DESY-04-096

DEUTSCHES ELEKTRONEN-SYNCHROTRON
in der HELMHOLTZ-GEMEINSCHAFT

DESY 04-096
ROM2F/2004/17
August 2004

Designing Generalized Statistical Ensembles for Numerical Simulations of Biopolymers

G. La Penna

Consiglio Nazionale delle Ricerche,
Istituto per lo Studio delle Macromolecole ISMac, Section of Genova, Italy
and
Magnetic Resonance Center, University of Firenze, Italy

S. Morante

Dipartimento di Fisica, Università di Roma "Tor Vergata", INFM, Unità di Roma 2, Italy

A. Perico

Consiglio Nazionale delle Ricerche,
Istituto per lo Studio delle Macromolecole ISMac, Section of Genova, Italy

G. C. Rossi

Dipartimento di Fisica, Università di Roma "Tor Vergata", INFN, Sezione di Roma 2, Italy
and
Deutsches Elektronen-Synchrotron DESY,
John von Neumann-Institut für Computing NIC, Zeuthen

ISSN 0418-9833

PLATANENALLEE 6 - 15738 ZEUTHEN

ISSN 0418-9833

ROM2F/2004/17
DESY 04-096
August 2004

Designing generalized statistical ensembles for numerical simulations of biopolymers

Giovanni La Penna *

Consiglio Nazionale delle Ricerche,

Istituto per lo Studio delle Macromolecole (ISMac), Section of Genova,

Via De Marini 6, 16149 Genova, Italy

and Magnetic Resonance Center, University of Firenze, Firenze, Italy

Silvia Morante †

Dipartimento di Fisica, Università di Roma Tor Vergata

INFM, Unità di Roma 2

Via della Ricerca Scientifica, 00133 Roma, Italy

Angelo Perico ‡

Consiglio Nazionale delle Ricerche,

Istituto per lo Studio delle Macromolecole (ISMac), Section of Genova,

Via De Marini 6, 16149 Genova, Italy

Gian Carlo Rossi §

Dipartimento di Fisica, Università di Roma Tor Vergata

INFN, Sezione di Roma 2

Via della Ricerca Scientifica, 00133 Roma, Italy

John von Neumann Institute for Computing, DESY Zeuthen, Germany

(Dated: 10th June 2004)

* Author for correspondence; E-mail: lapenna@ge.ismac.cnr.it
† E-mail: morante@roma2.infn.it
‡ E-mail: perico@ge.ismac.cnr.it
§ E-mail: rossig@roma2.infn.it

Abstract

Conformational properties of polymers, like average dihedral angles or molecular α-helicity, display a rather weak dependence on the detailed arrangement of the elementary constituents (atoms). We propose a computer simulation method to explore the polymer phase-space using a variant of the standard *multi-canonical* method, in which the density of states associated to suitably chosen configurational variables is considered in place of the standard energy density of states. This configurational density of states is used in the Metropolis acceptance/rejection test when configurations are generated with the help of a hybrid Monte Carlo algorithm. The resulting configurational probability distribution is then modulated by exponential factors derived from the general principle of the maximal constrained entropy by requiring that certain average configurational quantities take preassigned (possibly temperature dependent) values. Thermal averages of other configurational quantities can be computed by using the probability distributions obtained in this way. Moments of the energy distribution require an extra canonical sampling of the system phase-space at the desired temperature, in order to locally thermalize the configurational degrees of freedom.

As an application of these ideas we present the study of the structural properties of two simple models: a bead-and-spring model of polyethylene with independent hindered torsions and an all-atom model of alanine and glycine oligomers with 12 aminoacids in vacuum.

Keywords: Polymers; Statistical Mechanics; Information theory; Computer simulations.

I. INTRODUCTION

The study of the Statistical Mechanics of polymers, consisting of monomers of specific nature, is becoming more and more important in chemical and biological technologies. Polymers like proteins,[1] nucleic acids,[2] polysaccharides[3] and synthetic materials,[4] display features that strongly depend on their detailed physico-chemical properties like, for instance, the degree of flexibility of specific chemical linkages, the charge density on the monomer atoms and the structure of the hydrogen bond network between monomers either close or far away in the sequence.

The prediction of the structure, or family of structures, characteristic of a sequence of monomers in well specified experimental conditions is, nowadays, one of the most investigated structural problems posed in this field.[5,6]

The enormous complications associated, even within classical physics, with the atomistic description of the specific interaction among the elementary components of the polymer can only be handled by numerical simulations. Clever algorithms have been devised to efficiently explore the configuration space available to the system and different types of *ensembles* have been invented and numerically implemented, starting from molecular dynamics (MD) and Monte Carlo (MC) methods. As is well known, MD and MC simulations explore the *micro-canonical* ensemble and the *canonical* ensemble of the system, respectively. Other kinds of ensembles, which may be collectively indicated with the name of *generalized ensembles*, [7–13] can also be introduced and employed for the study of the thermodynamic properties at equilibrium. In principle, under standard ergodicity assumptions, all these ensembles can yield equivalent physical information.

Despite the fact that many powerful mathematical tools are available today within the polymer theory, reliable calculations of only a limited set of statistical averages or configurational distributions can be obtained and studied as functions of a few parameters encoding the many complications of the polymer structure and inter-atomic interactions.[14–16] Therefore, the idea of combining computer simulations with the knowledge of the statistical averages of a few configurational quantities, that may be available either from theory or from experiments, emerges.

The most difficult problem that is encountered in numerical simulations is related to the limited and strongly biased sampling of the configurational space occurring when the temperature is lower than the critical temperature of the model. By "critical temperature" we generically mean the temperature above which the system is in the disordered phase. Below the critical temperature the system remains trapped in local minima, within energy barriers that are rarely overtaken by

thermal fluctuations.

Most of the widely used computational methods designed to avoid this problem, like simulated annealing[17,18], genetic algorithms,[19] and stochastic tunnelling[20] do not always permit the computation of statistical averages in well defined ensembles (*i.e.* those that are the statistical representatives of the desired experimental conditions). On the other hand, the use of generalized ensembles within MD and MC simulation strategies results in a powerful approach to predict the statistical properties of atomistic models of complex molecules in a wide range of temperatures and order parameter values. Algorithms designed to access ensembles of the *multi-canonical* type, *i.e.* the kind of generalized ensembles that exploit the information on the (potential) energy density of states, are well assessed and made rather effective if used in conjunction with the replica-exchange method. In many interesting instances it is possible to computationally monitor order parameters of geometrically constrained molecular models of polymers in a fairly large temperature range. Problems, however, arise when all the degrees of freedom of the models, which include high-frequency vibrations, are taken into account,[21] as it is necessary to do in order to treat condensed phases and explicit solvents: large variations of the potential energy are observed associated with these modes even for tiny configurational changes. Indeed large potential energy changes associated to stiff terms in the Hamiltonian (like bond stretching and bending potentials) or inter-atomic contacts cause very low acceptance in the exchange of temperatures between replicas and lack of convergence of the multi-canonical weight computation.[21–23]

A. Aim of the paper

The aim of this work is to design generalized ensembles suited for computer simulations, not based on monitoring the system potential energy, but some appropriately chosen configurational variable about which theoretical and/or experimental information are available in terms of averages, in some range of the experimentally relevant physical parameters, like temperature or density.

In a recent paper,[24] it has been shown that information on average molecular structural quantities can be introduced in configurational random walks by generalizing the standard maximal entropy method to find the maximum of the so-called "cross entropy"[25] functional under the constraint that the statistical average of certain, appropriately selected, configurational quantities, take their (experimentally or theoretically) known values. The well-known result of the constrained maximization procedure is that the best estimate of the thermal configurational probability is

yielded by a "modulation" of the original random walk probability distribution, in terms of a factor which turns out to have the usual exponential functional form.

As an alternative approach to the *multi-overlap* simulation[26] and various kinds of biasing potentials,[27–30], in this work we will further develop and straighten the approach of ref. 24 by imposing constraints coming from assigning well defined values to selected configurational quantities on an "improved" configurational probability distribution. The latter is recursively obtained, starting from some initial random walk, by successively generating configurations with weight provided by the density of states associated with the selected configurational quantities. At each step of the recursion the current distribution is used to get an estimate of this density of states and the logarithm of its inverse is in turn used to generate, by a standard Metropolis acceptance/rejection algorithm, a less biased probability distribution. The procedure is repeated until stability is obtained. In this way, as the recursion goes on, histograms of the configurational quantities used to estimate the current density of states become flatter and flatter because the configurations rarely reached at the beginning are successively sampled with higher and higher probability, owing precisely to the low value of their weight in the current density of states.

The principle of constrained maximal entropy is then applied to the configuration probability distribution obtained as the last refinement of the above loop. The least-biased estimate of the actual probability distribution of the system, which is "nearer" to the latter and fulfills the constraints relatively to the expectation values of the selected configurational quantities, is obtained by maximizing the cross-entropy functional[25] under the specified constraints.

B. The problem of temperature

A non-trivial issue within the approach we have described is the introduction of temperature. Actually, getting the (correct) temperature dependence of the whole set of average configurational quantities of a macromolecule is a formidable problem which goes beyond our present capabilities. The much more modest aim of this work is to offer a procedure suitable for computing the configurational probability distribution of a model system at a given temperature, when, as it is usually the case, only a few average quantities, at a small set of temperatures, are known either from theory or experiments.

The problem of introducing temperature in the computational scheme we advocate is solved by injecting the necessary information about it in two complementary ways. First of all, in the con-

strained maximal entropy method we can use as inputs for the chosen configurational quantities, values taken at the different, possibly available, temperatures.[24] This indirectly leads to a T dependence of the resulting configurational probability, which, however, is not sufficient to overcome the problem of the lack of "canonicity" of the ensuing energy distribution. The reason is that, due to complexity of the energy landscape, it easily happens that states of the system which only slightly differ in their atomic configurations have instead largely different values of the potential energy. It is then natural, as a second step, to try to enforce the correct Boltzmann probability distribution by further performing standard hybrid Monte Carlo (HMC) simulations at the desired temperature starting with the available set of configurations. According to the details of the procedure, one may need to reweight the collected configurations, when computing thermal averages, to get rid of possible biases due to the particular acceptance/rejection criterion adopted in the HMC step.

C. Applications and plan of the paper

We will discuss the application of the statistical method we have just presented to two simple macromolecule models. The first one is a model of polyethylene chain, described as a bead-and-spring chain with independent hindered torsions and no excluded volume. The cosine of the dihedral angle between the two planes formed by three consecutive bonds, averaged along the chain, is taken as the configurational variable with respect to which the relevant density of states is constructed. The temperature dependence of this quantity is known exactly from theory.[15]

The second system we will study is an all-atom model of polyalanine and polyglycine chains with 12 monomers in vacuum: the same strategy as in the previous case is used here, except that information about the tridimensional structure of the systems is introduced through the knowledge of the average molecular α-helicity. From the comparison between the behaviour of potential energy and entropy of the two macromolecules, as functions of the value of their average α-helicity, one can extract a quantitative information on the relative propensity of the two systems to form α-helical structures.

The plan of the paper is the following. In section II we will explain the theoretical and the computational tools we employ in this work and we will discuss how temperature can be introduced in the approach we propose. In section III we will describe the application of our method to the study of a polyethylene chain, while in section IV the same computational strategy will be applied to all-atom models of polyalanine and polyglycine molecules in vacuum. Conclusions and some

considerations regarding future lines of development can be found in section V. Certain technical details are given in an Appendix.

II. THEORETICAL CONSIDERATIONS

A useful notion of "cross-entropy" can be introduced in situations where partial information about the statistical properties of a system is available. As suggested in ref. 25, we may formalize the discussion in the following way. Let w_i be the weight of the microscopic state i, corresponding to a given macrostate of the system. To fix the ideas we can imagine that the macrostate is specified in terms of the value of its total energy, E, the total number of elementary components, N, and the volume V, where the components are confined. What we have in mind here is the construction of the Statistical Mechanics of the system in terms of the notion of micro-canonical ensemble. The probability, $\mathcal{P}_i$, to find the system in the microstate i is

$$\mathcal{P}_i = \frac{w_i}{W}, \quad W = \sum_i w_i, \tag{1}$$

where W is the number of microstates of the system compatible with the assigned values of E, N and V. The total entropy of the system is

$$S = k_B \log W, \tag{2}$$

k_B being the Boltzmann constant. Let us now suppose that a certain amount of information is available at some "mesoscopic" level. Labeling mesostates by the index γ, we have for their statistical weight the formula

$$w_\gamma = \sum_{i\in\gamma} w_i. \tag{3}$$

In analogy with Eq. (2) we will call $S_\gamma = k_B \log w_\gamma$ the entropy of the mesostate γ. Using this formula, the probability of a mesostate becomes

$$\mathcal{P}_\gamma = \frac{w_\gamma}{\mathcal{Z}} = \frac{1}{\mathcal{Z}} \exp(S_\gamma/k_B), \quad \mathcal{Z} = \sum_\gamma \exp(S_\gamma/k_B), \tag{4}$$

where $\mathcal{Z}$ is the partition function of the system. From the definition of S_γ and Eq. (3) one immediately concludes that $\mathcal{Z} = W$.

The key observation at this point is that the total entropy of the system (Eq. (2)) can be written in the suggestive form

$$S = k_B \log \mathcal{Z} = \sum_\gamma \mathcal{P}_\gamma [S_\gamma - k_B \log \mathcal{P}_\gamma]. \tag{5}$$

Despite the fact that Eq. (5) is an identity (to show this use the previous equations for $\mathcal{P}_\gamma$ and S_γ), it can be exploited to set up a useful variational principle. We can imagine, in fact, to be in the following situation. We have some approximate information on the weights, $\widetilde{w}_\gamma$, of appropriately defined mesostates of the system and preassigned values for a number of thermal averages, $\langle \vec{A} \rangle \equiv \langle A^{(k)} \rangle, k = 1, \ldots, M$. From the knowledge of the quantities $\widetilde{w}_\gamma$ we can construct the probability

$$\widetilde{\mathcal{P}}_\gamma = \frac{\widetilde{w}_\gamma}{\sum_{\gamma'} \widetilde{w}_{\gamma'}} \tag{6}$$

of finding the system in the mesostate γ. The (approximate) information about its entropy is then expressed by the quantity $\widetilde{S}_\gamma = k_B \log \widetilde{w}_\gamma$.

The problem of finding the least-biased expression of the probabilities $\mathcal{P}_\gamma$ that is nearer to the $\widetilde{\mathcal{P}}_\gamma$ and satisfies the conditions

$$\langle A^{(\ell)} \rangle = \sum_\gamma \mathcal{P}_\gamma A_\gamma^{(\ell)} = \alpha^{(\ell)}, \quad \ell = 1, \ldots, M \tag{7}$$

is solved by determining the maximum of what, following ref. 25, we will call the "cross-entropy" functional

$$\begin{aligned} S_c[\mathcal{P}, \widetilde{\mathcal{P}}] &= \sum_\gamma \mathcal{P}_\gamma [\widetilde{S}_\gamma - k_B \log \mathcal{P}_\gamma] = k_B \sum_\gamma \mathcal{P}_\gamma [\log \widetilde{w}_\gamma - \log \mathcal{P}_\gamma] = \\ &= -k_B \sum_\gamma \mathcal{P}_\gamma \log \frac{\mathcal{P}_\gamma}{\widetilde{\mathcal{P}}_\gamma} - k_B \log \widetilde{\mathcal{Z}} \end{aligned} \tag{8}$$

under the constraints imposed by the Eqs. (7). The above is a well posed variational problem, as it follows by noticing that the functional $S_c[\mathcal{P}, \widetilde{\mathcal{P}}] + k_B \log \widetilde{\mathcal{Z}}$ is non-positive (provided $\mathcal{P}_\gamma$ and $\widetilde{\mathcal{P}}_\gamma$ are normalized to unit, *i.e.* satisfy $\sum_\gamma \mathcal{P}_\gamma = \sum_\gamma \widetilde{\mathcal{P}}_\gamma = 1$)[43] and vanishes when $\mathcal{P}_\gamma = \widetilde{\mathcal{P}}_\gamma$. Introducing the constraints (7) through the M Lagrange multipliers, $\vec{\lambda} \equiv (\lambda^{(k)}, k = 1, \ldots, M)$, the solution for $\mathcal{P}_\gamma$ is easily found to be

$$\mathcal{P}_\gamma = \frac{\widetilde{\mathcal{P}}_\gamma}{\mathcal{Z}(\vec{\lambda})} \exp\Big[-\sum_{k=1}^{M} \lambda^{(k)} A_\gamma^{(k)} \Big], \tag{9}$$

$$\mathcal{Z}(\vec{\lambda}) = \sum_\gamma \widetilde{\mathcal{P}}_\gamma \exp\Big[-\sum_{k=1}^{M} \lambda^{(k)} A_\gamma^{(k)} \Big], \tag{10}$$

where the parameters $\vec{\lambda}$ are consistently determined by solving the equations

$$\sum_\gamma \mathcal{P}_\gamma A_\gamma^{(\ell)} = \frac{1}{\mathcal{Z}(\vec{\lambda})} \sum_\gamma \widetilde{\mathcal{P}}_\gamma \exp\Big[-\sum_{k=1}^{M} \lambda^{(k)} A_\gamma^{(k)} \Big] A_\gamma^{(\ell)} = \alpha^{(\ell)}, \ \ell = 1, \ldots, M. \tag{11}$$

Introducing back the solution for $\widetilde{\mathcal{P}}_\gamma$ into Eq. (8), one gets for the cross-entropy at the maximum

$$S_c^{(m)}[\vec{\alpha}] = k_B\Big[\log\frac{\mathcal{Z}(\vec{\lambda})}{\widetilde{\mathcal{Z}}} + \sum_{k=1}^{M}\lambda^{(k)}\alpha^{(k)}\Big]\,. \tag{12}$$

The procedure we have followed (sometimes called for short the MEC procedure in the rest of the paper) parallels exactly the usual constrained maximal entropy method. Similarly to what is done in more standard cases, we assume that any increase of information one might get about a system reflects itself in a decrease of its informational entropy, which in our case we identify with the cross-entropy functional defined in Eq. (8). Thus within the set of all possible configurational probability distributions satisfying the constraints (7), the least-biased one is to be considered the one which maximizes our "ignorance", hence S_c.

A. Application to numerical simulations

To make precise the following discussion and make contact with the previous rather general notations, we must first of all specify what is to be meant in this context by the index γ. To this end we observe that in actual simulations any configurational probability distribution (which is a function of the system coordinates) will be necessarily described by a histogram with a finite number of entries. Thus γ will have to be identified with the index that labels these entries. In the following pages, however, for short we will keep using a continuum-like notation where we replace γ by the set of coordinates, $\{\mathbf{r}\}$, that specify the state of the system, and the sum over γ by the integral over all the system coordinate space, *i.e.*

$$\gamma \to \{\mathbf{r}\}\,, \qquad \sum_\gamma \to \int\prod d\mathbf{r}\,. \tag{13}$$

We now proceed to construct the appropriate probability distribution, $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$ (the analog of Eq. (6)), that will allow us to go over to the improved one, $\mathcal{P}[\{\mathbf{r}\}]$, once equations like (11) are imposed within the constrained maximal entropy scheme. The construction of $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$ will be done recursively, starting from an initial random walk distribution, $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$, obtained as discussed in ref. 24. The probability associated to random walks was called there "meta" configurational probability distribution, a name that we will occasionally use also in this paper. For completeness we recall in the Appendix the construction of $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$.

B. A recursive construction of $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$

For reasons explained in the Introduction we want to construct an improved configurational probability distribution based on the knowledge of the "A-density" of states, $g_A(a)$, associated to a given configurational quantity, $A[\{\mathbf{r}\}]$. $g_A(a)$ is defined by the equation

$$g_A(a) = \int \prod d\mathbf{r}\, \delta\Big(a - A[\{\mathbf{r}\}]\Big)\,, \tag{14}$$

which naturally generalizes the usual formula for the density of states

$$n(u) = \int \prod d\mathbf{r}\, \delta\Big(u - U[\{\mathbf{r}\}]\Big)\,. \tag{15}$$

Eq. (14) can be further extended to define the cross $\{\vec{A}\}$-density of states, $\{\vec{A}\} \equiv (A^{(k)}, k = 1, \ldots, M)$, through the equation

$$g_{\{\vec{A}\}}(\vec{a}) = \int \prod d\mathbf{r} \prod_{k=1}^{M} \delta\Big(a^{(k)} - A^{(k)}[\{\mathbf{r}\}]\Big)\,. \tag{16}$$

Of course only an approximate knowledge of the density of states is possible, as the number of generated configurations is never infinite nor unbiased. This is so because the available CPU-time is limited and there are always constraints and biases in the way the configurational space is explored. For instance, in the procedure described in the Appendix there is a maximum allowed temperature in the procedure employed to extract the particle velocities, though no upper bound to their absolute magnitude. Moreover, different steady states are obtained if different (but always time-reversible) algorithms are used to numerically implement the MD steps or the shape of the suggesting probability is modified (though keeping it invariant under the sign inversion of the particle velocities and independent of the configuration).

In order to (at least partially) overcome these difficulties, one can envisage an iterative procedure to construct a better probability distribution, $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$, based on a systematically improved knowledge of some cross $\{\vec{A}\}$-density of states. For illustrative purposes let us discuss the case $M = 1$. The procedure yielding the required probability distribution is rather straightforward. From the random walk distribution, $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$, discussed in the Appendix one starts by constructing a zero-th order approximation of the A-density of states, $g_A^{(0)}(a)$, by simply discretizing Eq. (14). This is done by partitioning in an appropriate number of bins, say N_A, the variability range of the quantity A and counting the number of times any given bin value, $a_{(j)}, j = 1, \ldots, N_A$, is obtained while sweeping through all the configurations that make up $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$.

At this point a new sampling of the system configurational space is performed with some suggesting probability, P_{sug} and an acceptance/rejection probability, $P_{acc}^{(0)}$, given by the multi-canonical-like formula[9]

$$P_{acc}^{(0)}(a \to a') = \min\left[1, \frac{\exp[-Q_n^{(0)}(a')]}{\exp[-Q_n^{(0)}(a)]}\right], \tag{17}$$

where $Q_n^{(0)}(a)$ is an order n polynomial interpolation of $\log[g_A^{(0)}(a)]$. This procedure satisfies the detailed balance condition if P_{sug} is invariant under the exchange of its arguments, as explained in the Appendix. Furthermore it is so obtained to preferentially accept configurations with previously low probability, *i.e.* with previously low values of $g_A^{(0)}(a)$.

The configurations collected in this way give rise to a new meta configuration probability distribution, $\widetilde{\mathcal{P}}^{(1)}$. The set of newly generated configurations can now be used to compute the next approximation, $g_A^{(1)}$, to the A-density of states. The latter is then employed to again sample the system phase-space with an acceptance/rejection algorithm like the one in Eq. (17), but with $Q_n^{(0)}(a) = \log[g_A^{(0)}(a)]$ replaced by $Q_n^{(1)}(a) = \log[g_A^{(1)}(a)]$. The loop is iterated m times until no change in the $\vec{\lambda}$'s is appreciable while passing from the m-th to the $m+1$-th iteration. The last estimates of the meta distribution and A-density of states will be simply called $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$ and $\widetilde{g}_A(a)$, respectively. $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$ is finally used in the constrained maximal entropy method described at the beginning of this section (see Eq. (9)).

For the thermal average of the configurational quantity, $B[\{\mathbf{r}\}]$, one finally gets

$$\langle B\rangle_{\vec{\lambda}} = \frac{1}{\mathcal{Z}(\vec{\lambda})}\sum_{i=1}^{N_c}[\widetilde{g}_{\{\vec{A}\}}(\vec{a}_i)]^{-1}\exp\left[-\sum_{k=1}^{M}\lambda^{(k)}a_i^{(k)}\right]b_i\,, \tag{18}$$

$$\mathcal{Z}(\vec{\lambda}) = \sum_{i=1}^{N_c}[\widetilde{g}_{\{\vec{A}\}}(\vec{a}_i)]^{-1}\exp\left[-\sum_{k=1}^{M}\lambda^{(k)}a_i^{(k)}\right], \tag{19}$$

where N_c is the total number of collected configurations and we have set for short $b_i = B[\{\mathbf{r}_i\}]$ and $a_i^{(k)} = A^{(k)}[\{\mathbf{r}_i\}]$.

We conclude by noticing that we should have appended an index A also to $\widetilde{\mathcal{P}}$, because in any realistic situation this probability distribution would somehow depend on the particular set of configurational variables we had decided to consider in the construction. To lighten the notation we do not do it. In principle such a dependence would disappear after a fully exhaustive (*i.e.* ergodic) exploration of the system phase-space has been carried out.

C. Introducing temperature

So far temperature has not yet been brought on stage. Introducing the notion of temperature for a complex (fully flexible) system, like a polymer, is a delicate issue, because of the observation we already made that configurations only slightly different in their atomic spatial arrangement may have largely different potential energies. Consequently, as it turns out, it becomes more and more difficult to get the correct (Boltzmannian) energy distribution of the total available energy among the many degrees of freedom of the system as the temperature increases (despite the fact that at high temperature overcoming energy barriers may become easier).

This is the reason why, within the standard multi-canonical approach, introducing the temperature through the modulation of the energy density of states by the Boltzmann factor does not yield sufficiently satisfactory results, as soon as the number of beads (hence the number of degrees of freedom) of the polymer exceeds a certain value and/or the temperature is above the order-disorder phase transition.[21]

It is precisely to overcome this kind of problems that we are proposing in this paper to work in a generalized ensemble where configurations are in the first place generated according to the density of states associated to some set of configurational quantities, rather than to energy. In this way, however, it might seem that the fundamental statistico-mechanical relation between the energy of the system and its temperature is put at stake. In fact, for the purpose of studying, say, biopolymers, which after all are neither isolated systems, nor do they work at equilibrium, this state of affairs may not be a problem and can be dealt with along the lines described below.

Indeed, in the framework we are developing there is room for the introduction of a sensible notion of temperature. We do it in two separate steps. Temperature can be injected in the statistics if we know how the preassigned values of the thermal averages depend on T. In this case the conditions (7) take the form $\langle \vec{A} \rangle = \vec{\alpha}_T$. This knowledge will induce a T dependence ($\vec{\lambda} = \vec{\lambda}(\beta)$, $\beta = 1/k_B T$) on the values of the Lagrange multipliers that are obtained by solving Eqs. (11). Through Eqs. (18) and (19) this dependence will be passed to the thermal average of any other configurational quantity one wishes to compute.

The temperature dependence induced in this way is not enough, however, to produce the correct T-behaviour of quantities that require an accurate thermalization of all the degrees of freedom of the system, like *e.g.* the moments of the (potential) energy distribution. Thus a local, extra thermalization has to be carried out. This is done in the following way. Starting from each one of

the already recorded configurations (those making up the available meta probability distribution, $\widetilde{\mathcal{P}}[\{\mathbf{r}\}]$), one performs a number of hybrid MC steps with velocities extracted from a Maxwell-Boltzmann distribution at the desired temperature, T. At the end of the MD moves of each hybrid MC step configurations are subjected to a standard Metropolis test with acceptance/rejection probability

$$P_{acc}[\{\mathbf{r}_1,\mathbf{v}_1\} \to \{\mathbf{r}_2,\mathbf{v}_2\}] = \min\left[1, \frac{\exp(-\beta H[\{\mathbf{r}_2,\mathbf{v}_2\}])}{\exp[(-\beta H[\{\mathbf{r}_1,\mathbf{v}_1\}])}\right], \tag{20}$$

where $H[\{\mathbf{r},\mathbf{v}\}]$ is the total (kinetic plus potential) energy of the system. In this way configurations are smoothly thermalized at the desired temperature.

The set of the newly obtained hybrid MC trajectories provides the configuration samples that should be used in computing expectation values of potential energy functions, $f(U)$. For such quantities formulae exactly analogous to Eqs. (18)–(19) are valid, which we report here for the completeness

$$\langle f(U)\rangle_\beta = \frac{1}{\mathcal{Z}(\beta)} \sum_{i=1}^{N_{\rm HMC}} [\widetilde{g}_{\vec{A}}(\vec{a}_i)]^{-1} \exp\left[-\sum_{k=1}^{M} \lambda^{(k)}(\beta) a_i^{(k)}\right] f(U_i)\,, \tag{21}$$

$$\mathcal{Z}(\beta) = \sum_{i=1}^{N_{\rm HMC}} [\widetilde{g}_{\vec{A}}(\vec{a}_i)]^{-1} \exp\left[-\sum_{k=1}^{M} \lambda^{(k)}(\beta) a_i^{(k)}\right], \tag{22}$$

In Eqs. (21)–(22) $N_{\rm HMC}$ is the number of the collected hybrid MC configurations. The quantities $U_i = U[\{\mathbf{r}_i\}]$ and $a_i^{(k)} = A^{(k)}[\{\mathbf{r}_i\}]$ are the values of the potential energy and configurational variables, $A^{(k)}$, respectively, recorded along the hybrid MC trajectory. This approach will be called thermalized constrained maximal entropy, TMEC hereafter.

The following alternative thermalization procedure can also be envisaged. Exploiting the knowledge of the density of states $\widetilde{g}_A(a)$ a single generalized-ensemble hybrid Monte Carlo trajectory[31] can be constructed using the acceptance/rejection rule criterion

$$P_{acc}[\{\mathbf{r}_1,\mathbf{v}_1\} \to \{\mathbf{r}_2,\mathbf{v}_2\}] = \min\left[1, \frac{\exp(-\beta H[\{\mathbf{r}_2,\mathbf{v}_2\}] - \widetilde{Q}_n[a(\mathbf{r}_2)]}{\exp[(-\beta H[\{\mathbf{r}_1,\mathbf{v}_1\}] - \widetilde{Q}_n[a(\mathbf{r}_1)]}\right], \tag{23}$$

where $\widetilde{Q}_n(a)$ is the (already determined) order n polynomial interpolation of $\log \widetilde{g}_A(a)$ (see Eq. (17)). No use of the constrained maximal entropy strategy is employed here, so no exponential factors of the type appearing in Eqs. (21) and (22) will enter in the formulae for thermal averages. Thus compared to the TMEC approach, the construction of the generalized-ensemble hybrid MC (in the following GEHMC for short) trajectory does not require any *a priori* knowledge of the system, except for the A-density of states, $g_A(a)$, which is supposed to have been estimated either

as explained above or by some other possibly more sophisticated algorithm. It should, however, be said that the use of the Maxwell-Boltzmann velocity distribution at fixed inverse temperature β produces strongly correlated moves that tend to leave the system trapped in states where $g_A(a)$ is at a local minimum.

It is our experience that, although thermal averages of configurational variables do not change appreciably if the TMEC equilibration is omitted, energy dependent quantities significantly improve their T-behaviour, if the latter step is carried out. It must also be observed that consistency requires that the temperature appearing in the Metropolis test (Eq. (20)) be equal to the temperature at which the constraints (Eq. (11)) are imposed. To underline this obvious, but important fact we have explicitly indicated in all our formulae the β dependence of the parameters $\vec{\lambda}$.

III. SINGLE POLYETHYLENE CHAIN

A simple polyethylene model was studied in ref. 24, using a straightforward constrained maximal entropy type approach. Here, the same model is summarized and previous results are improved according to the random walk refinement and thermalization procedure described in the previous section.

The polyethylene chain we have considered consists of n units representing methylene fragments joined by bonds. Bond stretching and angle bending motions are described by harmonic springs, while the potential for dihedral angles is expanded as a Fourier cosine sum. The coefficients in the sum are those used for early simulations of united-atom models of alcanes.[32] All bond lengths, valence angles and dihedral angles along the chain will be considered equivalent, as the monomers making the polymer are all identical.

An exact expression for the average square end-to-end polymer distance $\langle h^2 \rangle$ of a polymer with n methylene units can be given.[15] Explicit formulae show that $\langle h^2 \rangle$ can be written in terms of three configurational average quantities: the average of the square bond length, $\langle l^2 \rangle$, the average of $\cos\theta$, where θ is the valence angle, and the average of $\cos\phi$, where ϕ is the dihedral angle between three consecutive C-C bonds. In ref. 24 the average end-to-end square distance was extracted from simulations in which configurational random walk probability distributions were modulated (according to the constrained maximal entropy method) exploiting only the information coming from the temperature dependence of the configurational average of the quantity $\langle \cos\phi \rangle$, which is exactly known from theory. The average square end-to-end distance values obtained in

this way have been compared to the known theoretical behaviour, finding a rather good agreement in the explored range of temperatures ($T = 100$ K to $T = 600$ K) and with n up to 40.

Despite this agreement, the information contained in the probability distribution gathered in the simulation proved not to be sufficient to allow for the construction of the correct end-to-end distance distribution function.

As a first step in the direction of constructing the correct configurational distribution function, one can start by simplifying the model, keeping bond lengths and valence angles fixed to their equilibrium values (0.1526 nm and 112.4 degrees, respectively). In this situation the use of generalized-ensemble Monte Carlo (GEMC for short) methods of the multi-canonical type looks adequate. The rigid geometry model was studied in ref. 24, employing the SMMP program.[33] Results on the configurational distribution densities of the fully flexible model obtained in the next section will be compared with those derived using the generalized-ensemble MC probability distribution of the rigid model. In the explored temperature range (100 K $< T <$ 600 K) the configurational distribution densities of the two models should not be very different.

A. Constructing the meta probability distribution

In this section we want to show that by making use of the meta configuration probability distribution, $\widetilde{\mathcal{P}}$, associated to some, suitably chosen, configurational variable and constructed as explained in Sect. II, one can get satisfactory distribution functions of configurational quantities, even for the fully flexible model of polyethylene, *i.e.* for the model where the force-field is taken to include all the degrees of freedom of the system, namely bond stretching, angle bending and dihedral deformation modes.

As recalled in the Appendix, the initial configurational random walk, $\widetilde{\mathcal{P}}^{(0)}$, is constructed by collecting 10^5 configurations of the system generated in MD runs each made of 10^3 iterations with a time-step of 1 fs. Every run is performed starting with velocities extracted from a Maxwell-Boltzmann distribution at a temperature randomly chosen in the interval from 0 K to 1000 K. Further details of this construction are reported in the Appendix and in ref. 24. It was shown in this work that, if the initial random walk, where all the generated configurations are accepted, is used in computing thermal averages, the configurations that are expected to give the most important contributions to the low temperature average of the end-to-end square distance are rarely encountered. This fact is signaled by large uncertainties in the computed values of constrained

maximal entropy λ parameters. This is how, in this setting, the difficulties of properly sampling the system configurational space at low temperatures emerge.

In the present paper we want to improve, in the physically motivated way we discussed before, the initial random walk. As a relevant configurational variable with respect to which we compute the associated density of states, we decided to take $A = \sum_i \cos\phi_i/N_d$, where i runs over the N_d equivalent dihedral angles of the molecule. This quantity was chosen because the cosine of the dihedral angle is the most important parameter characterizing the polymer stiffness and elongation.

As we said, the recursive construction illustrated in Sect. II B uses the $\widetilde{\mathcal{P}}^{(0)}$ distribution as the initial approximation. Already this zero-th order distribution is capable of adequately sampling configurations with $\langle\cos\phi\rangle$ in the range from -0.8 to 0. To go to the next iteration a fifth-order polynomial, $Q_A^{(0)}(\cos\phi)$, was fitted through the (thirteen) points making up $\log[g_A^{(0)}(\cos\phi)]$ in this range. This is done only to have a continuous function as probability distribution for proposing configurations in the next iteration. As usual in multi-canonical weight construction, beyond the limits of the sampled range the logarithm of the density of states is extended as a linear function in the relevant variable (here the average $\cos\phi$ along the chain) with a continuous derivative at the matching points. One proceeds in the iteration until convergence is obtained.

At each step the current meta probability distribution, $\widetilde{\mathcal{P}}^{(m)}$, is subjected to the exponential modulation as required in the constrained maximal entropy method (see Eq. (11)). In this step the theoretically known temperature dependence of $\langle\cos(\phi)\rangle$ is exploited.[15] As a result, the computed λ parameter will be a function of the temperature.

In Fig. 1 the λ parameters successively obtained, while progressively improving the meta probabilities, are displayed as a function of the inverse temperature, β. It is clearly seen that the gain in information sensitivity in going from the second iteration, $\widetilde{\mathcal{P}}^{(1)}$, to the third, $\widetilde{\mathcal{P}}^{(2)}$, is very small. This suggests that the firstly refined meta probability is already adequately sampling the relevant part of the conformational space of the system in the 100-600 K temperature range.

In Fig. 2 we compared the exact (*i.e.* theoretically computed) dependence of $\langle h^2\rangle$ on the temperature with results coming from different simulations and/or modeling strategies, namely generalized-ensemble MC (GEMC), constrained maximal entropy (MEC) and generalized-ensemble hybrid MC (GEHMC). Generalized-ensemble MC and constrained maximal entropy methods give comparably good results. This means that, as expected on the basis of the stretching and bending force constants used in the model, the full flexibility is not necessary in order to get agreement with theory. On the contrary, data for the fully flexible model, coming from

the fixed temperature generalized-ensemble hybrid MC trajectories depart significantly from the correct results already below 350 K (see comment below Eq. (23)).

In the constrained maximal entropy and generalized-ensemble MC cases the increase in the average chain extension visible below 200 K is well reproduced and originates from contributions of almost fully extended (all *trans*) configurations. In order to quantify this effect, in Fig. 3 the distribution, $P = P(h^2)$, of the square end-to-end distance h^2, is plotted for three temperatures, 100 K, 300 K and 600 K, in panels (A), (B) and (C), respectively. We compare them with the same kind of distributions obtained after the thermalization step described by Eq. (23). Differences are small, except at the lowest temperature, where the generalized-ensemble MC distribution somewhat deviates from the other two. This is due both to statistical fluctuations and the smaller flexibility of the model which was employed in that case. We recall that bond stretching and angle bending modes were frozen at their equilibrium value. The area under the rightmost peak of panel (A) in the generalized-ensemble MC distribution yields an estimate of the all-*trans* population. Next peak to the left should be related to the all-but-one-*trans* population. It can be seen from the distribution at $T = 100$ K that, due to full bond flexibility, in the constrained maximal entropy and thermalized constrained maximal entropy (TMEC) cases there is no clear-cut separation between the fully extended (all-*trans*) configurations (geometrically the square end-to-end distance of the all-*trans* configuration is 24.4 nm^2) and those with only one single *gauche* dihedral angle.

B. Towards a canonical probability distribution

As we have previously explained, the evaluation of averages of quantities related to the potential energy requires a further thermalization step. The expectation value of the potential energy and its 2nd momentum at various temperatures have been evaluated by using the TMEC thermalization method, summarized in Eqs. (20)–(22).

More precisely this is done in the following way. One tenth of the 10^5 configurations obtained after the iterative procedure bringing from the initial random walk to the final meta probability distribution, $\widetilde{\mathcal{P}}$, have been selected and taken as starting points for the hybrid MC thermalization. From each such configuration a total of 100 HMC steps were performed at every one of the temperatures used in the constrained maximal entropy modulation. In each HMC step the system was evolved with short MD runs (10 iterations). In total 10^6 configurations were collected at every single temperature and a sample of uniformly drawn 10^5 configurations were used for the analysis.

In Fig 4 we display as functions of temperature the quantities $\langle U \rangle$ (panel (A)), and $\langle U_t^2 \rangle$ (panel (B)), where U_t is the torsional contribution to the potential energy. Ensemble averages are compared with the exact results. The agreement is fairly good, with the residual difference decreasing as the thermalization length of each hybrid MC run is increased (data not shown). It is also useful to compare with the results, also shown in Fig. 4, obtained by averaging U and U_t along the generalized-ensemble hybrid Monte Carlo trajectory. We see that the agreement with the exact results is very satisfactory and even better than in the TMEC case, despite the fact that, as we have seen before, the average molecular size, monitored through the average square end-to-end distance, is incorrectly reproduced as soon as $T < 300$ K (see Fig. 2). This happens despite the fact that even at low temperatures the "entropic $\widetilde{Q}$-term" in the acceptance probability (Eq. (23)) quickly drives the molecule towards the most relevant configurations. The reason is that, when the temperature is too low, configurations that still give a significant contribution to the average end-to-end distance are not sampled adequately because of the height of energy barriers and the fact that the temperature is kept fixed throughout the whole procedure (*i.e.* both while proposing a configuration and in the acceptance/rejection step). Viceversa, by using the approach summarized by Eqs. (20)–(22) such significant contributions to the average end-to-end distance are included, though at the expenses of a less accurate description of the potential energy moments.

The main conclusion of this detailed analysis is that distributions of structural quantities and energy are quite independent from each other. For instance, going back to Fig. 3, we observe that the TMEC thermalization step only slightly modifies the constrained maximal entropy h^2-distributions, thus showing that the latter is already strongly constrained by the average value of $\cos\phi$ through the appropriate exponential modulation. Once this information has been injected, the settling down of the average potential energy due to thermalization and equipartition has little structural effect.

IV. OLIGOPEPTIDE CHAINS

The more complex molecular chains become (as, for instance, because of the presence of more and more complicated type of interactions, like steric exclusions, hydrogen bonds, electrostatic potentials) the larger is the variety of structures that are accessible to the system. Well defined structures or structural basins are in thermal equilibrium at any given temperature and the question arises whether one may be able to characterize the resulting configurational probability distribu-

tions by monitoring the expectation values of some suitably chosen set of physical quantities.

In the case of polypeptides some of these equilibrium structures have been identified by minimizing local steric repulsions or dispersive attractive interactions.[34] These simple methods are effective in a situation where only a subset of the many variables describing the polypeptide conformation can be reliably used for its description. In most cases the selected variables are the two dihedral angles, C(i-1)-N(i)-Cα(i)-C(i) (ϕ) and N(i)-Cα(i)-C(i)-N(i+1) (ψ), of each constituent aminoacid.[15,35] Successively the idea has been largely developed by introducing more and more details in the description of interatomic forces.[36]

The exploration of atomistic force-fields for polypeptides and small proteins has recently achieved quite a remarkable level of efficiency thanks to the important developments in implementing various kinds of multi-canonical and replica-exchange algorithms.[37–40]

In this paper we assume that the outcome of experimental observations carried out in different environmental conditions on a molecular system or the available theoretical results can be summarized in the knowledge of a set of average properties. In this setting we would then like to address here the following questions. What is the amount of entropy reduction when such knowledge is added to a situation where little or no extra information is available? Can we develop a consistent scheme where thermodynamical quantities, like energy and free energy, can be computed once certain structural information are known and introduced to constraint the system configurational probability distribution?

Two oligopeptides, Ala_{12} and Gly_{12}, have been used in this paper as paradigmatic examples to investigate our ability in answering the previous questions by the methods of Sect. II. As the relevant configurational variable we decided to take the α-helicity of the molecule, N_α, where the latter is defined as the number of residues with $260° \leq \phi \leq 320°$ and $293° \leq \psi \leq 353°$.

A. Conformational analysis

We want to study what are the implications of assuming through experiments (or theory) information are available on $\langle N_\alpha \rangle$ at two structurally characteristic temperatures. One is taken to be T_h, *i.e.* the temperature where the helicity is at its maximum and 12 residues, on average, can be identified as being in an α-helix state. T_h is a low temperature which corresponds to a *highly* ordered phase, where the order parameter (that we can identify with $\langle N_\alpha \rangle / 12$) is large, *i.e.* near to 1. The second temperature, that we will call T_l, is the temperature where the average helicity is

zero, as no residue is found in an α-helix state. T_l is a high temperature which corresponds to a *low* order (disordered) phase where the order parameter is very near to zero. Between these two extreme conditions we may assume that the molecule will be found with almost equal probability in any one of its possible configurations. As it has been argued in Sect. II, this situation is approximately described by the "un-modulated" meta probability distribution, $\widetilde{\mathcal{P}}(\mathbf{r})$, with the latter improved making reference to the density of states defined by the average helicity.

The interaction potential of the system has been modeled as it is usually done in all-atom MD simulations of peptides. The AMBER-94 force-field,[41] with relative dielectric permittivity $\epsilon_r = 1$ was used. An atom based cut-off of 0.9 nm was introduced for Lennard-Jones and electrostatic interactions, while at the same time cut-off errors in the electrostatic interactions were reduced by screening (with a screening constant equal to 2 nm^{-1}) and charge-neutralization for pair interactions, according to the algorithm proposed in ref. 42. The N and C terminal groups were modeled as neutral groups adding acyl- and methyl-amide groups at the N and C termini, respectively.

We have acquired a trajectory of 10^5 configurations by constructing a meta probability distribution in the manner described in the previous example of polyethylene, but with $\cos\phi$ replaced by the average molecular α-helicity, N_α, as the relevant configurational variable upon which the iterative procedure illustrated in Sect. II B is based. This meta probability distribution was then modulated through the constrained maximal entropy method by requiring it to lead to the desired values of $\langle N_\alpha \rangle$. The interval 0 to 12 was divided in 100 parts and 100 constrained maximal entropy configurational probability distributions were constructed each corresponding to one of the values, n_α, taken by $\langle N_\alpha \rangle$. Both for polyalanine and polyglycine, one iteration was enough to obtain convergence of the procedure. These modulated distributions have been used in the analysis that follows.

As a first step of our investigation, we studied the effect of injecting information concerning the average helicity of the molecule on the meta probability distribution, $\widetilde{\mathcal{P}}$, with consequent cross-entropy reduction, by monitoring how the modulating constrained maximal entropy factors affected the final distribution. This was done in three selectively chosen cases: $\langle N_\alpha \rangle = 0$ (panels A and D of Figs. 5 and 7), no constraint on $\langle N_\alpha \rangle$, *i.e.* $\lambda = 0$ (panels B and E) and $\langle N_\alpha \rangle = 12$ (panels C and F).

In Fig. 5 we show the histograms of the logarithm of the probability of finding the dihedral angles ϕ and ψ of residue 5 of the alanine (panels A to C) and glycine (panels D to G) oligomers within intervals of $\pm 10°$. The three pairs of histograms correspond to the three different types of

constraints we have imposed on the value of the oligomer average helicity, as explained above. Black squares correspond to values of the logarithm equal to -1. Each successive lighter (gray) colour represents a decrease by one unit.

Naturally, configurations where the modulating factor is large contribute more to build up the desired value of the average helicity and these configurations can be considered to represent adequately the whole set of statistically relevant configurations.

The relevant conformations corresponding to no helicity information and to low average helicity are more widely distributed for glycine than for alanine. For the alanine oligomer certain regions of the plot are excluded because of the steric repulsion between methyl groups, while others are more densely populated because of the dispersive attraction between methyl groups.

It is important to remark that the glycine oligomer has access to both the α-helix and the β-sheet ($\phi, \psi \sim 180°$) regions, while the latter is empty for the alanine oligomer.

In Fig. 6 the histograms of the helicity distribution obtained in the three cases considered above are displayed for alanine (A-C) and glycine (D-F) oligomers. It may be observed that the distributions corresponding to minimum (A and D) and maximum (C and F) helicity look very much the same when comparing the two aminoacids, while the distributions with no helicity information, which correspond to un-modulated meta probabilities, show the most significant differences.

We also see that, even using the un-modulated meta probability distribution, the alanine oligomer is partially structured in an α-helix conformation (see also the discussion below), although the peptidic linkages in the helical state appear distributed along the chain and are not always consecutive. It is remarkable that under identical physical conditions the glycine oligomer shows a much lower propensity to form helical structures than alanine does. In the case of glycine, in fact, conformations with only one residue in a helical state are the mostly represented ones. In other words, when no constraint on the value of the average helicity is imposed, the phase-space region where the molecule finds itself in a α-helix conformation is significantly populated in the case of Ala_{12}, but not in the case of Gly_{12}. We may interpret this fact by saying that the helix-formation propensity of Ala_{12} is already partially contained in the nature of the force-field model we have employed, irrespectively of the temperature.

As for Gly_{12}, the presence of a non negligibly small population in the $\phi, \psi \sim 180°$ region is an interesting indication of its larger propensity, compared to Ala_{12}, to form β-sheet structures.

In Fig. 7, configurations randomly selected among those plotted in Figs. 5 and 6, *i.e.* corresponding to the maximum value of the modulating factor, are displayed. The pictures shown in

panels B and E represent random samples of the structures of the alanine and glycine oligomers, respectively, chosen among the configurations of the unconstrained meta probability distribution, *i.e.* where the modulating factor is constant ($\lambda = 0$). For alanine, most of the structures found in the unconstrained meta distribution, display nascent O(i)$\cdots$H-N(i+4) hydrogen bonds typical of an α-helix motif, as well as methyl-methyl contacts. The structure displayed in Fig. 7B precisely represents this kind of situation. The same type of hydrogen bonds are not as frequently represented in glycine, where in most of the configurations the molecular chain is definitely less structured, as it is seen in Fig. 7E.

Not surprisingly, for both Ala_{12} and Gly_{12} the constraint of high helicity (panels C and F) enhances the population of residues in an α-helix conformation and the importance of structures where the array of backbone hydrogen bonds is almost completely established together with methyl-methyl contacts (when the latter are available – alanine). Viceversa, the constraint of having low average helicity lowers the population in the high α-helix region and increases the population of other regions, more significantly than in the absence of any helicity constraint (no information).

The case of zero helicity (panels A and D) shows that also an order parameter lower than the value one can estimate in the case of the unconstrained meta distribution from panels B and C of Fig. 6, introduces significant information. As expected, in the case of zero helicity the representative structures rarely display the interactions typical of an α-helix. Since there are not many ways to break all of them, given the other structural constraints of the chain, the density of helical configurations is reduced by increasing the population of non-helical regions in the ϕ/ψ map already represented in the unconstrained statistics without creating new adequately populated regions.

In Sect. IV C the implication of the above analysis for devising a strategy aimed at computing free energy differences will be presented and discussed.

B. Energy and temperature

The application reported in Sect. III, concerning the different role and use of the constrained maximal entropy and thermalized constrained maximal entropy approaches in the study of a simple polyethylene model, suggests that when a collective variable is assigned an average value that is temperature dependent, the thermally accessible potential energy values will also be correspondingly biased. This circumstance also emerges in the cases of the two oligopeptides we are

discussing in this section.

The average value of the potential energy U has been computed for both alanine and glycine oligomers taking, as starting points for the hybrid MC thermalization trajectories, 10^4 configurations, randomly sampled from the random walk probability distribution, recursively refined as described in Sect. II B. Thermalization has been performed at temperatures uniformly distributed in the range from 50 K to 650 K, separated by 50 K. Each hybrid MC trajectory consists of 100 MD runs, where every run is a sequence of 10 iterations with a time-step of 1 fs.

In Fig. 8, the average of the potential energy, $\langle U \rangle$, is plotted as a function of λ for the above values of T for the alanine (A) and glycine (B) oligomers in vacuum. The values of the potential energy are normalized to $dRT/2$, where R is the gas constant and $d = 3N_a - 3$ is the number of degrees of freedom of the molecule, N_a being the number of atoms. We can see that for alanine $\langle U \rangle$ tends to be independent of λ for temperatures larger than about 350 K, while for glycine the same behaviour shows up at lower temperatures. To analyze this phenomenon in further details, in Fig. 9 the average potential energy difference, $\Delta\langle U \rangle$ (always in units of $dRT/2$), between the ordered (λ corresponding to helicity 12) and the disordered (λ corresponding to vanishing helicity) states is plotted as function of temperature, for alanine (squares) and glycine (circles), respectively.

We recognize from the figure that $\Delta\langle U \rangle$ for Ala_{12} is twice the value of Gly_{12}, indicating that the alanine oligomer undergoes the helix-coil transition at a temperature lower than the glycine oligomer by almost a factor of two. Moreover, for the glycine oligomer the energy difference becomes smaller than $dRT/2$ at $T \sim 100$ K, while for alanine the same phenomenon occurs at a somewhat higher value, $T \sim 150$ K. The above observations suggest that for temperatures beyond a critical value, characteristic of the molecule, the decrease in average potential energy, consequent of having imposed that the average helicity has some given value, is negligible compared to the approximate thermal contribution, $RT/2$, associated with each degree of freedom. Beyond this critical temperature, in fact, effects coming from configurational information constraints are no longer visible and equipartion takes over.

Other models of aminoacid oligomers have been studied employing many of the existing variants of the multi-canonical MC simulation strategies.[38,40] For the alanine oligomer in vacuum, a helix-coil transition temperature in the range 400-450 K was found. A less visible transition for glycine at temperatures around 200 K was not related to the formation of an α-helical state but rather to a molecular collapse, and an extended α-helix structure was never found for the glycine oligomeer. Although the TMEC method proposed in the present paper cannot predict the transition

temperature of a given molecule, it allows a fair comparison between different molecules as for their propensity to give rise to definite conformational transitions. Compared to previous works, in the approach we propose we can rather accurately monitor the coil/helix transition of both alanine and glycine oligomers, because extended α-helical structures are significantly represented in the modulated configurational probability distributions that we construct. As a result, we can also address the problem of computing the change of average potential energy in the transition between the two states (see below Sect. IV C).

When we look at the behaviour of the average helicity, $\langle N_\alpha \rangle$, as a function of λ at different temperatures, the situation completely changes compared to what we have observed in the case of the average potential energy. In Fig. 10 we see that the values of the average helicity corresponding to different temperatures (the bunch of curves labeled by TMEC in the figure) only slightly differ among themselves, precisely because information on helicity is injected in the meta probability distribution through the maximal entropy modulation factor. At all temperatures thermalization produces a small increase in helicity that is completely negligible with respect to the effect of changing the preassigned value that constraints the magnitude of $\langle N_\alpha \rangle$. Beyond values of the helicity of the order of $6 \div 8$, the MEC curve is only slightly above the TMEC curves at all temperatures.

For glycine temperature does not affect helicity values lower than 6, which shows that the constrained random walk (MEC curve in Fig. 10) already gives a rather good representation of the probability distribution of chain configurations with low helicity. The most evident difference between alanine and glycine oligomers (see Figs. 6) is that, when $\lambda = 0$, *i.e.* when no information on helicity is provided, helicity is larger for alanine (~ 8) than for glycine (~ 1). This confirms again that the propensity to form helical segments is larger for polyalanine than for polyglycine, irrespective of the temperature.

C. Computing entropy and free energy differences

In Sect. II we have computed the maximum value attained by the cross-entropy functional, once constraints have been imposed on the random walk probability distribution in the form of preassigned values for the thermal averages of certain physical quantities. The maximum, $S_c^{(m)}$, is expressed by Eq. (12). Up to an irrelevant constant, this quantity can be identified with the informational entropy associated with that particular state of the system. This is a very important

observation, as it allows us to use Eq. (12) to compute entropy differences between pair of states. In fact, we recall that, by extending Shannon observation that any new information one can get on a system reduces its entropy, the constrained maximal entropy principle relies on the idea that the least biased way to exploit the new information is to have the minimal decrease of the (informational) entropy. In this context it seems natural to assume that the thermodynamic entropy differences between any two states can be evaluated computing the corresponding informational entropy difference.

With these premises, within the simplified approach we have described in Sect. IV A, we want to compute the entropy difference between the molecular states h and l, previously defined as the states of *high* and *low* order of an oligopeptide, respectively. We recall that the state h is characterized by a high value of $\langle N_\alpha \rangle_h = n_h \sim 12$ and a low temperature, T_h, while state l is characterized by a low value of $\langle N_\alpha \rangle_l = n_l \sim 0$ and a high temperature, T_l. For the difference $\Delta S_{hl} = S(n_h) - S(n_l)$, using Eq. (12) with $M = 1$, we get

$$\Delta S_{hl} = S(n_h) - S(n_l) = k_B[\log \mathcal{Z}(\lambda_h) + \lambda_h n_h - \log \mathcal{Z}(\lambda_l) - \lambda_l n_l] \,, \tag{24}$$

where $\lambda_{h/l} = \lambda(T_{h/l})$. Strictly speaking, ΔS_{hl}, as defined by the equation above is not the canonical entropy difference, as the canonical entropy should be expressed as a function of the energy of the state. We will, however, assume that ΔS_{hl} gives a reliable estimate of it, *i.e.* we will assume the validity of the approximate relation

$$S(n_h) - S(n_l) \simeq S(\langle U \rangle_h) - S(\langle U \rangle_l) \,. \tag{25}$$

This assumption is suggested by the following line of arguments.

- If the temperature dependence of the preassigned thermal averages, that determine the constrained maximal entropy modulation factors of the meta probability distribution, is known, we can reconstruct the temperature dependence of $\langle U \rangle$ (and viceversa) through the formula

  $$\langle U(T) \rangle \simeq \langle U(\lambda(T)) \rangle \,, \tag{26}$$

  where $\lambda(T)$ is the value that results from solving the consistency equations for the Lagrange multiplier;

- On the other hand, using the possible knowledge on $\langle U(T) \rangle$ in the constrained maximal entropy procedure, together with preexisting conformational information (at the same temperature), would not appreciably modify the final form of the modulated meta probability distribution.

A comment is in order here. The assumption underlying the previous considerations is that adding information about $\langle U(T)\rangle$ on the modulated meta distribution tends to affect the latter less and less, as the number of configurational constraints increases. Ideally, if one could impose constraints so as to completely fix the configuration of the molecule, the expectation value of U (and, indeed, of any other physical quantity) could be exactly computed.

The Helmoltz free energy difference between the states h and l is

$$\Delta F = F_h - F_l = \langle U\rangle_h - \langle U\rangle_l - [T_h\, S(\langle U\rangle_h) - T_l\, S(\langle U\rangle_l)]\,, \tag{27}$$

which, using Eq. (26), can be rewritten in the form

$$\begin{aligned}\Delta F &= \langle U(\lambda_h)\rangle - \langle U(\lambda_l)\rangle - [T_h\, S(\langle U(\lambda_h)\rangle) - T_l\, S(\langle U(\lambda_l)\rangle)] \equiv \\ &\equiv \Delta F_U + \Delta F_S\,.\end{aligned} \tag{28}$$

It is important to realize that, once the two temperatures, T_h and T_l, and the values of a set of order parameters, $\langle\vec{A}\rangle = (\langle A^{(k)}\rangle, k = 1,\ldots,M)$, at the chosen temperatures, are known (in the example we are discussing, where $M = 1$, the configurational variable, A, on which the whole construction is based, is the average molecular α-helicity), one can compute the potential energy difference, $\langle U(\lambda_h)\rangle - \langle U(\lambda_l)\rangle$, hence ΔF_U, and the entropy difference $S(\langle A\rangle_h) - S(\langle A\rangle_l)$, but not immediately the entropic contribution to the free energy, *i.e.* the quantity ΔF_S. Indeed, the entropy difference can be computed by using the modulated meta probability distribution which results from the constrained maximal entropy procedure. The potential energy variation can be evaluated after thermalizing the modulated meta distribution at the two temperatures, T_h and T_l, and then separately computing the two potential energy expectation values.

Despite the fact that one cannot directly access the full Helmoltz free energy difference, one can get a useful inequality as for the magnitude of the entropic term, ΔF_S, using again Eq. (25). From $T_h \leq T_l$ and the second law of thermodynamics (entropy is a state function which grows monotonically with the amount of "disorder" of the state), one immediately derives the following chain of inequalities

$$\begin{aligned}&T_h\,[S(n_h) - S(n_l)] \simeq T_h\,[S(\langle U\rangle_h) - S(\langle U\rangle_l)] \leq \\ &\leq T_h\, S(\langle U\rangle_h) - T_l\, S(\langle U\rangle_l) \leq \\ &\leq T_l\,[S(\langle U\rangle_h) - S(\langle U\rangle_l)] \simeq T_l\,[S(n_h) - S(n_l)]\,.\end{aligned} \tag{29}$$

The relation (29) thus tells us that the quantity ΔF_S must fall inside the interval $[T_h\Delta S, T_L\Delta S]$. This inequality can be rather useful in practice. We will, in fact, show below that in some cases

it can provide a very tight bound to the possible values of ΔF_S. We will illustrate this point in the example of alanine and glycine oligomers. In both cases we will assume the low-order state, l, to be characterized by the parameters $T_l = 450$ K and small average molecular α-helicity. For simplicity we will take $\lambda_l = 0$. In this way the configurations that describe the low-order state are those obtained by thermalizing at high temperature, $T_l = 450$ K, the meta probability distributions, refined as described in Sect. II B. The state h is assumed to live at $T_h = 50$ K and to be characterized by the value of λ_h that corresponds to the highest possible helicity. In Table I, the computed values of ΔF_U and the bounds for ΔF_S, corresponding to the $l \rightarrow h$ (coil $\rightarrow$ helix) transition, are shown for the two oligomers. In both cases the energetic contribution, ΔF_U, is negative, while the entropic contribution ΔF_S is bound by two positive numbers, in accordance with what is expected when order increases in the transition. We remark that, although, as we said above, we are not in the position to compute the actual value of ΔF_S, we can nevertheless get useful information about it from our approach. We see, in fact, that the entropic contribution is negligible in the case of the alanine oligomer and between 2% to 20% of the energetic contribution in the case of the glycine oligomer.

We also notice that for alanine, the energetic contribution is larger in absolute value than for glycine, due to the contribution of dispersive attractive interactions between methyl groups that are absent in the case of glycine. These interactions are effective also when $\lambda = 0$, because in this situation alanine average helicity is 8, hence substantially larger than that of glycine which stays around 1 at all temperatures (see Fig. 6E).

The fact that ΔF_U is more negative for alanine than for glycine can be explained in the following way. The decrease of potential energy upon structure formation is expected to be significant when the interacting sites in the ordered state have been confined to a limited region of the conformational space. This is the condition imposed by the constraint on the average helicity, which is effective when such interaction sites are structurally present, as in the case of alanine.

The entropy decrease upon ordering is far less predictable, both because of conformational exclusions in the disordered state and residual flexibility in the ordered state. However, from Figs. 5 and 6, we realize that, when a high helicity value is forced on the system, the helicity distributions of alanine (C) and glycine (F) are very similar, while in the disordered state (B and E) glycine displays highly dense population in regions that are unaccessible to the alanine oligomer.

The sparser conformations of glycine contribute to low helicity and high potential energy (no hydrogen bonds), while the less sparse conformations of alanine contribute to helicity in the range

$6 \div 12$. Therefore, the fact that the entropic contribution, ΔF_S, to the Helmoltz free energy is larger for glycine than for alanine should be interpreted in terms of a higher "localization" of configurations in the conformational map upon ordering when moving from state l to h.

We end this section by comparing some of our results with those previously obtained on the same molecular systems employing appropriately tuned multi-canonical simulation strategies.[38,40] The most striking difference between the two sets of results in visible in the computed value of the potential energy variation in the coil to helix transition. Our result for the alanine oligomer is definitely larger (in absolute value) than what was observed by using the ECEPP/2 force-field (-650 kJ/mol vs -160 kJ/mol). We interpret this difference as due to differences in force-field used in the simulations, having recognized that the AMBER force-field strongly favours the α-helix. However, we wish to observe that no comparison of potential energy variations between alanine and glycine oligomers was possible in previous works, because no α-helical state for the glycine oligomer was ever detected.

Qualitatively the physico-chemical description of the transition is quite similar in our approach and in that of refs. 38,40 The contribution to the free energy variation due to the potential energy is for both systems (Ala and Gly) negative and more negative for Ala than for Gly, while the entropic contribution is in both cases positive. The only difference is that the variation of the entropic contribution in the ECEPP/2 multi-canonical simulations is competitive with the variation of potential energy contribution. As we explained above, this fact should be attributed to the different force-fields that have been employed in the simulations.

V. CONCLUSIONS

In this paper we have presented a new computational strategy for the study of statistical properties of polymers, based on a three-step construction of the configurational probability distribution of the system.

1. One starts from a random sampling of the system configurational space, in which configurations are generated by MD moves with velocities taken from a Maxwell-Boltzmann distribution at a randomly chosen temperature in the range from 0 to 1000 K.

2. The ergodicity of the resulting distribution is iteratively improved by means of a multi-canonical-like procedure. The main novelty of the present approach is that the accep-

tance/rejection test in this step is based on the density of states, $g_{\vec{A}}$, associated to a set of selected configurational quantities, $\vec{A}$, rather than to the density of states associated with the potential energy, as it is most often done. At each iteration the current meta distribution is modulated by exponential factors which are determined by requiring that the quantities $\vec{A}$ (or other suitably chosen configurational quantities) have certain (possibly temperature dependent) values. This modulation is the result of maximizing the cross-entropy functional which encodes the amount of knowledge we (pretend to) have on the system.

3. If required, as it is necessary to do if averages associated with the potential energy and its moments have to be computed, one proceeds to thermalize at a desired temperature the set of configurations collected in the meta trajectory. The thermalization is performed by means of (short) hybrid Monte Carlo simulations starting from configurations uniformly sampled within the existing meta trajectory.

Two simple molecular systems have been chosen as examples for the validation of the method. A bead-and-spring model of polyethylene with independent hindered torsions, that can be solved exactly, and an all-atom model of Ala_{12} and Gly_{12} oligomers in vacuum, extensively studied numerically. It is remarkable that in both cases, many of the aspects of the Statistical Mechanics of the model systems could be adequately explored and elucidated using information on only a single configurational variable. The success of the construction is strongly related to the real amount of information encoded in such a pivotal variable that should, therefore, be carefully chosen. The success of the applications we have described shows that when the selected configurational variable is known (either through experiments or theory) as function of the temperature or other relevant environmental parameters, the information that is gathered can be used to construct a fairly accurate probability distribution of the molecule configurations.

We have observed that, once structural information have been injected into the meta configuration probability distribution, the effect of the thermalization step on configurational averages is negligible. Or in other words that the change in the shape of the distributions due to constraints imposed on structural variables can be more significant than the effect of a mere modification of the temperature. The small sensitivity of configurational distributions to temperature can be related to the fact that we have neglected all possible temperature dependence in modeling the classical force-field of the oligomers: all the force-field parameters do not depend explicitly on temperature and they encode almost properties of simple molecules at room conditions.

An important outcome of the method is that it consistently allows to estimate the change in free energy due to changes in average structural parameters.

The procedure devised to obtain the meta configuration probability distribution is the most critical step in the whole method. In this work the basic new idea was to construct a steady state flow within the space of configurations where the weight of each configuration would tend to become almost constant for the given force-field. But any other method may be used, provided the resulting configurational probability distribution is stationary with respect to the selected process of generating and collecting configurations. The key point is that this meta distribution contains information about which regions are forbidden and which other ones should become the seed of further accumulation of population.

The main virtue of the method we are advocating in this paper is that it is ready to include in the meta distribution *a priori* information not only in terms of interatomic interactions, but also in terms of contributions from complicated mean fields or entropy that can significantly modify the density of states. Partial information on free energy variations, that may be available from theory, can be included in the construction of the meta distribution via the constrained maximal entropy procedure allowing for the identification of the important structures that are compatible with such information. In conclusion, instead of asking the numerical approach to yield the whole Statistical Mechanics of the system, we propose to combine detailed information coming from numerical investigations of the configurational space with the knowledge we might have on the nature and the role of selected structural parameters that significantly affect the behaviour of the system.

Acknowledgments - G.L.P. wishes to thank prof. Piero Procacci (University of Florence, Italy) for suggestions regarding the manuscript. G.C.R. thanks the Humboldt Foundation for partial financial support and NIC at Desy-Zeuthen for hospitality while this work was completed.

Appendix

In this Appendix, following ref. 24, we recall the construction of the probability distribution $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$, which enters as a first step in the iterative construction of Sect. II B.

Random walk configurations are generated by sequential MD moves of fixed length by taking as initial system coordinates the coordinates of the last stored configuration and as initial particle velocities the values extracted from a Maxwell-Boltzmann distribution at a random temperature, uniformly chosen at each MD step within zero and a high-temperature limit (1000 K in this work).

This procedure obeys the detailed balance principle and generates a time independent (stationary) conditional probability, $P_c[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}]$, where $\{\mathbf{r}_1\}$ and $\{\mathbf{r}_2\}$ are the initial and final configurations of each move. Although unknown, P_c is perfectly well defined and generates an acceptable initial probability distribution, $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$.

To prove that the above procedure obeys the principle of detailed balance, let P_{sug} and P_{acc} be the probabilities of suggesting and accepting the move $\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}$, respectively. The conditional probability P_c of the move is

$$P_c[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}] = P_{sug}[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}]\, P_{acc}[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}]\,. \tag{30}$$

Different choices for P_{acc} are possible. Let us start by discussing the case $P_{acc} = 1$, *i.e.* the case in which all configurations generated by the algorithm we have described are accepted. From

$$P_{sug}[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}] = D[\{\mathbf{r}_1, \mathbf{v}_1\}]\, G_{MD}[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}]\,, \tag{31}$$

where D is the probability of starting with a given set of particle positions and velocities and G_{MD} is the deterministic MD propagator, one gets

$$P_c[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}] = D[\{\mathbf{r}_1, \mathbf{v}_1\}]\, G_{MD}[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}]\,. \tag{32}$$

Since we are recording all the configurations that are generated, we can safely assume that the distribution D does not depend on the coordinates, $\{\mathbf{r}_1\}$, and that for every particle the probability of having a velocity $\mathbf{v}$ is identical to that of having $-\mathbf{v}$. The important observation here is that, if all the atomic velocities in the configuration $\{\mathbf{r}_2\}$ are inverted, the deterministic MD propagator G_{MD} drives the particle positions exactly from $\{\mathbf{r}_2\}$ back to $\{\mathbf{r}_1\}$. This is true also for all the time-reversible MD algorithms that approximate the deterministic evolution in actual MD simulations. Therefore, every move $\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}$ has a corresponding opposite move $\{\mathbf{r}_2\} \to \{\mathbf{r}_1\}$ with identical probability. This is sufficient for P_c to satisfy the detailed balance principle.

If $P_{acc} \neq 1$, P_c still satisfies the principle of detailed balance, provided P_{acc} does not depend on velocities. Moreover, if we put, as usual

$$P_{acc}[\{\mathbf{r}_1\} \to \{\mathbf{r}_2\}] = \min\left[1, \frac{e^{-F[\{\mathbf{r}_2\}]}}{e^{-F[\{\mathbf{r}_1\}]}}\right], \tag{33}$$

where $F[\{\mathbf{r}\}]$ is a given positive function of the particle coordinates, then the generated probability distribution will be related to the probability distribution $\widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]$, constructed before, by the formula

$$\widetilde{\mathcal{P}}^{(0)}_{(F)}[\{\mathbf{r}\}] \propto \widetilde{\mathcal{P}}^{(0)}[\{\mathbf{r}\}]\, e^{-F[\{\mathbf{r}\}]}\,. \tag{34}$$

[1] C. Branden and J. Tooze, *Introduction to protein structure* (Garland Publishing Inc., London, UK, 1999).

[2] W. Saenger, ed., *Principles of Nucleic Acid Structure* (Springer-Verlag, New York, USA, 1984).

[3] V. S. R. Rao, P. K. Qasba, P. V. Balaji, and R. Chandrasekaran, *Conformation of Carbohydrates* (Harwood Academic Publishers, Amsterdam, The Netherland, 1998).

[4] W. L. Mattice and U. W. Suter, eds., *Conformational Theory of Large Molecules. The Rotational Isomeric State Model in Macromolecular Systems* (John Wiley & Sons, New York, USA, 1994).

[5] B. Kuhlman and D. Baker, Curr. Opin. Struct. Biol. **14**, 89 (2004).

[6] U. H. E. Hansmann, *New Optimization Algorithms in Physics* (VCH-Wiley, New York, USA, 2004), chap. Protein Folding in Silico - The Quest for Better Algorithms.

[7] B. A. Berg, J. Stat. Phys. **82**, 323 (1996).

[8] B. A. Berg, Fields Institute Communications **26**, 1 (2000), see also cond-mat/9909236.

[9] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers (Pept. Sci.) **60**, 96 (2001).

[10] U. H. E. Hansmann, Comp. Phys. Comm. **147**, 604 (2002).

[11] C. Bartels and M. Karplus, J. Phys. Chem. B **110**, 8254 (1999).

[12] Y. M. Rhee and V. S. Pande, Biophys. J. **84**, 775 (2003).

[13] Y. Okamoto, J. Mol. Graph. Mod. **22**, 425 (2004).

[14] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, USA, 1989).

[15] P. J. Flory, *Statistical Mechanics of Chain Moleculs* (Hanser, New York, USA, 1989).

[16] A. Grosberg, A. R. Khokhlov, and A. R. Khoklov, *Statistical Physics of Macromolecules* (Springer-Verlag, New York, USA, 1997).

[17] S. Kirkpatrick, C. Gelatt, Jr., and M. Vecchi, Science **220**, 671 (1983).

[18] Y. Okamoto, *Encyclopedia of Optimization* (Kluwer Academic, Dordrecht, The Netherlands, 2001), vol. III, chap. Monte Carlo Simulated Annealing in Protein Folding, pp. 425–439.

[19] J. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbour, USA, 1975).

[20] W. Wenzel and K. Hamacker, Phys. Rev. Lett. **82**, 3003 (1999).

[21] B. A. Berg, G. La Penna, V. Minicozzi, S. Morante, and G. C. Rossi, *Proceedings of the MODSIM 2003 Conference* (Modelling and Simulation Society of Australia and New Zeland, Townsville, Australia, 2003), vol. 4, chap. Multi-canonical Algorithms for Folding Processes, pp. 1967–1972,

http://mssanz.org.au/modsim03/modsim2003.html.

[22] B. A. Berg, Phys. Rev. Lett. **90**, 180601 (2003).

[23] B. A. Berg, Comp. Phys. Comm. **153**, 397 (2003).

[24] G. La Penna, J. Chem. Phys. **119**, 8162 (2003).

[25] P. Attard, J. Stat. Phys. **100**, 445 (2000).

[26] B. A. Berg, H. Noguchi, and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).

[27] M. Marchi and P. Ballone, J. Chem. Phys. **110**, 3697 (1999).

[28] E. Paci, M. Vendruscolo, C. Dobson, and M. Karplus, J. Mol. Biol. **324**, 151 (2002).

[29] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. USA **20**, 12562 (2002).

[30] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2003).

[31] A. Fishcer, F. Cordes, and C. Schütte, J. Comp. Chem. **19**, 1689 (1998).

[32] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comp. Phys. **23**, 327 (1977).

[33] F. Eisenmenger, U. H. E. Hansmann, S. Hayryan, and C.-K. Hu, Comp. Phys. Comm. **138**, 192 (2001).

[34] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. USA **37**, 205 (1951).

[35] D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers* (Academic Press, New York, USA, 1970).

[36] P. von Ragué Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner, eds., *The Encyclopedia of Computational Chemistry* (John Wiley & Sons, Chichester, UK, 1998).

[37] Y. Okamoto and U. H. E. Hansmann, J. Phys. Chem. **99**, 11276 (1995).

[38] A. Mitsutake and Y. Okamoto, J. Chem. Phys. **112**, 10638 (2000).

[39] A. E. Garca and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **24**, 13898 (2003).

[40] B. S. Kinnear, M. F. Jarrold, and U. H. E. Hansmann, J. Mol. Graph. Mod. **22**, 397 (2004).

[41] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. J. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).

[42] D. Wolf, P. Keblinski, S. R. Phillpot, and J. Eggebrecht, J. Chem. Phys. **110**, 8254 (1999).

[43] To prove this statement it is enough to note the relation $\sum_\gamma \mathcal{P}_\gamma \log \mathcal{P}_\gamma/\widetilde{\mathcal{P}}_\gamma = \sum_\gamma \widetilde{\mathcal{P}}_\gamma \left[\mathcal{P}_\gamma/\widetilde{\mathcal{P}}_\gamma \log \mathcal{P}_\gamma/\widetilde{\mathcal{P}}_\gamma - \mathcal{P}_\gamma/\widetilde{\mathcal{P}}_\gamma + 1\right] \geq 0$, which follows from the elementary inequality $x \log x \geq x - 1$, valid for any finite, non-negative x.[25]

Table I: Helmoltz free energy differences (in kJ/mol) between the state l ($T = 450$ K and $\lambda = 0$) and the state h ($T = 50$ K and average helicity $\simeq 12$) of alanine and glycine oligomers. Statistical errors on ΔF_U are on the last digit and are shown in brackets.

AA	ΔF_U	ΔF_S
Ala_{12}	-650 (3)	$1 \div 9$
Gly_{12}	-480 (4)	$10 \div 90$

Figure captions

- Figure 1 - The parameter λ of the polyethylene chain as a function of the inverse temperature β at progressively improved meta distributions: first iteration (MEC-0) solid line; second iteration (MEC-1) dashed line; third iteration (MEC-2) dotted line.

- Figure 2 - Average end-to-end square distance, $\langle h^2 \rangle$, of polyethylene chain with 40 methylene units as a function of the temperature, T: exact result (□); generalized-ensemble (multi-canonical) Monte Carlo (GEMC, ○); maximal constrained entropy (MEC, △); generalized-ensemble hybrid Monte Carlo (GEHMC, ▽). Errors (not shown) are ± 0.16 and ± 0.32 nm^2 for GEMC and MEC data, respectively.

- Figure 3 - Distribution of the square end-to-end distance, $P = P(h^2)$, obtained from generalized-ensemble (multi-canonical) Monte Carlo (GEMC, solid line), maximal constrained entropy (MEC, dashed line) and maximal constrained entropy thermally equilibrated by hybrid Monte Carlo (TMEC, dotted line) simulations. GEMC curves are at $T =$ 100 K (A), 300 K (B) and 600 K (C), while MEC and TMEC curves are obtained using as a constraint the value of $\langle \cos \phi \rangle$ at the corresponding GEMC temperatures.

- Figure 4 - (A) Average total potential energy, $\langle U \rangle$, and (B) square torsional potential energy, $\langle U_t^2 \rangle$, as functions of temperature, T: exact results (□); generalized-ensemble hybrid Monte Carlo (GEHMC, ▽); thermalized maximal constrained entropy (TMEC, △). In the TMEC procedure, errors (not shown) are $\pm 3 \times 10^{-2}$ and $\pm 5.5 \times 10^{-2}$ in units of 10^{-2} kj/mol and $(10^{-2}\,\mathrm{kj/mol})^2$ in panels (A) and (B), respectively.

- Figure 5 - Logarithm of the probability of finding the dihedral angles ϕ and ψ of the residue 5 of Ala_{12} (panels A, B, C) and Gly_{12} (panels D, E, F) within intervals of $\pm 10°$. Results for the MEC trajectory with the constraint $\langle N_\alpha \rangle = 0$ are shown in panels A and D; those for the MEC trajectory with no constraint in panels B and E and those for the MEC trajectory with the constraint $\langle N_\alpha \rangle = 12$ in panels C and F. Colours are as explained in the text.

- Figure 6 - Lego plot of the probability, $P = P(n)$, of finding n residues in an α-helix state. Different panels refer to the different trajectories and oligomers as described in the caption of Fig. 5.

- Figure 7 - Typical oligomer configurations randomly selected among those corresponding to the maximum value of the modulating factor in the plots of Fig. 5.

- Figure 8 - Average potential energy per degree of freedom as function of λ at different temperatures for Ala_{12} (panel A) and Gly_{12} (panel B); from top to bottom temperatures are 50, 150, 250, 350, 450, 550 and 650 K.

- Figure 9 - Difference of average potential energy between ordered ($n_\alpha = 12$) and disordered ($n_\alpha = 0$) states for Ala_{12} (squares) and Gly_{12} (circles) as function of temperature.

- Figure 10 - Average α-helicity as function of λ at different temperatures for Ala_{12} (A) and Gly_{12} (B); thick curve corresponds to the maximal constrained entropy (MEC) trajectory, thin curves to the thermalized maximal constrained entropy (TMEC) trajectory at the same set of temperatures as in Fig. 8.

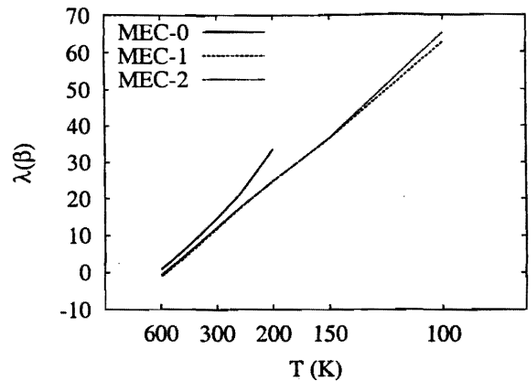


Figure 1: G. La Penna, "Designing generalized statistical..."

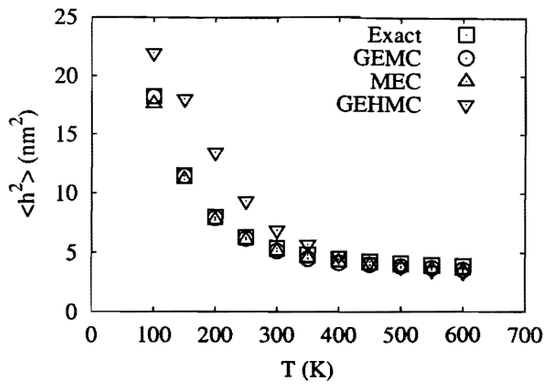


Figure 2: G. La Penna, "Designing generalized statistical..."

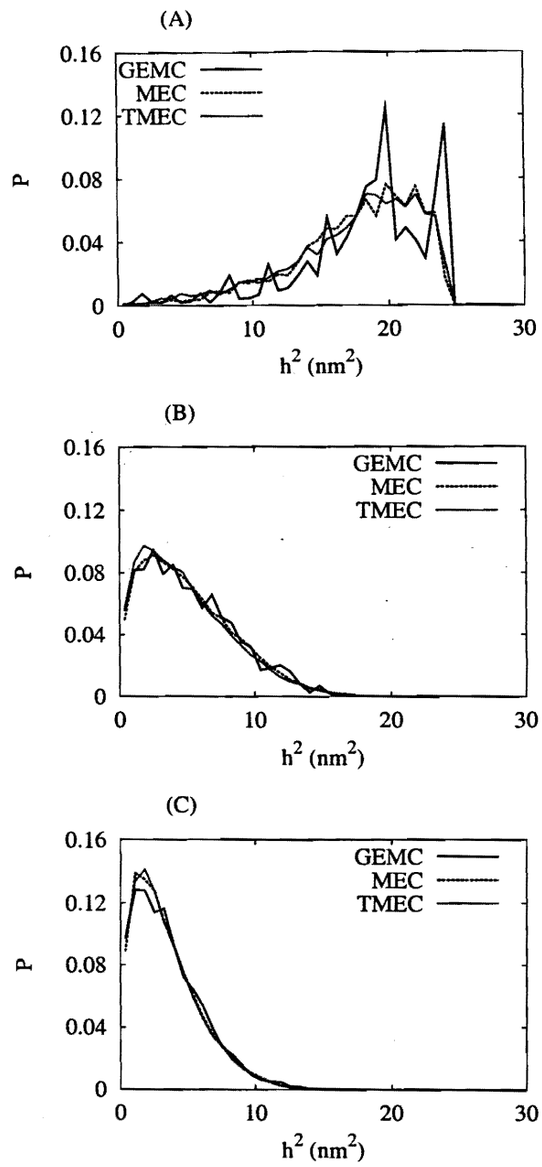


Figure 3: G. La Penna, "Designing generalized statistical..."

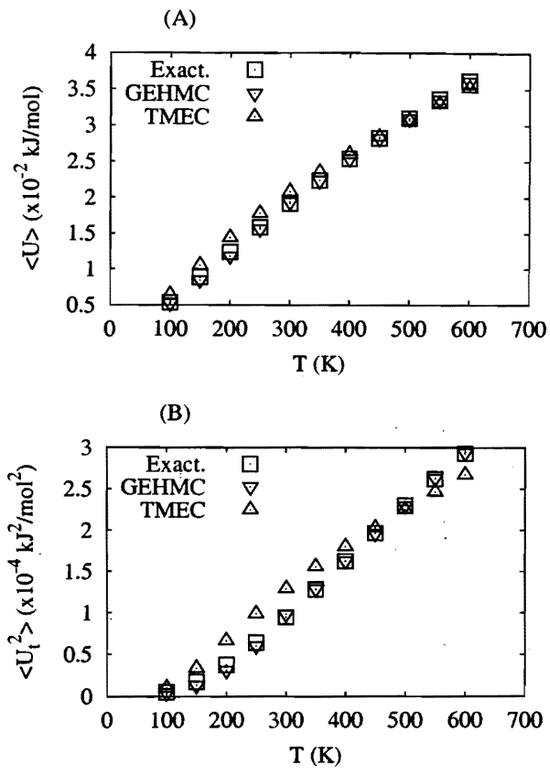
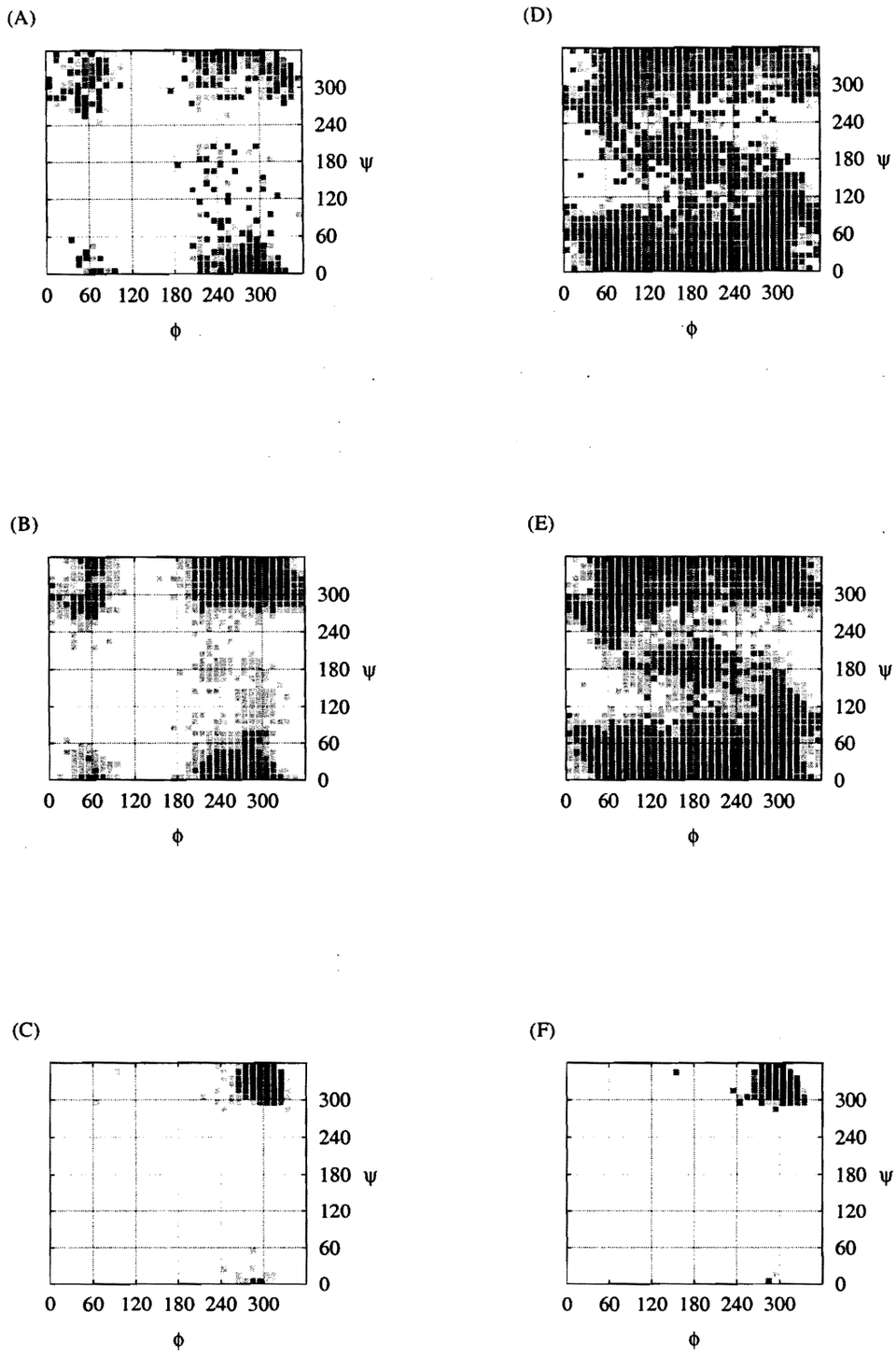


Figure 4: G. La Penna, "Designing generalized statistical..."


(A)
(B)
(C)
(D)
(E)
(F)
0 60 120 180 240 300
ϕ
0 60 120 180 240 300
ψ

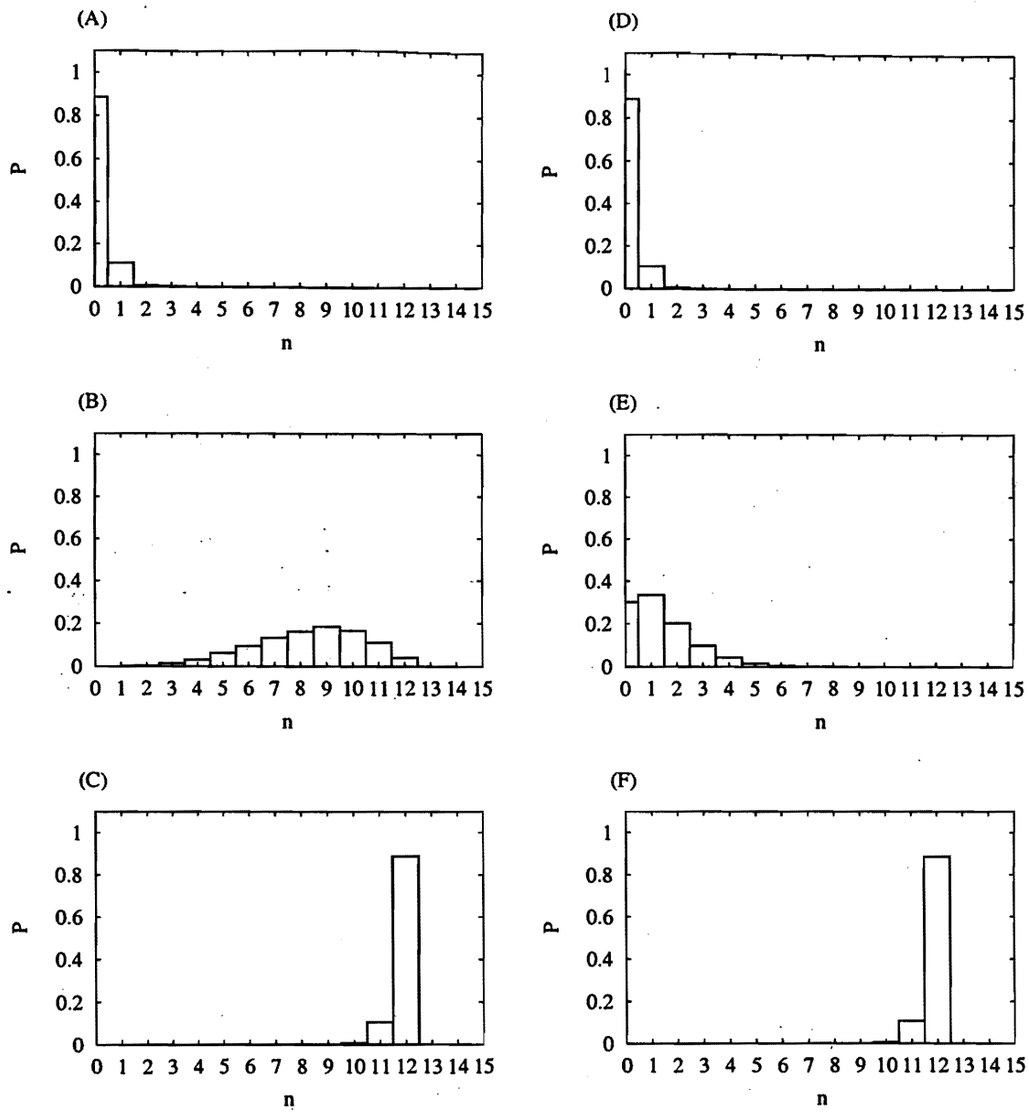
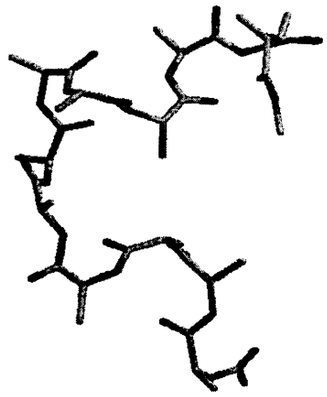
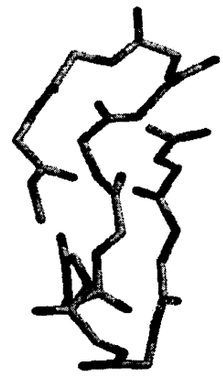


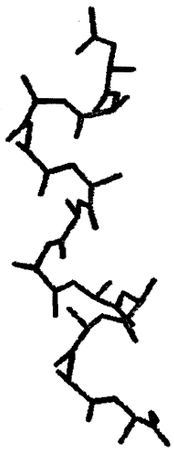
Figure 6: G. La Penna, "Designing generalized statistical..."



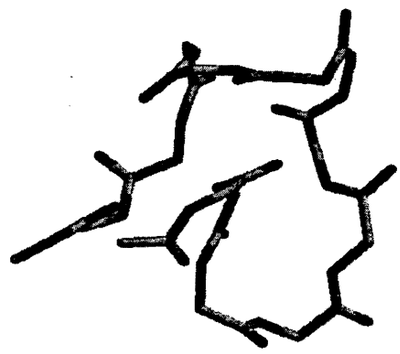
(A)



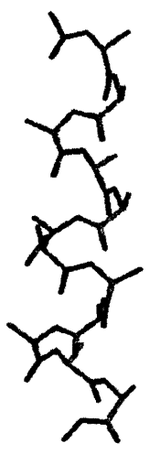
(D)



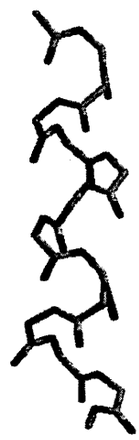
(B)



(E)



(C)



(F)

Figure 7: G. La Penna, "Designing generalized statistical..."

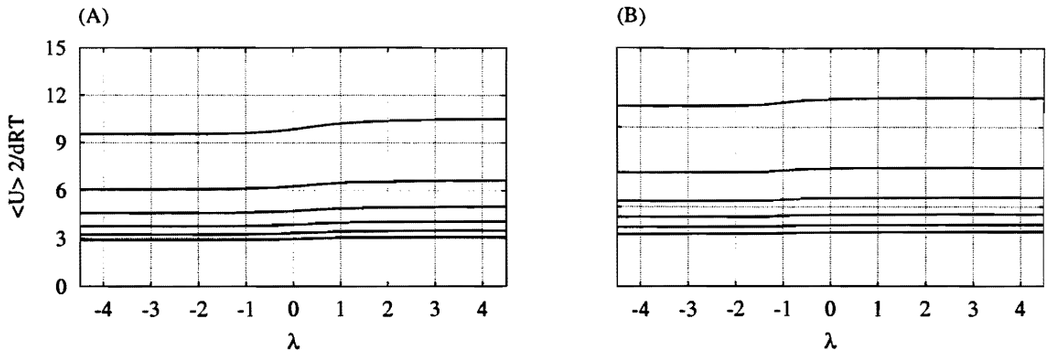


Figure 8: G. La Penna, "Designing generalized statistical..."

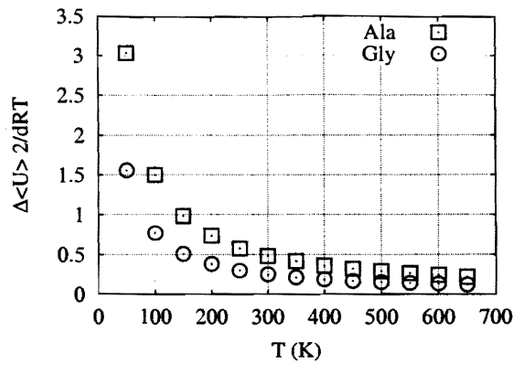


Figure 9: G. La Penna, "Designing generalized statistical..."

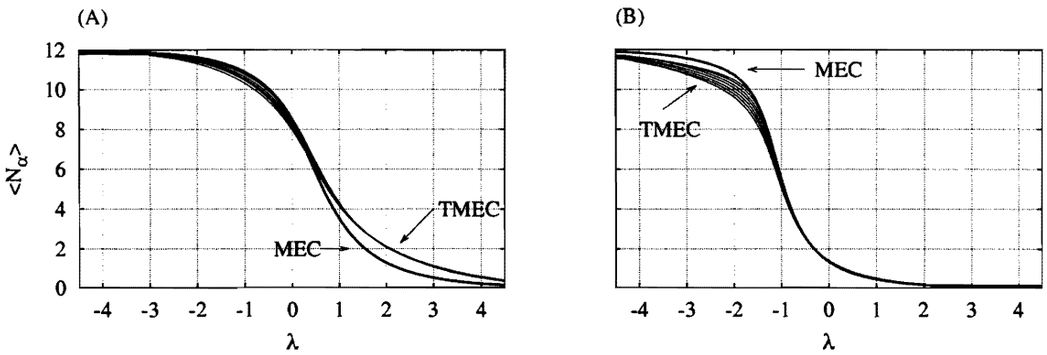


Figure 10: G. La Penna, "Designing generalized statistical..."