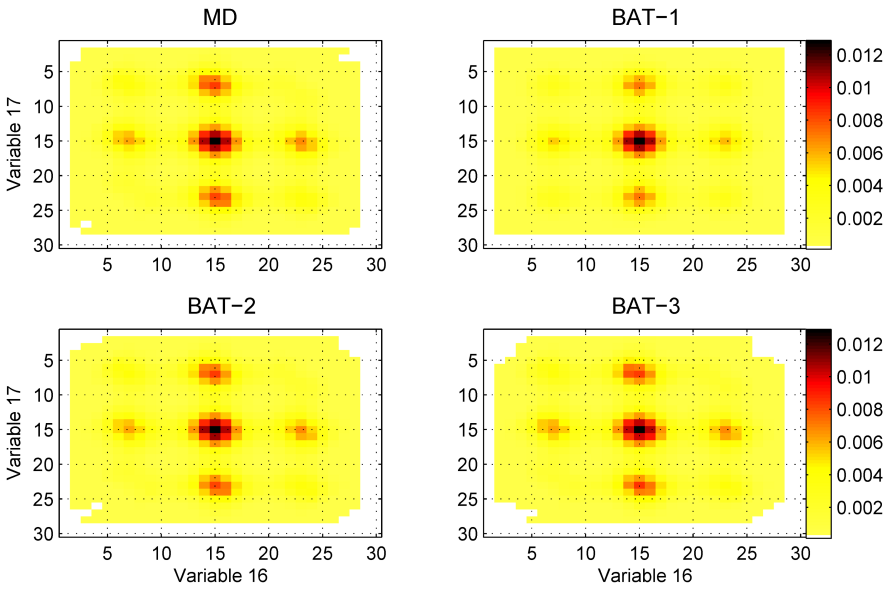


Figure 3.13: Comparison of doublet marginals from MD with BAT-1, BAT-2 and BAT-3 sampling of variables 16 and 17 of nonane, corresponding to torsions 1-2-3-4 and 2-3-4-5 respectively (Figure 3.2). The RMSDs of the sampled doublet marginal pdfs are 0.0106, 0.0015 and 0.0018 for BAT-1, BAT-2 and BAT-3, respectively. Cells are colored on a linear scale of probability, except that cells with identically zero probability are colored white.

3.4.5 Comparing sampled conformations with MD conformations

The part of the N-dimensional configuration space accessible to the present sampling algorithms is determined by the zero probability bins in reference pdfs used. In one extreme case, if the reference distributions were populated using a single MD conformation, then all reference distributions would have zero probability in all bins except one, and only the single MD conformation would be sampled at all levels.

It was also noted that the sampling algorithms can generate novel conformations. In all of the present test systems, over 99% of the $2 - 5 \times 10^5$ sampled conformations are new relative to the 5×10^6 original MD conformations used to populate the reference marginals. (Conformations were compared after mapping the coordinates to bin numbers in the discrete space.) This small degree of overlap is intuitively reasonable, given the large number of potential conformations, $\sim B^N$, where $B = 30$ is the number of bins used to discretize each coordinate and N is the dimensionality. Here N is 12, 21 and 32 for cyclohexane, nonane and host-guest complex, respectively, so the number of conformations ranges from $\sim 10^{17}$ to $\sim 10^{47}$. To better judge whether the high fraction of novel conformations generated by the sampling algorithm is reasonable, we did the following test on the MD trajectories of cyclohexane and nonane. For both molecules, we counted the number of conformations in the first 5×10^5 MD that were repeated in the following 4.5×10^6 conformations. We found no repeats for nonane, while $< 1\%$ of the first 5×10^5 of cyclohexane were repeated, consistent with the low overlap between the sampled with corresponding MD conformations.

3.5 Discussion

This chapter introduced algorithms for sampling molecular conformations in a manner that includes correlations up to a desired order by means of superposition approximations, and employs the algorithms to test the importance of correlations in capturing conformational fluctuations. We find that incorporating higher order correlations systematically improves the distributions of sampled conformations as compared with the MD conformations, and that conformations sampled via superposition approximations at the doublet or triplet levels resemble MD conformations rather well, for one or both of the BAT or Cartesian coordinate systems. This observation supports the hypothesis that molecular fluctuations may be described to good approximation in the absence of high-order correlations. This assessment relies on the results obtained for the three molecular systems considered here, but we expect the picture to be similar for other molecular systems of similar size and type. It would clearly be of interest to know whether the same is true for larger systems, such as proteins.

This study was motivated by evidence that configurational entropy may be approximated to good accuracy without accounting for high-order mutual information terms (see Section 3.1). Comparing sampled conformations, as done here, provides a more stringent test of our hypothesis than merely comparing entropy values, because the same entropy could be obtained for two very different conformational distributions. The present study confirms that neglecting high-order correlations still allows generation of reasonably good conformations. By using conditional distributions based on mixed-SA (Section 2.4.3), the present sampling algorithms can be generalized to use a selected set of higher-order reference pdfs. The present approach therefore provides a framework for investigating the contributions of selected correlations to fluctuations in a multi-dimensional system.

The SA-l sampling algorithms sample points in the high dimensional space which are simultaneously consistent with all of the given reference pdfs. In principle, this could have been accomplished by using standard ancestral sampling to sample directly from the N-dimensional SA-l distribution constructed using the reference samples. In practice, however, constructing the conditional pdfs to carry out such an approach would require all marginal pdfs of the SA-l, and this would be intractable because of their high dimensionality and because all variables in SA-l are coupled with one another. Another approach to sample directly from the SA-l distribution could be to use Gibbs sampling, but that is likely to be computationally inefficient because successive samples in Gibbs sampling are correlated.

The present approach is attractive for molecular conformational sampling for several reasons. First, it retains the key feature of SA-l distributions – a functional form in terms of product of all reference distributions. This ensures that the samples are simultaneously consistent with all the pdfs used and limits the sampled conformations to regions of configuration space for which the reference marginal pdfs are populated. This helps to avoid sampling conformations with grossly unrealistic energies, while still allowing construction of new conformations, i.e., ones not present in the MD samples used to construct the reference distributions.

Second, the sampling distribution of the present algorithms is normalized, unlike the SA-l distribution (except for the trivial case of SA-1). Thus the sampling distribution could be used as a proposal distribution and reweighted according to any N-dimensional distribution, including the SA-l and the Boltzmann distribution, using standard importance sampling methods [59]. This property will be used in the next chapter to compute the normalization of the Boltzmann distribution, that is, the configurational integral of a

molecule.

Third, unlike other methods of sampling from high dimensional distributions, such as Gibbs sampling, MC and MD, successive samples from the SA-l sampling are uncorrelated with each other: in effect, the algorithm has no memory of prior conformations. This prevents the sampling from getting trapped in particular regions of the configurational space, such as in low-energy wells of the physical energy surface with high barriers and should allow better coverage of the conformational space. Note that, although the conformational region accessible to the sampling algorithms is larger than the region sampled by the MD simulation, the quality of the MD sampling used to build the reference marginals is still important. This is because the boundaries of the region accessible to sampling are a function of the MD samples. For example, consider a molecule that can be in an open or a closed conformation, depending on the value of a single torsion. If none of the MD conformations are in the open state, so that the open state torsion values are never observed, then SA-l sampling cannot access the open state. On the other hand, this property could be used to focus the sampling in particular regions of the configurational space which could be advantageous, e.g., in computing the absolute free energy of a particular conformational state (see item 1 of Section 5.2).

It is worth elaborating on potential weaknesses of the present approach as well. One is the possibility of generating null samples. However this occurred very rarely in the present tests. The likelihood of generating null samples might be reduced by using a reduced set of reference pdfs through the mixed-SA approach and by using more data to populate the reference pdfs. Second, the sampling distribution depends upon the sampling sequence, and the optimal sampling order – one maximising the similarity between the sampling distribution ($\tilde{p}_N$) and the true distribution (p_N) – is not known *a priori*, although,

if required, the sequence-dependence could be removed simply by randomizing the order of sampling.

Chapter 5 discusses potential strategies to scale the sampling algorithms presented here to larger systems and for improving the quality of the sampled conformations.

Chapter 4

Free Energy Calculation using Superposition Approximation Based References

4.1 Introduction

The calculation of the free energy of a molecule is a key problem in computational chemistry and biophysics, with applications to conformational stability [60], solvation [61] and molecular recognition [19, 62]. Systems of interest in this regard include relatively small (<100 atoms) drug-like molecules, moderately sized supramolecular systems (100-1000 atoms), and proteins with thousands of atoms. The challenge of calculating molecular free energies stems chiefly from the complexity and high dimensionality of the energy surfaces involved [63]. In one approach, the problem is recast as a calculation of the free energy difference between the system of interest and a reference system whose free energy is known. The effectiveness of this reference system approach increases as the conformational probability distribution of the reference system more closely approximates that of the physical system of interest. Contrariwise, if the reference distribution has little overlap with the physical system, then it will be very difficult to obtain a reliable, converged result. As a consequence, the choice of reference system is of critical importance.

In some calculations of this type, the free energy of the reference system can be computed analytically, as in the harmonic or "Einstein solid" reference system for solids [64, 65] or the ideal gas reference system for liquids [66, 67]. A reference system approach to computing the absolute free energy of a molecule was apparently first employed by Stoessel and Nowak [68], whose reference system was a collection of independent harmonic oscillators centered at atomic coordinates. More recently, simulation data have begun to be used to construct numerical, rather than analytical, reference systems for free energy calculations [69, 70]. In one approach, MD simulation data is used to set up a harmonic reference system whose free energy is computed using normal mode analysis [70, 71]. In another approach distributions of internal coordinates observed in a simulation are used to set up the reference system [69]. Reference systems set up using simulation data can be advantageous, because they can better capture the flexibility and inhomogeneity of biomolecules, thereby increasing conformational overlap with the physical system. They also avoid the need for *ad hoc* tuning of adjustable parameters, such as spring constants.

The work presented in this chapter is inspired by the reference system method of Zuckerman and coworkers [69], which uses molecular dynamics simulation data to construct one-dimensional pdfs of internal bond, angle and torsion coordinates. Their reference distribution is simply the product of these 1-D pdfs, and thus is effectively the singlet level superposition approximation (SA-1) to the Boltzmann distribution. Samples drawn from this reference distribution are used to compute the configurational integral of the molecule. The singlet level reference system is relatively simple to construct, but does not capture correlations among the internal coordinates, and this neglect of correlations reduces its overlap with the physical Boltzmann distribution. In particular, as the system size (dimensionality) increases, there is a rapid rise in the fraction of conformations

sampled from the reference system that have high energies in the physical system, mainly due to steric clashes among the atoms. This can lead to poor convergence of the free energy estimate, restricting the applicability of the method to small molecules with weak correlations, such as linear chains with weak non-bonded interactions.

The previous chapter presented the SA-l based approximations to the high-dimensional Boltzmann distribution of a molecule which can account for specified correlations among the internal coordinates. We also presented algorithms to sample molecular conformations from SA-l based approximations constructed using marginal pdfs of up to order l. Importantly for the present application, the sampled distribution is normalized by construction, allowing us to set up a reference system with known free energy. As a consequence, the free energy of the physical system can be computed as the known free energy of the reference system plus the free energy difference between the physical and reference systems. In this work, we use the SA-l based sampling distribution as the reference canonical distribution, and as done previously [69], the free energy difference is estimated using thermodynamic perturbation with samples drawn from the reference distribution. The main result of this chapter is that incorporating pairwise correlations among all internal coordinates in the reference system dramatically improves the convergence of the free energy estimates, in comparison to the original singlet level reference system.

The following section describes the underlying theory and overall approach, including the definition of the discretized reference system and the approach to computing the free energy of the physical system using samples from the reference system. Section 4.3 section then details the implementation of the free energy calculations for molecular systems. Section 4.4 evaluates the method numerically on a series of molecular test systems of

increasing complexity which were used as test cases for a similar calculation by Zuckerman and coworkers with an uncorrelated reference distribution [72]. Finally, Section 4.5 assesses the significance of the results. This chapter is based on Ref [73].

4.2 Theory and Approach

We wish to compute a molecule's configuration integral,

$$Z_P = \int_{\Gamma_P} \exp\left(-\beta\, U_P(\boldsymbol{\xi})\right) J(\boldsymbol{\xi})\, d\boldsymbol{\xi} \tag{4.1}$$

where $\boldsymbol{\xi}$ denotes the vector of N internal coordinates, β is the inverse temperature and $J(\boldsymbol{\xi})$ is the Jacobian. The potential energy of the molecule, $U_P(\boldsymbol{\xi})$, is termed the physical energy and, here, is given by a molecular mechanics force field. The integral is over the region Γ_P corresponding to a conformational state of interest; e.g. a tertiary structure of a peptide, or the bound state of a protein-small molecule system. The absolute free energy, F_P, of the molecule is given by

$$\beta F_P = -\ln Z_P\,. \tag{4.2}$$

Note that Z_P is treated as a dimensionless number, and the units of F_P are set by $k_B T$. In this chapter, the BAT internal coordinate system is used with $\boldsymbol{\xi} = (\mathbf{b}, \mathbf{a}, \mathbf{t}) \in \mathbb{R}^N$ where $\mathbf{b}$, $\mathbf{a}$ and $\mathbf{t}$ are vectors denoting M-1 bond-lengths, M-2 bond angles and M-3 torsions, respectively, M being the number of atoms in the molecule. The Jacobian for the BAT coordinate system, which is independent of the torsion angles, is given by [74, 75, 53]

$$J(\mathbf{b}, \mathbf{a}, \mathbf{t}) = b_3 \prod_{i=4}^{M} \left(b_i^2 \,\sin a_i\right). \tag{4.3}$$

Figure 4.1 illustrates the specification of the BAT coordinates for test molecules in this chapter, using propane as an example.

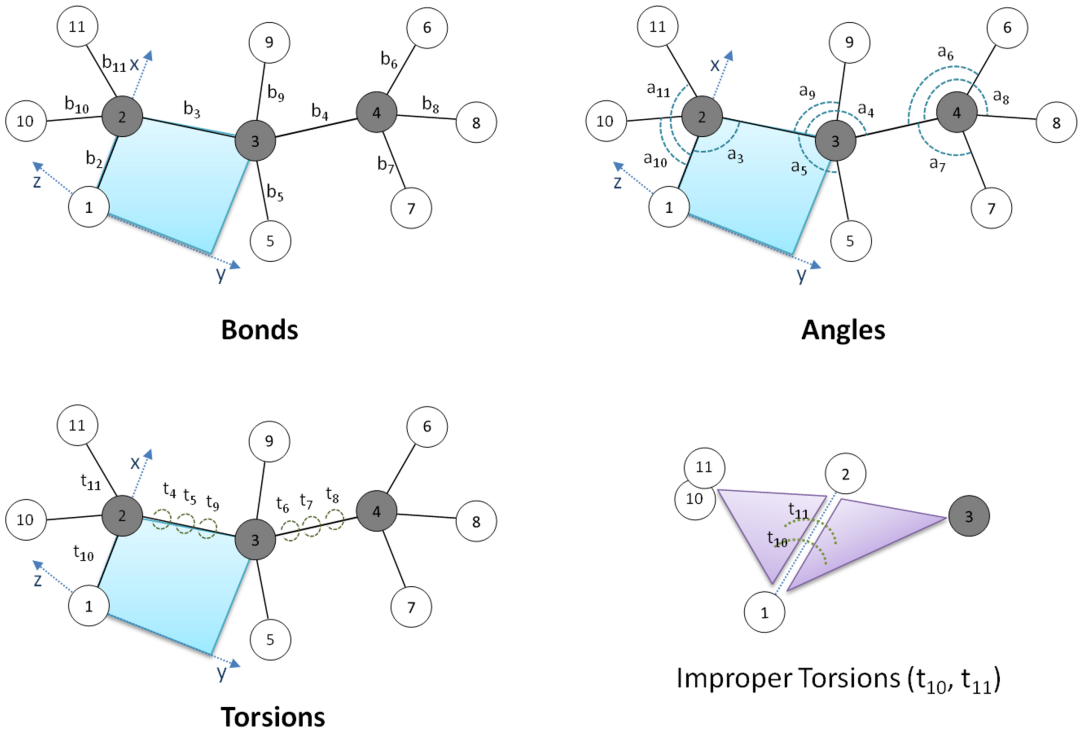


Figure 4.1: Internal coordinate systems illustrated using propane. The anchored Cartesian system is defined in terms of three root atoms such that atom 1 is at the origin, atom 2 is on the x-axis and atom 3 is in the x-y plane (shaded in blue). For the peptides in Figure 4.2, a C-terminal hydrogen is labeled as atom 1 and two subsequent carbon atoms in the chain are labeled, in order, as atoms 2 and 3. The bond, angle and torsion coordinates in the BAT system are labeled. In all, propane has $11 \times 3 - 6 = 27$ internal coordinates specified by 10 bond lengths (top left), 9 bond angles (bottom left) and 8 torsion angles (bottom right). Torsion angles for atoms 10 and 11 are improper torsions (top right), and the others are proper torsions. Shaded circles: carbon; unshaded circles: hydrogen.

4.2.1 Discretized reference systems

The reference system is defined in terms of the SA-l based sampling distributions which are set up by discretizing the N-dimensional BAT conformational space and using MD simulation data to populate low-order reference pdfs as described in Section 3.3 of the previous chapter. A conformation, in the discrete space is denoted by $\mathbf{X} = (X_1, \ldots, X_N) \in \mathbb{Z}^N$, with $X_i \in \{1, .., B\}$ denoting a bin number for the i-th coordinate. (In this chapter, the distinction between upper and lower-case variables is not needed, as the context makes the sense unambiguous.) A discrete-space conformation $\mathbf{X}$ maps to the continuous space conformation $\boldsymbol{\xi}(\mathbf{X})$ as per Eq. 3.21. The MD simulation data are used to construct the N singlet pdfs, $p(X_i)$, and the $N(N-1)/2$ doublet pdfs, $p(X_i, X_j)$ where $i < j \in 1, .., N$. The singlet and doublet reference systems are set up in terms of the SA-1 and SA-2 based sampling distributions, which are given, respectively, by

$$\tilde{p}_N^{(1)}(\mathbf{X}) = p(X_1) \times .. \times p(X_N) \tag{4.4}$$

and

$$\tilde{p}_N^{(2)}(\mathbf{X}) = p(X_1, X_2) \times p^{(2)}(X_3|X_1, X_2) \times .. \times p^{(2)}(X_N|X_1, .., X_{N-1})\,. \tag{4.5}$$

Conformations are sampled from these reference distributions via algorithms described in Section 3.2. In the case of doublet level sampling, which depends on the order in which the coordinates are sampled, the torsion coordinates are sampled first, followed by the bond-angles and, finally, the bond-lengths. We define the potential energy function for the reference systems as:

$$U_R^{(l)}(\mathbf{X}) \equiv \begin{cases} -\frac{1}{\beta} \ln \tilde{p}_N^{(l)}(\mathbf{X}) \ , \ \text{if } \tilde{p}_N^{(l)}(\mathbf{X}) \neq 0 \\ \infty \qquad\qquad\qquad , \text{if } \tilde{p}_N^{(l)}(\mathbf{X}) = 0 \end{cases} \tag{4.6}$$

where the second case accounts for the fact that the sampling probability for some conformations may be zero. Based on Eq. 4.6, conformations drawn from $\tilde{p}_N^{(l)}$ are effectively sampled from the canonical distribution corresponding to the reference energy function. Since the reference energy function is defined in the discrete space, the partition function for the reference system is given by the discrete sum

$$Z_R^{(l)} = \sum_{\mathbf{X}_i \in \Omega} \exp(-\beta\, U_R^{(l)}(\mathbf{X}_i)) = \sum_{\mathbf{X}_i \in \Omega} \tilde{p}_N^{(l)}(\mathbf{X}_i) = 1 \tag{4.7}$$

where Ω denotes the set of all possible conformations in the discretized conformational space. The last equality derives from the normalization of the sampling distribution (item 2 of Section 3.2.4). Therefore, the free energy of the reference systems, $\beta F_R = -\ln Z_R^{(l)}$, is identically zero for both singlet and doublet references. Therefore, the physical free energy can be obtained as the free energy difference between the physical system and the reference systems.

4.2.2 Estimation of the physical free energy

The configurational integral in Eq. 4.1 can be approximated as a sum over the states of the discretized conformational space,

$$Z_P \approx \bar{Z}_P \equiv \Delta V \sum_{\mathbf{X}_i \in \Omega} \exp\left(-\beta\, U_P(\,\boldsymbol{\xi}(\mathbf{X}_i)\,)\right)\;\; J(\,\boldsymbol{\xi}(\mathbf{X}_i)\,) \tag{4.8}$$

where $\Delta V \equiv \delta_1 \times .. \times \delta_N$ is the volume of a cell in the discretized BAT space, and the Jacobian is given by Eq. 4.3. Defining the effective physical energy in the discretized space as

$$\bar{U}_P(\mathbf{X}) \equiv U_P(\boldsymbol{\xi}(\mathbf{X})) \; - \frac{1}{\beta} \ln J(\boldsymbol{\xi}(\mathbf{X})) - \frac{1}{\beta} \ln \Delta V \tag{4.9}$$

gives

$$\bar{Z}_P = \sum_{\mathbf{X_i} \in \Omega} \exp\left(-\beta \bar{U}_P(\mathbf{X_i})\right) \; . \tag{4.10}$$

Note that the last term of Eq. 4.9 is independent of the conformation and is effectively a constant offset to the physical energy depending on the discretization. The approximation in Eq. 4.10 goes to the continuous integral of Eq. 4.1 in the limit of infinitely fine discretization ($\delta_i \to 0$) if the set of discrete space conformations, Ω, covers the desired region of the continuous conformational space, Γ_P. Since the reference partition function, $Z_R^{(l)}$ is one (Eq. 4.7), the discrete sum approximation, $\bar{F}_P$, of the physical free energy, F_P, can be written as

$$\beta \bar{F}_P = -\ln \frac{\bar{Z}_P}{Z_R^{(l)}} \,. \tag{4.11}$$

Defining the energy difference for conformation $\mathbf{X}$ as

$$\Delta U^{(l)}(\mathbf{X}) \equiv \bar{U}_P(\mathbf{X}) - U_R^{(l)}(\mathbf{X}) \tag{4.12}$$

we can write the ratio of the partition functions in Eq. 4.11 in the form of a thermodynamic perturbation [76]

$$\beta \bar{F}_P^{(l)} = -\ln \left\langle \exp(-\beta\, \Delta U^{(l)}) \right\rangle_{\text{ref}\,(l)} \tag{4.13}$$

where the superscript on the left-hand side acknowledges the potential dependence of the computed physical free energy on the reference system, as explained in the next subsection. The perturbation estimate is computed using samples drawn from the singlet or doublet reference canonical distributions as

$$\begin{aligned} \beta \bar{F}_P^{(l)} &\doteq -\ln \left(\frac{1}{N_R} \sum_{n=1}^{N_R} \exp(-\beta\, \Delta U^{(l)}(\mathbf{X}_{i(n)})) \right) \\ &\equiv \beta \hat{F}_P^{(l)} \end{aligned} \tag{4.14}$$

where $\mathbf{X}_{i(n)} \sim \tilde{p}_N^{(l)}$ denotes the conformation corresponding to the n-th sample, and N_R is the total number of reference samples. Note that conformations with infinite reference energy, which have zero probability in the reference distributions, are never sampled, and therefore, the energy difference in Eq. 4.14 is finite for all samples.

For clarity, we note that this discussion has developed the following series of approximations,

$$\beta F_P \approx \beta \bar{F}_P \approx \beta \bar{F}_P^{(l)} \approx \beta \hat{F}_P^{(l)} \tag{4.15}$$

where the first approximation is associated with the discretization of the coordinate space, the second approximation indicates potential dependence on the reference system and the final approximation results from the finite number of reference samples used in practice. The superscripts again denote the singlet ($l = 1$) or doublet ($l = 2$) reference systems used here.

The approach outlined here is equivalent to the importance sampling method [59], where samples drawn from one distribution (here, the reference distributions) are reweighted to compute averages with respect to the distribution of interest (here, the physical distribution). In the context of importance sampling, the quantity $e^{-\beta\, \Delta U^{(l)}}$ is referred to as the importance weight, and the reference distribution as the proposal distribution [50].

4.2.3 Bias and convergence of the free energy estimate

The bias and convergence of the free energy estimate, $\bar{F}_P^{(l)}$, of Eq. 4.14 can be understood by analyzing its asymptotic, or infinite sampling, limit:

$$\begin{aligned} \lim_{N_R \to \infty} \exp\left(-\beta \hat{F}_P^{(l)}\right) &= \lim_{N_R \to \infty} \frac{1}{N_R} \sum_{n=1}^{N_R} \exp\left(-\beta\, \Delta U^{(l)}(\mathbf{X}_{i(n)})\right) \\ &= \sum_{\mathbf{X}_i \in\, \Omega} \tilde{p}_N^{(l)}(\mathbf{X}_i) \exp\left(-\beta\, \Delta U^{(l)}(\mathbf{X}_i)\right). \end{aligned} \tag{4.16}$$

The second equality uses the fact that, in the limit of infinite sampling, the reference samples are distributed according to $\tilde{p}_N^{(l)}$. Dropping conformations with zero probability in the reference distributions from the summation, and substituting $\Delta U^{(l)}$ from Eq. 4.12

gives

$$\begin{aligned}\lim_{N_R\to\infty} \exp(-\beta\,\hat{F}_P^{(l)}) &= \sum_{\mathbf{X}_i\in\Omega;\tilde{p}_N^{(l)}\neq 0} \tilde{p}_N^{(l)}(\mathbf{X}_i)\exp\left(+\beta\,U_R(\mathbf{X}_i)\right)\exp\left(-\beta\bar{U}_P(\mathbf{X}_i)\right) \\ &= \sum_{\mathbf{X}_i\in\Omega;\tilde{p}_N^{(l)}\neq 0} \tilde{p}_N^{(l)}(\mathbf{X}_i)\frac{1}{\tilde{p}_N^{(l)}(\mathbf{X}_i)}\exp\left(-\beta\bar{U}_P(\mathbf{X}_i)\right) \\ &= \sum_{\mathbf{X}_i\in\Omega^{(l)}} \exp\left(-\beta\bar{U}_P(\mathbf{X}_i)\right) \qquad (4.17)\end{aligned}$$

where Eq. 4.6 is used in the second step and $\Omega^{(l)}$ denotes the set of conformations accessible to l-level sampling. Comparing Eq. 4.17 with Eq. 4.10, one can infer that the estimate in Eq. 4.14 will be asymptotically biased if $\Omega^{(l)}$ is a proper subset of Ω, because the asymptotic limit of the perturbation estimate does not include contributions from conformations not belonging to $\Omega^{(l)}$. Also, since the contribution of the missed conformations is strictly positive, the asymptotic free energy increases as the conformational region $\Omega^{(l)}$ shrinks. Therefore, since the doublet reference system is expected to have a smaller set of accessible conformations than the singlet reference (item 5 of Section 3.2.4), in the asymptotic limit we have

$$\bar{F}_P \leq \bar{F}_P^{(1)} \leq \bar{F}_P^{(2)}\,. \qquad (4.18)$$

However, note that, due to the exponential in Eq. 4.17, the free energy is dominated by conformations with low physical energy. Therefore, if $\Omega^{(l)}$ contains these dominant conformations, the asymptotic bias will be low. Also, due to their larger Boltzmann factors, the dominant conformations are more likely to be sampled from the physical distribution. Therefore, if the conformational overlap between samples from the reference and physical distributions is high, then the bias is likely to be low and convergence faster.

In the present study, we assess conformational overlap in terms of overlap in the distributions of the force-field energies computed for the reference and the physical samples.

The overlap of the doublet reference with the physical distribution is likely to be greater than that of the singlet reference due to the incorporation of pair correlations. Therefore, the doublet free energy estimate is expected to converge more rapidly than the singlet estimate, though the asymptotic bias may be somewhat higher.

4.2.4 Boltzmann average using reference distributions

Although this chapter focuses on free energy calculation, it is of interest to point out a closely related potential application of the reference distributions. Samples from the reference distributions can also be used to compute the Boltzmann average of any function of the coordinates, such as the potential energy or the end-to-end distance of a linear chain molecule. The Boltzmann average of a function $f(\boldsymbol{\xi})$ is given by

$$\langle f \rangle = \frac{1}{Z_P} \int_{\Gamma_P} f\left(\boldsymbol{\xi}\right)\ \exp(-\beta\, U_P(\boldsymbol{\xi}))\, J(\boldsymbol{\xi})\ d\,\boldsymbol{\xi}\,. \tag{4.19}$$

which can be approximated as the discrete space sum

$$\langle f \rangle \approx\ \overline{\langle f \rangle} = \frac{1}{\bar{Z}_P} \sum_{\mathbf{X}_i \in\, \Omega} f\left(\mathbf{X}_i\right)\ \exp(-\beta\, \bar{U}_P(\mathbf{X}_i)) \tag{4.20}$$

where $\bar{U}_P(\mathbf{X})$ is given by Eq. 4.9. Multiplying and dividing by the l-level reference probability distribution and changing the summation to include only the accessible conformations gives

$$\begin{aligned}
\overline{\langle f \rangle} \approx \overline{\langle f \rangle}^{(l)} &= \frac{1}{\bar{Z}_P^{(l)}} \sum_{\mathbf{X}_i \in\, \Omega^{(l)}} f\left(\mathbf{X}_i\right)\ \frac{\tilde{p}_N^{(l)}\left(\mathbf{X}_i\right)}{\tilde{p}_N^{(l)}\left(\mathbf{X}_i\right)} \exp(-\beta\, \bar{U}_P(\mathbf{X}_i)) \\
&= \frac{1}{\bar{Z}_P^{(l)}} \sum_{\mathbf{X}_i \in\, \Omega^{(l)}} \tilde{p}_N^{(l)}\left(\mathbf{X}_i\right)\ f\left(\mathbf{X}_i\right)\ \exp\left(-\beta\, \Delta U^{(l)}(\mathbf{X}_i)\right) \\
&= \frac{\left\langle f\ \exp\left(-\beta\, \Delta U^{(l)}\right)\right\rangle_{\mathrm{ref}\,(l)}}{\left\langle \exp\left(-\beta\, \Delta U^{(l)}\right)\right\rangle_{\mathrm{ref}\,(l)}}
\end{aligned} \tag{4.21}$$

where $\bar{Z}_P^{(l)} = \exp(-\beta \bar{F}_P^{(l)})$ and $\bar{F}_P^{(l)}$ is given by Eq. 4.13. Finally, $\overline{\langle f \rangle}^{(l)}$ can be computed using samples from the reference distribution as

$$\overline{\langle f \rangle}^{(l)} \doteq \frac{\frac{1}{N_R} \sum_{n=1}^{N_R} f(\mathbf{X}_{i(n)}) \ \exp\left(-\beta \, \Delta U^{(l)}(\mathbf{X}_{i(n)})\right)}{\frac{1}{N_R} \sum_{n=1}^{N_R} \exp\left(-\beta \, \Delta U^{(l)}(\mathbf{X}_{i(n)})\right)} \tag{4.22}$$

where the denominator is same as the terms in the parenthesis of Eq. 4.14. As in the case of the free energy calculation above, in the asymptotic limit, the Boltzmann average computed using Eq. 4.22 will only include contributions from conformations accessible to the reference distribution. Also, based on Eq. 4.20, it can be seen that the convergence will be good if the accessible region includes conformations for which the product $f \ e^{-\beta \bar{U}_P}$ is large.

It is recognized that the Boltzmann average in Eq. 4.22 could also be computed directly by using the physical samples used to populate the reference pdfs. The present approach might nonetheless be useful because the reference sampling can be made computationally more efficient via distributed computing, and also more exhaustive, since it is not susceptible to getting trapped in local energy mimima of the physical energy surface. Also, in Chapter 5, we discuss a strategy for using a library of pdfs for constructing the reference distribution without using molecule specific simulation data (item 5 of Section 5.1), and a potential application of the Boltzmann average computed using the reference samples for computing the configurational entropy (item 3 of Section 5.2).

4.3 Methods

The procedure for setting up the singlet and doublet references pdfs using MD simulation data is identical to that described in Section 3.3. The additional steps for computing the free energy are:

(i) Sample N_R molecular conformations by singlet level sampling, and another N_R conformations by doublet level sampling, and compute the sampling probabilities associated with each set, that is, $\tilde{p}_N^{(1)}$ and $\tilde{p}_N^{(2)}$, respectively.

(ii) For each reference sample, compute the reference energy $U_R^{(l)}$ from Eq. 4.6, map the samples from the discretized BAT coordinate space to continuous BAT space using Eq. 3.21, compute the Jacobian using Eq. 4.3, construct the corresponding Cartesian molecular coordinates of the molecule, compute the force-field energy U_P, compute $\bar{U}_P$ from Eq. 4.9 and evaluate the energy difference $\Delta U^{(l)}$ from Eq. 4.12.

(iii) Finally, compute the free energy estimates, $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$, by applying Eq. 4.14 to the sets of energy differences corresponding to the samples from the two reference systems.

4.3.1 Molecular systems

We first validate the theory and implementation with tests on a simplified representation of all-atom propane (11 atoms), for which the free energy can be computed analytically. Starting with a standard force-field representation, we drop all nonbonded (Lennard-Jones and Coulombic) energy terms, as well as all bonded terms that do not correspond to the BAT coordinates used to specify the conformation. These simplifications decouple all internal coordinates so the multidimensional configurational integral factorizes into a product of one-dimensional integrals which may be computed analytically or numerically (see Appendix B) given the force-field parameters. However, propane is still a high-dimensional system, with 27 internal coordinates, and thus a useful test case. We then test the methodologies for full force-field representations of three peptides previously studied with a closely related method by Zhang *et. al.* [72]: alanine dipeptide (Ace-

Ala-Nme, 22 atoms, 60 BAT coordinates), dialanine (Ace-$(Ala)_2$-Nme, 32 atoms, 90 BAT coordinates) and tetra-alanine (Ace-$(Ala)_4$-Nme, 52 atoms, 150 BAT coordinates) where Ace is acetyl (CH_3-CO), Ala is Alanine (HN-C-CH_3-CO) and Nme is N-methylamide (NH-CH_3) (Figure 4.2). GAFF [11] force-field parameters were used in all cases. Force-field parameter files were generated with Amber AnteChamber [77] and converted to Gromacs format using amb2gmx from ffAMBER tools [78, 79]. Note that the force-field and temperature used in this study are different from those in Ref [72], so a quantitative comparison is not possible. For all molecules except propane, 10 sets of 50 ns vacuum

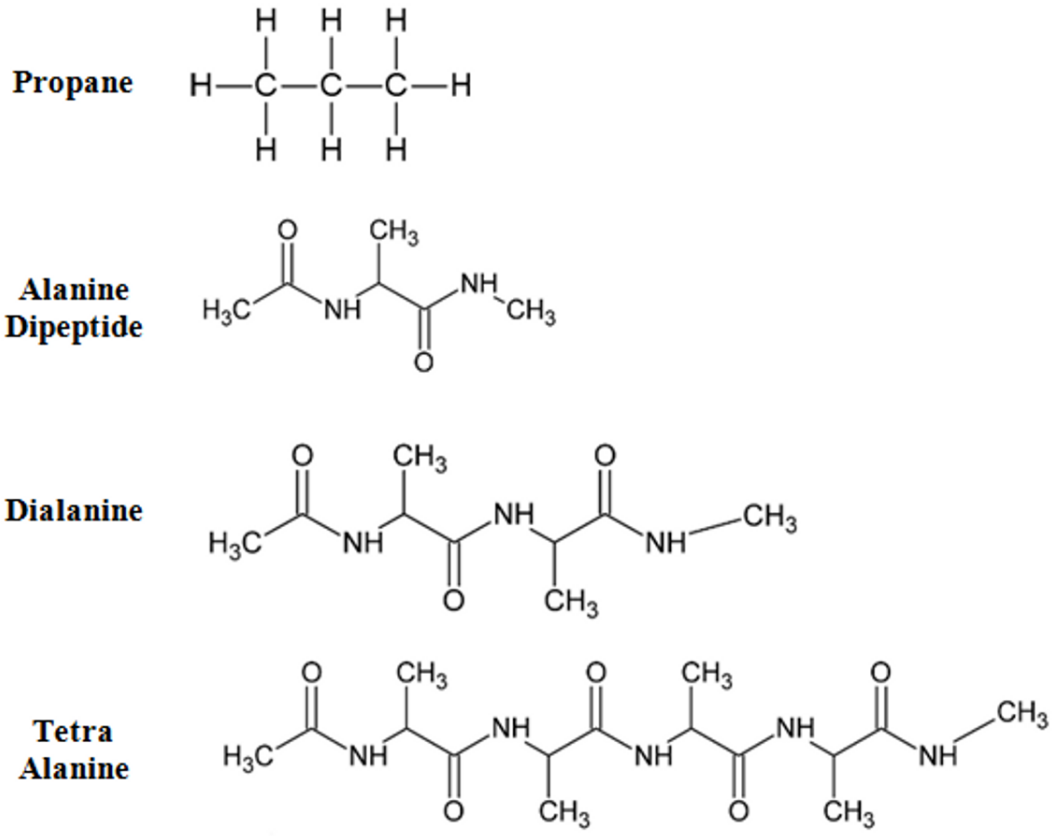


Figure 4.2: Chemical structures of the molecules used for testing the free energy method.

MD simulations were done at 1000 K using Gromacs 4.0.5 [80] with a time step of 1 fs; for propane, a single 100 ns run was carried out. Conformations were saved every picosecond to generate a total of 5×10^6 conformations for each peptide and 10^6 conformations for propane. Free energies are reported in units of k_BT (= 8.36 kJ/mol at 1000 K). Each BAT coordinate was discretized into B = 30 bins equally spaced between the minimum and maximum values found in the MD snapshots, and the coordinate snapshots were used to populate the singlet and doublet reference pdfs. These were used in turn to generate $N_R = 10^6$ samples for propane and $N_R = 5 \times 10^6$ samples for each peptide at both the singlet and doublet levels.

4.3.2 Assessment of free energy estimates

We monitor convergence of the two free energy estimates, $\hat{F}_P^{(1)}$ and $\hat{F}_P^{(2)}$, as a function of the number of reference samples N_R. Error analysis is done using the bootstrap method [81, 82] in which the original set of samples from the reference system are resampled with replacement to generate 100 new sets of samples, and the perturbation formula (Eq. 4.14) is applied to each data set. The mean and standard deviation of these 100 estimates are reported as the final free energy estimate and its uncertainty, respectively. We furthermore compare the distributions of the physical (force-field) energies of the original MD sampled and reference sampled conformations as a measure of the conformational overlap.

4.4 Results

4.4.1 Validation with simplified propane

Both free energy estimates for simplified propane converge to 74.77 k_BT, within 0.08 k_BT of the analytic free energy of 74.692 k_BT, as shown in Table 4.1, and appear to be well-converged, as shown in Figure 4.3 (the range of the vertical axis is 1 K_BT). Nonetheless, there is evidently a small bias in the estimates, which is likely due to the discretization and restriction of the conformational space, as discussed above, since the analytic partition function is computed from continuous integrals with full coordinate ranges instead of those observed in the simulation (see Appendix B). The positive sign of the small bias is consistent with the analysis of asymptotic bias in Section 4.2.3. Figure 4.4 plots the force-field energy distributions of conformations sampled from MD and the two references. The three energy distributions are virtually identical, indicating high similarity between the conformations sampled from the reference distributions and the physical distribution. Overall, the accuracy of the results for this simplified propane test, for which the free energy is available in a reliable analytic form, validates the theory and implementation of the free energy calculations.

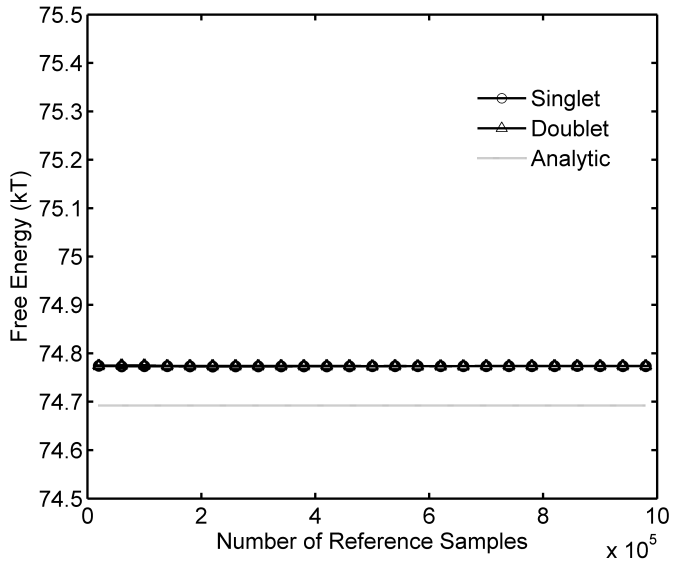


Figure 4.3: Convergence of free energy estimates for propane using a simplified force-field. The solid line indicates the mean of the estimate using 100 bootstrap samples (error bars of the singlet and doublet estimates are smaller than the thickness of the line).

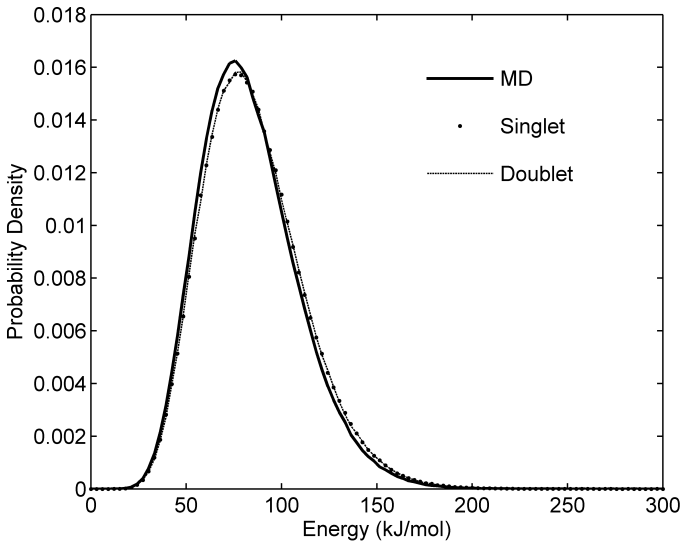


Figure 4.4: Normalized histogram (100 equally spaced bins) of potential energy of simplified propane conformations sampled by MD, singlet level sampling and doublet level sampling.

4.4.2 Peptides with full force-field representation

Analytic free energy values are not available for the three peptides (Figure 4.2) because the internal coordinates are coupled by force-field energy terms so that the high-dimensional configurational integral cannot be factorized. We therefore assess the reliability of the four free energy estimates obtained for each molecule (Table 4.1) by examining convergence plots (Figure 4.5) and the overlaps of the force-field energy distributions for the reference and the MD samples (Figure 4.6).

The central result of this study is that the doublet level reference systems lead to dramatically faster free energy convergence than the singlet level reference systems, as seen in Figure 4.5a-c and the bootstrap standard deviations in Table 4.1.

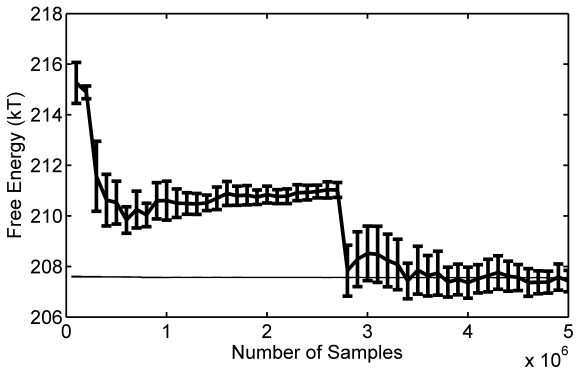
The excellent convergence of the doublet estimate is consistent with the strong overlap between the force-field energy distribution (Figure 4.6) of doublet reference samples with the MD samples from the physical Boltzmann distribution. The singlet reference systems yield much worse overlap with the physical distributions; indeed, the energies of over 99% of the singlet samples are greater than the maximum energy on the x-axis in Figure 4.6, so the singlet distributions are not graphed. These high energies result mainly from steric clashes. It is also worth remarking that, even for the doublet reference state, the fraction of high-energy samples increases with the size of the molecules, and the energy distribution correspondingly shifts toward higher energies. At the doublet level, the fraction of samples with energies greater than the maximum energy on the x-axis in Figure 4.6, were 0.9%, 2.4% and 10.5% for alanine dipeptide, dialanine and tetra-alanine, respectively.

Some of the convergence graphs display relatively long plateaus followed by sudden drops which occur whenever a sample with low energy difference is encountered [83]. This is

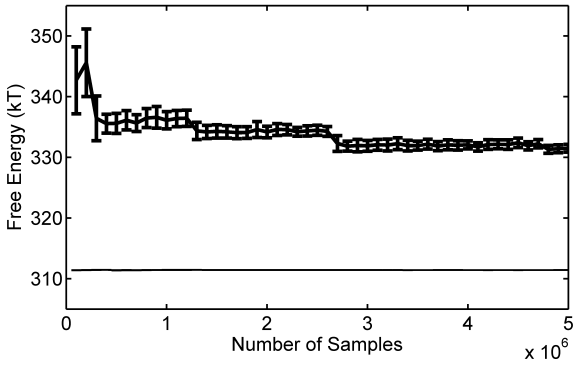
particularly evident in the singlet results for tetra-alanine (Figure 4.5). Such plateaus risk generating the deceptive appearance of a converged result. Thus, if a free energy estimate appears to be converged, but the energy overlaps are poor, then the apparent convergence may be illusory. On the other hand, if the overlap is extensive, then the convergence will be more credible.

Table 4.1: Mean and standard deviation (in parenthesis) of absolute free energy (in k_BT) of molecules from 100 bootstrap resampled data sets.

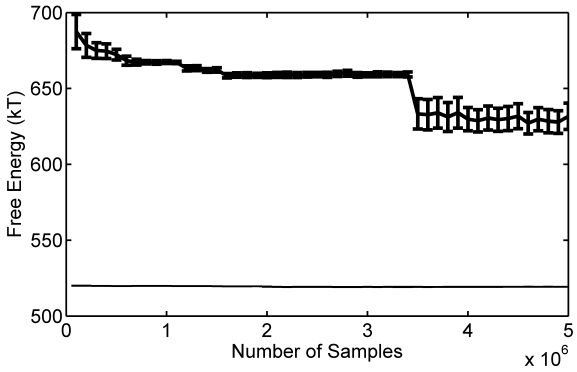
	Number of Atoms	Analytic	Singlet	Doublet
Simplified Propane	11	74.6920	74.77 (0.0002)	74.77 (0.0004)
Alanine Dipeptide	22	-	207.4 (0.8)	207.5 (0.005)
Dialanine	32	-	331.5 (1.3)	311.43 (0.01)
Tetra Alanine	52	-	631.7 (17.3)	519.3 (0.15)



(a) Alanine dipeptide

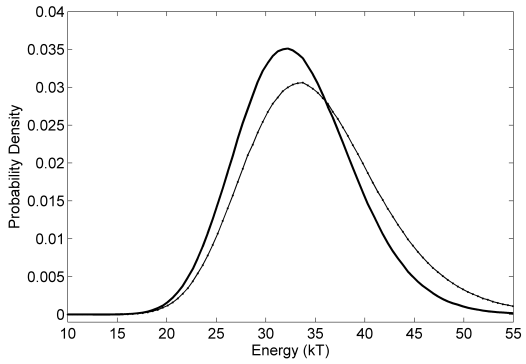


(b) Dialanine

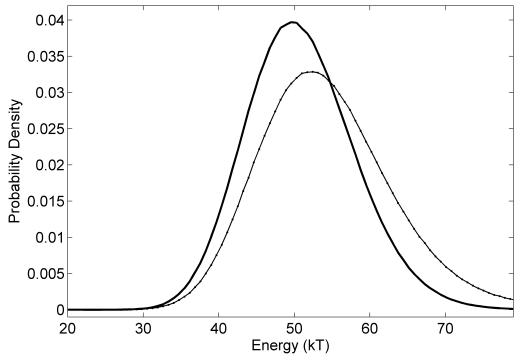


(c) Tetra-alanine

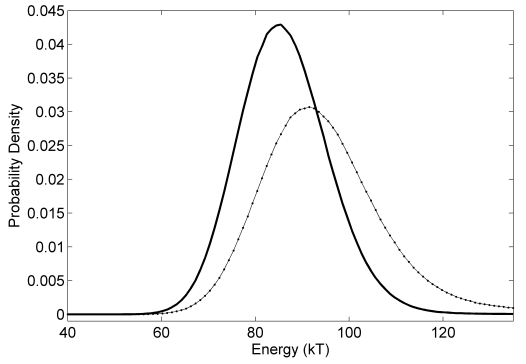
Figure 4.5: Convergence of free energy estimates using singlet (thick line) and doublet (thin line) reference system for (a) alanine dipeptide, (b) dialanine and (c) tetra-alanine. The solid line indicates the mean of the estimate using 100 bootstrap resampled data sets and the bars indicate the standard deviations (error bars on the doublet estimates are smaller than the thickness of the line).



(a) Alanine dipeptide



(b) Dialanine



(c) Tetra-alanine

Figure 4.6: Probability distribution of potential energy of conformations sampled by MD and by doublet level sampling for (a) alanine dipeptide: energies of 99.9% singlet samples and 0.1% doublet samples was greater than 500 kJ/mol, (b) dialanine: energies of all singlet samples and 1.3% doublet samples was greater than 800 kJ/mol, and (c) tetra-alanine: energies of all singlet samples and 10% doublet samples was greater than 1200 kJ/mol.

4.5 Discussion

This chapter has described a reference system method for computing the absolute free energy of a molecule, which achieves convergence with a relatively small number of conformational samples. The present approach builds on important prior work [69], in which simulations were used to build singlet pdfs of internal molecular coordinates and these pdfs in turn were used to define a reference system with a known free energy. The singlet level reference state was then used as a basis for the calculation of molecular free energies. The innovation in the present study grows from the use of the SA-based conformational sampling method (Chapter 3), which incorporates correlations of any given order, through the use of superposition approximations.

It is worth elaborating on the two sources of bias in the perturbation free energy estimates computed here. In order to set up the reference distributions, the conformational space is discretized. This leads to a bias since the method effectively computes the discrete-space approximation (Eq. 4.10) of the continuous configurational integral (Eq. 4.1). However, the results for propane, where the deviation of the computed discrete-space approximation from the continuous-space analytic free energy is $<0.1\%$, suggest that the bias due to the discretization is small. Based on these propane results, we expect the discretization bias for peptides to be small as well, since the force-field and discretization are similar. The second source of bias stems from the fact that the reference distributions are constructed from a finite set of MD data which restricts the conformational space accessible to SA-based sampling. However, although difficult to assess *a priori*, the bias due to the restrictions on the conformational space is expected to be low if the accessible region contains the low physical energy conformations that dominate the configurational integral.

As noted in Section 2.4.3, the present SA framework can be generalized to construct reference systems based upon selected second and higher order pdfs. It is thus worth noting that there can be a trade-off between the bias and convergence of the free energy estimate. As more pdfs are included, the conformational overlap increases, thereby speeding convergence, but the bias may increase because it is increasingly difficult to populate higher-order pdfs so the reference distribution may have more holes. The optimal SA-based reference will be one that uses the least number of higher order pdfs while maintaining sufficient overlap to achieve convergence with reasonable number of reference samples.

In prior work [69], the use of a singlet level reference system was found to limit the applicability of the reference system method to small molecules, because the overlap of the reference distribution with the physical ensemble falls with increasing system size. This limitation motivated the development of an innovative approach where the molecule is divided into fragments and samples are drawn separately for each fragment [72]. The fragments are then assembled to obtain the free energy of the full molecule. The present study shows that incorporating correlation into the reference system yields good convergence for larger molecules without recourse to the fragment-based method. Ultimately, a combined approach may be of value for still larger molecules, especially when correlations above second order are expected to be important.

One limitation of the present method is that it will likely need to be used in conjunction with an implicit solvent model, in order to limit the number of degrees of freedom to a computationally manageable number [84, 85, 86]. Indeed, even with an implicit solvent model, the computational requirements will be significant for most systems large enough to be of practical interest, such as proteins. Possible strategies for scaling

up the conformational sampling method to be applicable to much larger systems, such as small proteins, are discussed in Chapter 5.

Chapter 5

Future Directions

The SA-based conformational sampling method developed here generates physically reasonable conformations from a region of conformational space whose boundaries are set based on an initial set of Boltzmann distributed samples from a standard molecular simulation. Low order marginal pdfs are computed from the simulation data and used to construct an approximation to the full Boltzmann pdf. We saw that this SA-based distribution can have high overlap with the physical Boltzmann distribution. Moreover, since the sampling distribution is normalized, the samples can be reweighted to compute thermodynamic quantities of the system, such as the free energy, as seen in Chapter 4.

Multiple properties of the SA-based sampling make it potentially more attractive than simply drawing additional samples from the Boltzmann distribution via MD or MC methods. First, because the SA-based sampling does not account for higher order correlations (i.e., correlations of order greater than l), the range of conformations accessible to it is larger than that of the simulation data used to populate the reference pdfs. Second, the sampling is not obstructed by energy barriers on the physical energy surface, so highly dissimilar, yet low energy conformations can readily be sampled. Third, from a computational standpoint, SA-based sampling is naturally suited for distributed computing, so that the time required for generating a given number of samples can be

easily reduced. It is thus worth considering how this approach may be applied to real-world problems. In this chapter, then, we discuss strategies to extend the "proof-of-principle" studies, presented in the prior chapters, to larger molecular systems, and also further potential applications of these methods.

5.1 Accuracy and scale-up of SA-based sampling

A number of strategies could be employed to allow productive application of SA-based sampling to larger molecular systems, potentially small proteins with ∼2000 atoms. The effectiveness of the various strategies, and their combinations, should be assessed based upon the degree of speedup and memory savings they afford, and their conformational overlap with the physical distribution. The degree of overlap can be evaluated as done in Chapter 3, and also based upon the convergence properties of reference state free energy calculations as presented in Chapter 4.

(i) **Mixed-SA with select higher order reference** pdfs

The l-level sampling algorithm presented in Chapter 3 uses all reference pdfs of up to order l, leading to order N^l scaling of both the CPU time and memory requirements for sampling a conformation. This scaling will lead to prohibitive computational costs for sampling larger systems, primarily due to the prohibitive memory requirements at triplet and higher levels.

However, the number of higher order pdfs required to build an SA-based approximation may be reduced substantially by dropping the pdfs that correspond to weak correlations and that, therefore, are not critical for maintaining good overlap with the physical distribution. For instance, in a long chain molecule, joint distribution

of torsions that are far apart in sequence and three-dimensional space might not be strong. To identify the stronger correlations, heuristic rules such as close proximity in the bonded topology or in three dimensional space could be used, though these rules will likely be molecule-dependent. A more sophisticated and automatic approach could be to use low mutual information as an indicator of weak correlation within sets of coordinates [39, 87]. Once the pdfs that are least important, in this sense, have been identified, a mixed SA reference state can be set up by approximating the dropped pdfs in terms of their lower order marginals (see Section 2.4.3). Note that reducing the number of higher order reference pdfs by this approach will also reduce the cost of sampling a single conformation, since fewer multiplication operations will be required for computing the conditional distributions.

(ii) **Distributed computing**

The main computational tasks involved in the SA-based sampling are highly amenable to distributed computing. Thus, the MD or MC physical samples required to populate the reference pdfs can be generated by short simulations on multiple compute nodes, and the sampling from SA-based distributions can be trivially distributed over multiple compute nodes. Indeed, since successive samples are uncorrelated, no internode communication will be required during this sampling process.

(iii) **Enhanced sampling methods for generating physical samples**

To ensure sampling of the desired region of the conformational space, the physical samples used to populate the reference pdfs should roughly cover the region corresponding to the state of interest, e.g. the folded structure of a protein. The

coverage of the physical simulation may be improved by using enhanced sampling methods, such as temperature replica exchange [88, 89], other Generalized Ensemble methods [90, 91] and Markov State Models [92]; see Ref [93, 94, 95] for recent reviews. Moreover, the enhanced sampling algorithms in conjunction with modern computer hardware technologies [96, 97] and large scale distributed computing [98] can further help to obtain a good set of initial physical samples.

(iv) **Parametric representation of reference pdfs**

As noted above, the histogram representation of the reference pdfs can be expensive in terms of memory requirement. This cost could be moderated by representing the higher order marginals with a more sophisticated basis set, such as Gaussian or von Mises distributions, which require far fewer parameters than needed by the present histogram representation. Maximum-likelihood methods represent one possible approach to fitting these parametric models to the available simulation data [50].

(v) **Using generic pdfs for analogous coordinates**

The SA-based sampling method, as presented in previous chapters, requires a preliminary MD simulation of the molecule to set up the reference pdfs. This step limits the flexibility of the method, as MD simulations can be expensive and difficult to automate. These simulations would become unnecessary if we could establish a generic set of pdfs which can be used to construct the SA-based distributions for arbitrary molecules. This idea is based on the expectation that correlations among certain combinations of internal coordinates may be similar across molecules in the same chemical family. For instance, referring to the polypeptide test cases

(Figure 4.2), the correlations among backbone torsion angles of alanine dipeptide could be used for analogous coordinates in other polypeptides. In general, for proteins, correlations among adjacent backbone torsion may largely be determined by the local secondary structure. Since setting up the pdfs would then become one-time "offline" calculation, extensive computational resources could be devoted to it. For example, exhaustive simulations could be performed for short polypeptides in different secondary structure motifs to build a library of pdfs for proteins. A caveat to using a library based approach is that the various reference pdfs might not be consistent with one another, e.g., the 1-D marginal, $p(X_1)$, from $p(X_1, X_2)$ and $p(X_1, X_3)$ may not match, at least without additional steps to provide for consistency. Such inconsistencies might increase the fraction of null samples.

In a related approach, for a family of small molecules which differ by a few atoms, e.g. by varying an R-group, such as a congeneric series of drug-like ligands, one could build a "super molecule" using dummy atoms, such that the various compounds are obtained by deleting a subset of atoms of the super molecule. This strategy will work if the simulation of the super molecule is able to access conformations that are likely to be sampled by MD simulations of the individual molecules. This can be accomplished by using an "enveloping" potential energy function for simulating the super molecule, which would be obtained by combining the potential energy function of individual ligands [99], or by simply softening the barriers in the torsional energy terms of the super molecule. In this approach, all pdfs will be mutually consistent.

(vi) **Collective coordinates and coarse-grained representation**

Another approach for improving the accuracy of the sampled conformations is suggested by the dependence of the present results upon the choice of coordinate

system, since different coordinates systems result in different degrees of correlation, or coupling, among coordinates. Accordingly, other coordinate systems, such as principal components of the MD trajectory in Cartesian or BAT [100] coordinates, might better capture complex molecular fluctuations and other high-dimensional distributions in terms of tractable sets of low-order marginals.

A straightforward approach to reducing the dimensionality of a molecular system, e.g. protein, would be to use a coarse-grained representation and energy function. A coarse-grained approach will suffice if only the large scale conformational states, e.g. different loop conformations, of a protein are of interest, and not details such as side-chain orientations.

5.2 Applications

The SA-based conformational sampling algorithms and the associated absolute free energy calculation method could be useful for the following applications:

(i) **Conformational equilibrium**

The absolute free energy calculation method can be used to compute relative population for different conformational states [69, 101, 70, 71] which may be separated by high energy barriers. Consider a molecule, e.g. leucine dipeptide, which can take either an alpha or beta state depending on the values of the backbone torsions [69]. The two states essentially represent a partitioning of the conformational space. The relative population of the two states can be obtained in terms of the difference in the absolute free energies of the two states which can be obtained by populating the reference pdfs using conformations from a MD simulation confined

to the particular state [70]. Direct calculation of the relative population based on a converged equilibrium simulation could be more computationally expensive, since interconversion between the two states is restricted by the high energy barriers.

(ii) **Binding free energy and configurational integral calculation**

The methods developed here could be applied to the calculation of noncovalent binding affinities, a topic of active interest due to the importance of such methods in fields like drug discovery and catalysis. The reader is referred to Ref [102, 103, 19, 104, 105] for an overview and current status of the field. The reaction free energy for the non-covalent association of two molecules - i.e., their binding free energy – is

$$\Delta F_{bind} = F_P^{LR} - F_P^L - F_P^R \tag{5.1}$$

where the terms of the right-hand side denote the absolute free energy of ligand (L), receptor (R) and the complex (LR) [19]. (More formally, these quantities are standard chemical potentials, corresponding to a hypothetical dilute 1M solution [24].) The effect of solvent can be approximated by adding an implicit solvent energy term to the force-field potential energy function [84, 85, 86]. To compute the free energy of the complex, the internal coordinates will include six pseudo internal coordinates that specify the orientation and position of the ligand relative to the receptor. The distributions corresponding to these internal coordinates determine the "wiggle room" for the ligand in the binding pocket. Such an approach could potentially be automated and used for virtual screening of a set of ligands against a target molecule, a challenging task in computer aided drug design [106]. It would be interesting to compare the resulting binding free energies with those obtained by other computational approaches and with experimental measurements [62, 19, 103, 105].

(iii) **Configurational entropy**

This work was motivated by the observation that the configurational entropy of molecules could be computed rather accurately using the MIE at the doublet or triplet level. In Chapter 2, we saw that the l-level MIE of entropy, $S^{(l)}$, is the cross entropy of the SA-l distribution with respect to the Boltzman distribution. Another estimate of the configurational entropy, $\tilde{S}^{(l)}$, could be obtained as the cross-entropy of the SA-l based sampling distribution, with the Boltzmann distribution. Since the $\tilde{p}_N^{(l)}$distributions are normalized, unlike the superposition approximations, $\tilde{S}^{(l)}$ provides an upper bound on the true entropy, based on the Kullbach-Liebler inequality (Eq. 2.32). Moreover, we have seen that the doublet or triplet level sampling distribution can be a good approximation of the Boltzmann distribution. Therefore, we would expect the cross-entropy of the present sampling distributions to give a tighter bound as l increases. The $\tilde{S}^{(l)}$ estimate is essentially the Boltzmann average of the log of sampling probability

$$\tilde{S}^{(l)} = \left\langle -\ln \tilde{p}^{(l)} \right\rangle \tag{5.2}$$

which can be computed using reference samples by Eq. 4.22 of Section 4.2.4 with $f = -\ln \tilde{p}_N^{(l)}$. Thus, there is an interesting possibility of computing a novel estimator of the entropy which, like the MIE, is expected to converge on the true entropy but, unlike the MIE, always represents an upper limit of the true entropy.

(iv) **Conformational Search**

SA-based sampling could be used as part of a conformational search method in which sampled conformations are energy-minimized to find conformations corresponding to the low-energy minima of the physical energy surface. This concept takes

advantage of the fact that the SA-based sampling randomly combines different values of the coordinates from their permitted ranges allowing it to visit multiple energy wells. Such an approach could speed the discovery of stable, bioactive conformations of drugs and drug-like molecules. It could also help identify stable conformations of a protein-ligand complex in settings where there is reason to expect different conformational rearrangements of the binding site in response to different bound ligands. The SA-based samples could also be used to construct the a reservoir of conformations for use in Reservoir Replica Exchange [107], and in conformational analysis using higher level quantum chemistry calculation on the low-energy conformations of the empirical force-field [108].

Finally, we note that approximating a high-dimensional distribution in terms of low-order marginals is a common theme in the many fields of inquiry [50, 109, 110]. Although one often has enough data to compute low-order marginals, the available data are typically too sparse to allow a full, high-dimensional distribution to be evaluated. The work describes a novel approach to approximate the high-dimensional distribution in terms of tractable low-order marginals and, furthermore, to sample from this approximate distribution and compute averages with respect to the true distribution. This work may thus have applications in fields beyond statistical mechanics, including bioinformatics, structural biology, data mining, and machine learning.

Appendix A

Exponents of the superposition approximation at level l

The SA-l distribution, $p_N^{(l)}$, approximates an N-dimensional distribution p_N in terms of products of its marginal pdfs of highest order $l < N-1$ and has the form

$$\begin{aligned} p_N \approx p_N^{(l)} &= \mathscr{P}_{(N,l)}^{a(l;N,l)} \mathscr{P}_{(N,l-1)}^{a(l-1;N,l)} \times \ldots \times \mathscr{P}_{(N,l)}^{a(1;N,l)} \\ &= \prod_{j=l}^{1} \mathscr{P}_{(N,j)}^{a(j;N,l)} \end{aligned} \tag{A.1}$$

where $\mathscr{P}_{(N,k)}$ denotes the the product the C_k^N marginal distributions at order k of p_N. In this appendix, we use reverse induction to show that the SA-l for a general N and l is given by

$$p_N^{(l)} = \mathscr{P}_{(N,l)}^{+1} \mathscr{P}_{(N,l-1)}^{-(N-l)} \mathscr{P}_{(N,l-2)}^{+\frac{(N-l)(N-l+1)}{2!}} \times \ldots \times \mathscr{P}_{(N,1)}^{(-1)^{l-1}\frac{(N-l)\ldots(N-2)}{(l-1)!}} \tag{A.2}$$

so that the exponents in Eq. A.1 are given by

$$a(j;N,l) = (-1)^{l-j} \prod_{i=1}^{l-j} \frac{N-l+i-1}{i}\,. \tag{A.3}$$

The SA-l distribution is obtained by recursive application of the GKSA

$$\begin{aligned} p_k \approx p_k^{(k-1)} &= \mathscr{P}_{(k,k-1)}^{+1} \mathscr{P}_{(k,k-2)}^{-1} \times \ldots \times \mathscr{P}_{(k,1)}^{(-1)^{k-2}} \\ &= \prod_{j=k-1}^{1} \mathscr{P}_{(k,j)}^{(-1)^{k-1-j}} \end{aligned} \tag{A.4}$$

which expresses a k-dimensional pdf in terms of marginal pdfs of up to order $k-1$. The overall strategy of the proof is to show that applying the GKSA to each l-order pdf in the SA-l gives the SA-l for $l = l-1$.

We first derive an approximation, using the GKSA, to the product of all k-order pdfs in terms of producst of $k-1$ and lower order pdfs which has the form

$$\begin{aligned} \mathscr{P}_{(N,k)} &\approx \mathscr{P}_{(N,k-1)}^{b(k-1;N,k)} \mathscr{P}_{(N,k-2)}^{b(k-2;N,k)} \times \ldots \times \mathscr{P}_{(N,1)}^{b(1;N,k)} \\ &= \prod_{j=k-1}^{1} \mathscr{P}_{(N,j)}^{b(j;N,k)} . \end{aligned} \tag{A.5}$$

The exponents $b(j;N,k)$ are obtained as follows. The quantity $\mathscr{P}_{(N,k)}$ is a product of C_k^N pdfs of order k corresponding to the unique combinations of k variables out of the full N variables. On applying the GKSA, each k-order pdf in $\mathscr{P}_{(N,k)}$ generates C_j^k pdfs of order $j \leq (k-1)$. Therefore, the total number of pdfs generated at order j are

$$G_j = C_k^N \times C_j^k . \tag{A.6}$$

Given N variables, the number of possible pdfs at order j is $T_j = C_j^N$. Due to symmetry of the GKSA, G_j is a multiple of T_j and the ratio gives the magnitude of $b(j;N,k)$ in Eq. A.5,

$$\begin{aligned} |b(j;N,k)| &= \frac{G_j}{T_j} \\ &= \frac{C_k^N \times C_j^k}{C_j^N} \\ &= \frac{\frac{N!}{(N-k)!\ k!} \times \frac{k!}{(k-j)!\ j!}}{\frac{N!}{(N-j)!\ j!}} \\ &= \frac{(N-j)!}{(N-k)!(k-j)!} . \end{aligned} \tag{A.7}$$

To find the sign of $b(j;N,k)$, note that, in GKSA (Eq. A.4), the sign of the exponent of product of pdfs at the highest order, $\mathscr{P}_{(N,N-1)}$, is positive, and it alternates for pdfs of

subsequent orders. Therefore, in Eq. A.5 the exponent $b(k-1;N,k)$ is positive, $b(k-2;N,k)$ is negative, etc and the exponents in Eq. A.5 are given by

$$b(j;N,k) \;=\; (-1)^{k-1-j}\frac{(N-j)!}{(N-k)!(k-j)!}\,. \tag{A.8}$$

To illustrate the pattern of the exponents $b(j;N,k)$, we list the expressions of $\mathscr{P}_{(N,k)}$ for $k=2$ to 5:

$$\begin{aligned}
\mathscr{P}_{(N,2)} &= \mathscr{P}_{(N,1)}^{\frac{N-1}{1}} \\
\mathscr{P}_{(N,3)} &= \mathscr{P}_{(N,2)}^{\frac{N-2}{1}}\mathscr{P}_{(N,1)}^{-\frac{(N-2)(N-1)}{2!}} \\
\mathscr{P}_{(N,4)} &= \mathscr{P}_{(N,3)}^{\frac{N-3}{1}}\mathscr{P}_{(N,2)}^{-\frac{(N-3)(N-2)}{2!}}\mathscr{P}_{(N,1)}^{+\frac{(N-3)(N-2)(N-1)}{3!}} \\
\mathscr{P}_{(N,5)} &= \mathscr{P}_{(N,4)}^{\frac{N-4}{1}}\mathscr{P}_{(N,3)}^{-\frac{(N-4)(N-3)}{2!}}\mathscr{P}_{(N,2)}^{+\frac{(N-4)(N-3)(N-2)}{3!}}\mathscr{P}_{(N,1)}^{-\frac{(N-4)(N-3)(N-2)(N-1)}{4!}} \\
&\;\;\vdots \\
\mathscr{P}_{(N,l)} &= \mathscr{P}_{(N,l-1)}^{\frac{N-l+1}{1}}\mathscr{P}_{(N,l-2)}^{-\frac{(N-l+1)(N-l+2)}{2!}}\mathscr{P}_{(N,l-3)}^{+\frac{(N-l+1)(N-l+2)(N-l+3)}{3!}}\times\ldots\times\mathscr{P}_{(N,1)}^{-\frac{(N-l+1)\ldots(N-1)}{(l-1)!}}\,.
\end{aligned} \tag{A.9}$$

Substituting $\mathscr{P}_{(N,l)}$ from Eq. A.9 for the first factor in the RHS of the SA-l in Eq. A.2 gives

$$\begin{aligned}
&\Big(\quad \mathscr{P}_{(N,l-1)}^{\frac{N-l+1}{1}}\mathscr{P}_{(N,l-2)}^{-\frac{(N-l+1)(N-l+2)}{2!}}\times\ldots\times\mathscr{P}_{(N,1)}^{(-1)^{l-1}\frac{(N-l+1)\ldots(N-1)}{(l-1)!}}\Big) \\
\times\;&\Big(\mathscr{P}_{(N,l-1)}^{-(N-l)}\mathscr{P}_{(N,l-2)}^{+\frac{(N-l)(N-l+1)}{2!}}\times\ldots\times\mathscr{P}_{(N,1)}^{(-1)^{l-2}\frac{(N-l)\ldots(N-2)}{(l-1)!}}\Big) \\
=\;&\mathscr{P}_{(N,l-1)}^{+1}\mathscr{P}_{(N,(l-1)-1)}^{-(N-(l-1))}\times\ldots\times\mathscr{P}_{(N,1)}^{(-1)^{(l-1)-1}\frac{(N-(l-1))\ldots(N-2)}{((l-1)-1)!}} \\
=\;&p_N^{(l-1)}
\end{aligned} \tag{A.10}$$

which is the SA-l for $l=l-1$. Continuing this recursion l times we get the $p_N^{(1)}$ as $\mathscr{P}_{(N,1)}$, product of all 1-D pdfs, which is true. QED.

Appendix B

Exact free energy for simplified propane

We derive the partition function of propane, with $M = 11$ atoms, for a simplified energy function lacking non-bonded energy terms and possessing a single energy term corresponding to each of the $N = 3 \times M - 6 = 27$ BAT coordinates (Figure 4.1) used to specify the conformation. Thus, the energy function is given by

$$U_P(\mathbf{b}, \mathbf{a}, \mathbf{t}) = \sum_{i=2}^{M} U_b(b_i) + \sum_{i=3}^{M} U_a(a_i) + \sum_{i=4}^{M} U_t(t_i) \tag{B.1}$$

Harmonic functions are used for bond-stretch and angle-bend energy terms,

$$\begin{aligned} U_b(b) &= \frac{1}{2} k_b (b - b_{eq})^2 \\ U_a(a) &= \frac{1}{2} k_a (a - a_{eq})^2 \end{aligned} \tag{B.2}$$

and the Ryckaert-Bellemans [111] potential is used for torsions,

$$U_t(x) = \sum_{i=0}^{5} \mathrm{C}_i \left(\cos\left(t - \pi\right)\right)^i \tag{B.3}$$

where k_B, b_{eq}, a_{eq}, k_AandC_i are the force-field parameters. The configurational integral is then given by

$$\begin{aligned} Z_P &= \int \exp(-\beta\, U_P(\mathbf{b},\mathbf{a},\mathbf{t}))\, dV \\ &= \int \exp\left(-\beta \sum_{i=2}^{M} U_b(b_i)\right) \exp\left(-\beta \sum_{i=3}^{M} U_a(a_i)\right) \exp\left(-\beta \sum_{i=4}^{M} U_t(t_i)\right) \\ &\qquad b_3 \prod_{i=4}^{M} \left(b_i^2\, \sin a_i\right)\, db_2 db_3\, da_3\, \left(\prod_{i=4}^{M} db_i\, da_i\, dt_i\right) \end{aligned} \tag{B.4}$$

Grouping bond, angle and torsion variables gives

$$\begin{aligned} Z_P = &\int \exp\left(-\beta \sum_{i=2}^{M} U_b(b_i)\right)\, b_3\, b_4^2 .. b_M^2\, db_2 .. db_M \\ &\times \int \exp\left(-\beta \sum_{i=3}^{M} U_a(a_i)\right)\, \sin a_4 .. \sin a_M\, da_3 .. da_M \\ &\times \int \exp\left(-\beta \sum_{i=4}^{M} U_t(t_i)\right) dt_4 .. dt_M \end{aligned} \tag{B.5}$$

By separating variables further, we can write the partition function as a product of one-dimensional integrals of the form

$$\begin{aligned} \text{Bonds:}\quad & \int_0^{\infty} \exp(-\beta\, U_b(b))\, db \;\; ; \;\; \int_0^{\infty} b\, \exp(-\beta\, U_b(b))\, db \;\; ; \;\; \int_0^{\infty} b^2\, \exp(-\beta\, U_b(b))\, db \\ \text{Angles:}\quad & \int_0^{2\pi} \exp(-\beta\, U_a(a))\, da \;\; ; \;\; \int_0^{2\pi} \sin(a)\, \exp(-\beta\, U_a(a))\, da \\ \text{Torsions:}\quad & \int_0^{2\pi} \exp(-\beta\, U_t(t))\, dt \end{aligned} \tag{B.6}$$

all of which, except the integral over torsion coordinates can be computed analytically. The force-field parameters for propane are as follows. Parameters for the two bond types are

$$\begin{aligned} &\text{C-H}: \;\; k_b = 2.8225 \times 10^5\,, \quad b_{eq} = 0.1092 \\ &\text{C-C}: \;\; k_b = 2.5363 \times 10^5\,, \quad b_{eq} = 0.1535 \end{aligned} \tag{B.7}$$

The three angle types have parameters

$$\begin{aligned}
\text{H-C-C or C-C-H} &: k_a = 3.8828 \times 10^2\,, \quad a_{eq} = 1.9207 \\
\text{C-C-C} &: k_a = 3.8828 \times 10^2\,, \quad a_{eq} = 1.9309 \\
\text{H-C-H} &: k_a = 3.8828 \times 10^2\,, \quad a_{eq} = 1.8911
\end{aligned} \tag{B.8}$$

The units of spring constants, k_b and k_a, are in kJ/mol/nm2, equilibrium bond lengths, b_{eq} are in nm and equilibrium angles, a_{eq}, are in radians. Finally parameters for the two torsion types, in kJ/mol, are

$$\begin{aligned}
\text{H-C-C-C or C-C-C-H} &: \mathrm{C}_0 = 0.66944,\ \mathrm{C}_1{=}2.00832,\ \mathrm{C}_3 = -2.67776 \\
\text{H-C-C-H} &: \mathrm{C}_0 = 0.62760,\ \mathrm{C}_1{=}1.88280,\ \mathrm{C}_3 = -2.51040 \\
, &\qquad (\,\mathrm{C}_2 = \mathrm{C}_4{=}\,\mathrm{C}_5 = 0\,)
\end{aligned} \tag{B.9}$$

Substituting these parameters with $k_BT = 8.36$ kJ/mol corresponding to 1000 Kelvin gives the final free energy (in k_BT units) as

$$\begin{aligned}
F_P &= -\ln Z_P \\
&= 74.6920
\end{aligned} \tag{B.10}$$

Bibliography

[1] W. L. Jorgensen, "The many roles of computation in drug discovery," *Science*, vol. 303, no. 5665, p. 1813, 2004.

[2] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*. McGraw-Hill New York, 1989.

[3] D. A. McQuarrie, *Quantum chemistry*. Univ Science Books, 2008.

[4] H. M. Senn and W. Thiel, "QM/MM methods for biomolecular systems," *Angewandte Chemie International Edition*, vol. 48, no. 7, p. 1198, 2009.

[5] D. G. Truhlar, J. Gao, C. Alhambra, M. Garcia-Viloca, J. Corchado, M. L. Sànchez, and J. Villà, "The incorporation of quantum effects in enzyme kinetics modeling," *Accounts of Chemical Research*, vol. 35, no. 6, p. 341, 2002.

[6] A. Warshel, "Computer simulations of enzyme catalysis: Methods, Progress, and Insights," *Annual Review of Biophysics and Biomolecular Structure*, vol. 32, no. 1, p. 425, 2003.

[7] K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. Wollacott, L. M. Westerhoff, and K. M. Merz Jr, "The role of quantum mechanics in structure-based drug design," *Drug Discovery Today*, vol. 12, no. 17-18, pp. 725–731, 2007.

[8] A. J. Mulholland, "Modelling enzyme reaction mechanisms, specificity and catalysis," *Drug Discovery Today*, vol. 10, no. 20, p. 1393, 2005.

[9] J. W. Ponder and D. A. Case, "Force fields for protein simulations.," *Advances in Protein Chemistry*, vol. 66, p. 27, 2003.

[10] V. M. Anisimov, G. Lamoureux, I. V. Vorobyov, N. Huang, B. Roux, and A. D. MacKerell, Jr, "Determination of electrostatic parameters for a polarizable force field based on the classical Drude oscillator," *Journal of Chemical Theory and Computation*, vol. 1, no. 1, p. 153, 2005.

[11] J. Wang, R. Wolf, J. Caldwell, P. Kollman, and D. Case, "Development and testing of a general amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, pp. 1174, 1157, 2004.

[12] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 3, p. 712, 2006.

[13] A. D. Mackerell Jr, "Empirical force fields for biological macromolecules: overview and issues," *Journal of Computational Chemistry*, vol. 25, no. 13, p. 1584, 2004.

[14] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. M. Jr., "CHARMM general

force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *Journal of Computational Chemistry*, vol. 31, no. 4, p. 671, 2010.

[15] T. A. Halgren, "Potential energy functions," *Current Opinion in Structural Biology*, vol. 5, no. 2, p. 205, 1995.

[16] A. Rahman and F. H. Stillinger, "Molecular Dynamics Study of Liquid Water," *The Journal of Chemical Physics*, vol. 55, no. 7, p. 3336, 1971.

[17] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins.," *Nature*, vol. 267, no. 5612, p. 585, 1977.

[18] E. Shakhnovich, "Modelling protein folding: the beauty and power of simplicity," *Folding and Design*, vol. 1, no. 3, p. R50, 1996.

[19] M. K. Gilson and H.-X. Zhou, "Calculation of Protein-Ligand binding affinities," *Annual Review of Biophysics and Biomolecular Structure*, vol. 36, pp. 21–42, 2007.

[20] D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformational ensembles in biomolecular recognition," *Nature Chemical Biology*, vol. 5, no. 11, p. 789, 2009.

[21] P. I. Zhuravlev and G. A. Papoian, "Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework," *Quarterly Reviews of Biophysics*, p. 1, 2006.

[22] G. Kar, O. Keskin, A. Gursoy, and R. Nussinov, "Allostery and population shift in drug discovery," *Current Opinion in Pharmacology*, vol. 10, no. 6, p. 715, 2010.

[23] D. Chandler, *Introduction to Modern Statistical Mechanics*. Oxford University Press, USA, 1st ed., 1987.

[24] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, "The statistical-thermodynamic basis for computation of binding affinities: a critical review," *Biophysical Journal*, vol. 72, no. 3, p. 1047, 1997.

[25] J. G. Kirkwood and E. M. Boggs, "The radial distribution function in liquids," *The Journal of Chemical Physics*, vol. 10, p. 402, 1942.

[26] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids, Third Edition*. Academic Press, 3rd ed., 2006.

[27] I. Z. Fisher and B. L. Kopeliovich, "Improvement of superposition approximation in the theory of liquids," *Doklady Akademii Nauk SSSR*, vol. 133, p. 81, 1960.

[28] H. Reiss, "Superposition approximations from a variation principle," *Journal of Statistical Physics*, vol. 6, no. 1, pp. 39–47, 1972.

[29] N. N. Bugaenko, A. Gorban', and I. Karlin, "Universal expansion of three-particle distribution function," *Theoretical and Mathematical Physics*, vol. 88, no. 3, pp. 977–985, 1991.

[30] A. Singer, "Maximum entropy formulation of the kirkwood superposition approximation," *The Journal of Chemical Physics*, vol. 121, no. 8, p. 3657, 2004.

[31] P. Attard, O. G. Jepps, and S. Marcelja, "Information content of signals using correlation function expansions of the entropy," *Physical Review E*, vol. 56, p. 4052, 1997.

[32] H. Matsuda, "Physical nature of higher-order mutual information: Intrinsic correlations and frustration," *Physical Review E*, vol. 62, p. 3096, 2000.

[33] A. Baranyai and D. J. Evans, "Direct entropy calculation from computer simulation of liquids," *Physical Review A*, vol. 40, no. 7, pp. 3817–3822, 1989.

[34] D. C. Wallace, "On the role of density fluctuations in the entropy of a fluid," *The Journal of Chemical Physics*, vol. 87, p. 2282, 1987.

[35] R. Mountain and H. Raveché, "Entropy and Molecular Correlation Functions in Open Systems. II Two-and Three-Body Correlations," *The Journal of Chemical Physics*, vol. 55, p. 2250, 1971.

[36] W. McGill, "Multivariate information transmission," *IEEE Transactions on Information Theory*, vol. 4, no. 4, pp. 93–111, 1954.

[37] R. M. Fano, *Transmission of Information: A Statistical Theory of Communication*. The MIT Press, 1961.

[38] T. S. Han, "Multiple mutual informations and multiple interactions in frequency data," *Information and Control*, vol. 46, no. 1, pp. 26–45, 1980.

[39] B. J. Killian, J. Y. Kravitz, S. Somani, P. Dasgupta, Y. Pang, and M. K. Gilson, "Configurational entropy in protein-peptide binding: computational study of tsg101 ubiquitin e2 variant domain with an HIV-derived PTAP nonapeptide," *Journal of Molecular Biology*, vol. 389, p. 315, 2009.

[40] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[41] C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[42] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, p. 79, 1951.

[43] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 1991.

[44] T. Tsujishita, "On triple mutual information.," *Advances in Applied Mathematics*, vol. 16, p. 269, 1995.

[45] A. Bell, "The co-information lattice," *Tech. Rep. RNI-TR-02-1, Redwood Neurosci. Inst.*, 2003.

[46] S. Watanabe, "Information Theoretical Analysis of Multivariate Correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, p. 66, 1960.

[47] C. Chang and M. K. Gilson, "Free energy, entropy, and induced fit in host-guest recognition: calculations with the second-generation mining minima algorithm," *Journal of the American Chemical Society*, vol. 126, p. 13156, 2004.

[48] B. J. Killian, M. K. Gilson, and J. Y. Kravitz, "Extraction of configurational entropy from molecular simulations via an expansion approximation.," *The Journal of Chemical Physics*, vol. 127, 2007.

[49] S. Somani, B. J. Killian, and M. K. Gilson, "Sampling conformations in high dimensions using low-dimensional distribution functions," *The Journal of Chemical Physics*, vol. 130, p. 134102, 2009.

[50] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 1st ed. 2006. corr. 2nd printing ed., 2007.

[51] The MathWorks Inc., "MATLAB R2009b ver 7.5," 2009.

[52] K. S. Pitzer, "Energy levels and thermodynamic functions for molecules with internal rotation: II. unsymmetrical tops attached to a rigid frame," *The Journal of Chemical Physics*, vol. 14, no. 4, p. 239, 1946.

[53] C. Chang, M. J. Potter, and M. K. Gilson, "Calculation of molecular configuration integrals," *The Journal of Physical Chemistry B*, vol. 107, no. 4, pp. 1048–1055, 2003.

[54] R. Abagyan, M. Totrov, and D. Kuznetsov, "ICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation," *Journal of computational chemistry*, vol. 15, no. 5, pp. 488–506, 1994.

[55] S. K. Chang and A. D. Hamilton, "Molecular recognition of biologically interesting substrates: synthesis of an artificial receptor for barbiturates employing six hydrogen bonds," *Journal of the American Chemical Society*, vol. 110, no. 4, pp. 1318–1319, 1988.

[56] S. Goswami and R. Mukherjee, "Molecular recognition: a simple dinaphthyridine receptor for urea," *Tetrahedron Letters*, vol. 38, no. 9, pp. 1619–1622, 1997.

[57] A. D. MacKerell, Jr, D. Bashford, Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[58] J. Mazur and R. L. Jernigan, "Distance-dependent dielectric constants and their application to double-helical DNA," *Biopolymers*, vol. 31, no. 13, p. 1615, 1991.

[59] J. S. Liu, *Monte Carlo strategies in scientific computing*. Springer Verlag, 2008.

[60] F. M. Ytreberg and D. M. Zuckerman, "Peptide conformational equilibria computed via a single-stage shifting protocol," *The Journal of Physical Chemistry. B*, vol. 109, no. 18, pp. 9096–9103, 2005.

[61] M. R. Shirts and V. S. Pande, "Solvation free energies of amino acid side chain analogs for common molecular mechanics water models," *The Journal of Chemical Physics*, vol. 122, no. 13, p. 134508, 2005.

[62] H. Woo and B. Roux, "Calculation of absolute protein ligand binding free energy from computer simulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 19, pp. 6825–6830, 2005.

[63] T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G. J. Rylance, R. L. Johnston, and D. J. Wales, "How many dimensions are required to approximate the potential energy landscape of a model protein?," *Journal of Chemical Physics*, vol. 122, p. 84714, 2005.

[64] W. G. Hoover, "Thermodynamic properties of the fluid and solid phases for inverse power potentials," *The Journal of Chemical Physics*, vol. 55, no. 3, p. 1128, 1971.

[65] D. Frenkel and A. J. C. Ladd, "New Monte Carlo method to compute the free energy of arbitrary solids. application to the fcc and hcp phases of hard spheres," *The Journal of Chemical Physics*, vol. 81, no. 7, p. 3188, 1984.

[66] W. G. Hoover, "Use of computer experiments to locate the melting transition and calculate the entropy in the solid phase," *The Journal of Chemical Physics*, vol. 47, no. 12, p. 4873, 1967.

[67] L. M. Amon and W. P. Reinhardt, "Development of reference states for use in absolute free energy calculations of atomic clusters with application to 55-atom Lennard-Jones clusters in the solid and liquid states," *The Journal of Chemical Physics*, vol. 113, no. 9, p. 3573, 2000.

[68] J. P. Stoessel and P. Nowak, "Absolute free energies in biomolecular systems," *Macromolecules*, vol. 23, no. 7, pp. 1961–1965, 1990.

[69] F. M. Ytreberg and D. M. Zuckerman, "Simple estimation of absolute free energies for biomolecules," *The Journal of Chemical Physics*, vol. 124, p. 104105, 2006.

[70] M. D. Tyka, A. R. Clarke, and R. B. Sessions, "An efficient, path-independent method for free-energy calculations," *The Journal of Physical Chemistry B*, vol. 110, no. 34, p. 17212, 2006.

[71] M. Cecchini, S. V. Krivov, M. Spichty, and M. Karplus, "Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers," *The Journal of Physical Chemistry B*, vol. 113, no. 29, p. 9728, 2009.

[72] X. Zhang, A. B. Mamonov, and D. M. Zuckerman, "Absolute free energies estimated by combining precalculated molecular fragment libraries," *Journal of Computational Chemistry*, vol. 30, no. 11, pp. 1680–1691, 2009.

[73] S. Somani and M. K. Gilson, "Accelerated convergence of molecular free energy via superposition approximation-based reference states," *The Journal of Chemical Physics*, To be published.

[74] D. R. Herschbach, H. S. Johnston, and D. Rapp, "Molecular partition functions in terms of local properties," *The Journal of Chemical Physics*, vol. 31, no. 6, p. 1652, 1959.

[75] N. Go and H. A. Scheraga, "On the use of classical statistical mechanics in the treatment of polymer chain conformation," *Macromolecules*, vol. 9, no. 4, pp. 535–542, 1976.

[76] R. W. Zwanzig, "High-Temperature equation of state by a perturbation method. i. nonpolar gases," *The Journal of Chemical Physics*, vol. 22, p. 1420, 1954.

[77] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations.," *Journal of Molecular Graphics and Modelling*, p. 247, 2006.

[78] E. Sorin and V. S. Pande, "Exploring the helix-coil transition via all-atom equilibrium ensemble simulations," *Biophysical Journal*, vol. 88, no. 4, pp. 2472–2493, 2005.

[79] A. DePaul, E. Thompson, S. Patel, K. Haldeman, and E. Sorin, "Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics," *Nucleic Acids Research*, vol. 38, no. 14, pp. 4856–67, 2010.

[80] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for highly efficient, Load-Balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[81] A. M. Zoubir and B. Boashash, "Bootstrap methods and applications," *Signal Processing Magazine, IEEE*, vol. 15, no. 1, pp. 56–76, 1998.

[82] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*. Cambridge University Press, 1st ed., 1997.

[83] C. Jarzynski, "Rare events and the convergence of exponentially averaged work values," *Physical Review E*, vol. 73, no. 4, p. 46105, 2006.

[84] P. S. Shenkin, F. P. Hollinger, and W. C. Still, "The GB/SA continuum model for solvation. a fast analytical method for the calculation of approximate born radii," *Journal of Physical Chemistry A*, vol. 101, pp. 3005–3014, 1997.

[85] D. Sitkoff, K. A. Sharp, and B. Honig, "Accurate calculation of hydration free energies using macroscopic solvent models," *The Journal of Physical Chemistry*, vol. 98, pp. 1978–1988, 1994.

[86] J. Srinivasan, T. E. Cheatham III, P. Cieplak, P. A. Kollman, and D. A. Case, "Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices," *Journal of the American Chemical Society*, vol. 120, pp. 9401–9409, 1998.

[87] C. McClendon, G. Friedland, D. Mobley, H. Amirkhani, and M. Jacobson, "Quantifying correlations between allosteric sites in thermodynamic ensembles," *Journal of Chemical Theory and Computation*, vol. 5, pp. 2486–2502, 2009.

[88] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, pp. 141–151, 1999.

[89] A. Mitsutake, Y. Sugita, and Y. Okamoto, "Generalized-ensemble algorithms for molecular simulations of biopolymers," *Biopolymers*, vol. 60, pp. 96–123, 2001.

[90] J. G. Kim, Y. Fukunishi, and H. Nakamura, "Multicanonical molecular dynamics algorithm employing an adaptive force-biased iteration scheme," *Physical Review E*, vol. 70, no. 5, p. 57103, 2004.

[91] U. H. E. Hansmann and Y. Okamoto, "New Monte Carlo algorithms for protein folding," *Current Opinion in Structural Biology*, vol. 9, no. 2, p. 177, 1999.

[92] G. R. Bowman, D. L. Ensign, and V. S. Pande, "Enhanced modeling via network theory: Adaptive sampling of markov state models," *Journal of Chemical Theory and Computation*, vol. 6, no. 3, p. 787, 2010.

[93] M. Christen and W. F. van Gunsteren, "On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review," *Journal of Computational Chemistry*, vol. 29, no. 2, p. 157, 2008.

[94] A. Mitsutake, Y. Mori, and Y. Okamoto, "Multi-dimensional multicanonical algorithm, simulated tempering, replica-exchange method, and all that," *Physics Procedia*, vol. 4, p. 89, 2010.

[95] D. M. Zuckerman, "Equilibrium sampling in biomolecular simulation," *Annual Review of Biophysics*, vol. 40, no. 1, 2011.

[96] M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A. L. Beberg, D. L. Ensign, C. M. Bruns, and V. S. Pande, "Accelerating molecular dynamic simulation on graphics processing units," *Journal of Computational Chemistry*, vol. 30, no. 6, p. 864, 2009.

[97] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang, "Anton, a special-purpose machine for molecular dynamics simulation," *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2, pp. 1–12, 2007.

[98] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9 (1- 39)," *Journal of the American Chemical Society*, vol. 132, no. 5, p. 1526, 2010.

[99] C. Oostenbrink and W. F. van Gunsteren, "Free energies of binding of polychlorinated biphenyls to the estrogen receptor from a single simulation," *Proteins*, vol. 54, p. 237, 2004.

[100] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, "Dihedral angle principal component analysis of molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 126, p. 244111, 2007.

[101] D. J. Tobias and C. L. Brooks III, "Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results," *The Journal of Physical Chemistry*, vol. 96, no. 9, p. 3864, 1992.

[102] Y. Deng and B. Roux, "Computations of standard binding free energies with molecular dynamics simulations," *The Journal of Physical Chemistry B*, vol. 113, no. 8, p. 2234, 2009.

[103] C. Chipot and A. Pohorille, *Free energy calculations: Theory and applications in chemistry and biology*. Springer Berlin, 2007.

[104] H. X. Zhou and M. K. Gilson, "Theory of free energy and entropy in noncovalent binding," *Chemical Reviews*, vol. 109, no. 9, p. 4092, 2009.

[105] C. D. Christ, A. E. Mark, and W. F. van Gunsteren, "Basic ingredients of free energy calculations: A review," *Journal of Computational Chemistry*, vol. 31, no. 8, p. 1569, 2010.

[106] G. Schneider, "Virtual screening: an endless staircase?," *Nature Reviews Drug Discovery*, vol. 9, no. 4, p. 273, 2010.

[107] A. E. Roitberg, A. Okur, and C. Simmerling, "Coupling of replica exchange simulations to a non-Boltzmann structure reservoir," *The Journal of Physical Chemistry B*, vol. 111, no. 10, p. 2415, 2007.

[108] D. D. Claeys, T. Verstraelen, E. Pauwels, C. V. Stevens, M. Waroquier, and V. V. Speybroeck, "Conformational sampling of macrocyclic alkenes using a kennard-stone-based algorithm," *The Journal of Physical Chemistry A*, vol. 114, no. 25, p. 6879, 2010.

[109] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska, "Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data," *BMC bioinformatics*, vol. 5, no. 1, p. 118, 2004.

[110] A. Deshpande, M. Garofalakis, and R. Rastogi, "Independence is good: dependency-based histogram synopses for high-dimensional data," *SIGMOD Rec.*, vol. 30, pp. 199–210, 2001.

[111] J. P. Ryckaert and A. Bellemans, "Molecular dynamics of liquid alkanes," *Faraday Discussions of the Chemical Society*, vol. 66, p. 95, 1978.