From polymers to proteins: effect of side chains and broken symmetry in the formation of secondary structures within a Wang-Landau approach

Tatjana Škrbić*
Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca' Foscari di Venezia, Campus Scientifico, Edificio Alfa, via Torino 155,30170 Venezia Mestre, Italy

Artem Badasyan†
Material Research Laboratory, University of Nova Gorica, SI-5270 Ajdovscina, Slovenia

Trinh Xuan Hoang‡
Center for Computational Physics Institute of Physics, Vietnam Academy of Science and Technology, 10 Dao Tan St., Hanoi, Vietnam

Rudolf Podgornik§
Department of Theoretical Physics, J. Stefan Institute and Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana - SI-1000 Ljubljana, Slovenia

Achille Giacometti¶
Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca' Foscari di Venezia, Campus Scientifico, Edificio Alfa, via Torino 155,30170 Venezia Mestre, Italy

(Dated: January 5, 2016)

We study the equilibrium properties of a flexible homopolymers where consecutive monomers are represented by impenetrable hard spherical beads tangential to each other, and non-consecutive monomers interact via a square-well potential. To this aim, we use both replica exchange canonical simulations and micro-canonical Wang-Landau techniques. We perform a close comparative analysis of the corresponding results, and find perfect agreement between the two methods as well as the past results. The model is then refined in two different directions. By allowing partial overlapping between consecutive beads, we break the spherical symmetry and thus provide a severe constraint on the possible conformations of the chain. This leads to a single helix or a double helix ground state, depending on the range of the interactions. Alternatively, we introduce additional spherical beads at specific positions to represent the steric hindrance of the side chains in real proteins. The ground state in this case is found to be again a globule, thus showing that contrary to some recent suggestions, side chains alone are not sufficient to insure a protein-like structure. Finally, we show that a combination of the above two effects increases the stability of the obtaining secondary structures, in agreement with previous results. The fundamental role played by the range of the square-well attraction is highlighted, and it is shown that they play a role similar to that found in simple liquids and polymers. Perspectives in terms of protein folding are finally discussed.

* tatjana.skrbic@unive.it
† artem.badasyan@ung.si
‡ hoang@iop.vast.ac.vn
§ rudolf.podgornik@ijs.si
¶ achille.giacometti@unive.it

I. INTRODUCTION

Square-well (SW) potential has a long and venerable tradition in simple liquids [1], and has become a paradigmatic test-bench for more sophisticate new approaches. In early attempts of numerical simulations of liquids [2], it was initially used as a minimal model, alternative to Lennard-Jones potential, because it could be more easily implemented in a simulation code, and yet contained the salient features of pair potential for a liquid. It displays both a gas-liquid and liquid-solid transitions in the phase diagram, with results often quantitatively in agreement with real atomistic fluids [3]. For sufficiently short-range attraction, the gas-liquid transition becomes metastable and gets pre-empted by a direct gas-solid transition [4, 5]. Several variants of the SW model have also been proposed over the years in the framework of molecular fluids [6] and colloidal suspensions [7].

In the framework of polymer theory, the model is relatively less known, but it has experienced a renewed interest in the last two decades because it exhibits a reasonable compromise between realism and simplicity [8–12]. In this model, the polymer is formed by a monomer sequence of impenetrable hard-spheres, so that consecutive monomers are tangent to one another, and non-consecutive monomers additionally interact via a square-well interaction. The model can then be seen in a perspective as a variation of the usual freely-jointed-chain [13], with the additional inclusion of a short-range attraction and excluded volume interactions between different parts of the chain.

In spite of its simplicity, this model displays a surprisingly rich phase behavior, including a coil - globule and a globule - crystal transitions, that are the strict analog of the gas-liquid and liquid-solid transitions, respectively. Interestingly, a direct coil-dense globule transition is found for sufficiently short range attraction, pushing the analogy with the direct gas-solid freezing transition even further [9, 11].

The SW polymer model, henceforth referred to as model P (Polymer), can in principle be refined to mimic the folding of a protein, rather than the collapse of a polymer. However there are some crucial differences between synthetic polymers and proteins that should be taken into account. The first difference stems from the specificity of each amino acid monomer forming the polypeptide chain, that provides a selectivity in the intra-bonding arrangement. As amino acids are often classified according their polarities, one simple way of accounting for this effect at the minimal level is given by the so-called HP model [14], where the selectivity is enforced by partitioning the monomers in two classes, having *hydrophobic* and *polar* characters. Under the action of a bad solvent and/or low temperatures, the hydrophobic monomers will tend to get buried inside the core of the globule, in order to prevent contact with the aqueous solvent. The HP model has been shown to be very effective in on-lattice studies [15–17], to describe the folding process at least at a qualitative level. An alternative possible route that has been recently explored is the use of patchy interactions [18].

Another crucial difference, that will feature as a focus of the present study, stems from the observation that a chain, composed only of spherical beads backbone, is unable to capture the inherent anisotropy induced by the presence of side chains [19–21]. As we shall see, this difference can be accounted for within a refinement of the P model in different ways. In real proteins, side chains are typically directed parallel to the outward normal [22, 23]. Upon folding, they have to be positioned in a way to avoid steric overlap, a condition met by both alpha-helices and beta-sheets in real proteins. One possible refinement of a polymer model therefore hinges on the inclusion of additional spherical beads that would mimick this effect. We call this model as Polymer with Side Chains (PSC) model. The presence of side chains in the direction of the outward normal from the backbone additionally breaks the spherical symmetry of the original P model in favour of a cylindrical symmetry. This can be accounted for by allowing partial interpenetration of consecutive spherical backbone beads. This model will be referred to as Overlapping Polymer (OP), and will be the third model that we will consider in addition to the P and the PSC models. A final possibility is clearly given by combining the presence of side chains and the overlapping backbone spheres. This ultimate model will be denoted as Overlapping Polymer with Side Chains (OPSC) model.

In real proteins, all these effects drastically reduce the huge degeneracy of the polymer collapsed state, by progressively reducing the corresponding glassy nature of the original free energy landscape [24]. As a result, one eventually obtains a unique native state, rather than a multitude of local minima having comparable energies. Variants of the P model have also been used in the framework of Go-like models [25–28] routinely adopted in protein folding studies, where the amino acid specificities are enforced by including the native contact list into the simulation scheme.

One of the main difficulties involved in numerical simulations of long polymer chains, stems from the very large computational effort necessary to investigate its equilibrium properties. This is true both for using conventional canonical techniques [29, 30], as well as for more recently developed micro-canonical approaches, such as the Wang-Landau method [31]. Even in the simple P model, while high temperature behavior poses little difficulties, low-temperature/low-energy regions are much more problematic. With the canonical ensemble simulations the system frequently gets trapped into metastable states at low temperatures, and with the Wang-Landau method the low temperature results strongly depend on the ground state energy definition, as lower values of the ground state energy require increasingly larger computational efforts. It is then of paramount importance to discuss a correct implementation of such approaches within the framework of protein-like system and to make a critical assessment of

the reliability of the corresponding results.

The present paper presents a first step in this direction in the framework of the four proposed models (P,OP,PSC,OPSC). Using both replica exchange canonical Monte Carlo simulations [30] and Wang-Landau micro-canonical technique [31], we discuss results stemming from the four models that provide a link between conventional synthetic polymers and protein-like systems. In the P model, the comparative analysis with previous results both on the Wang-Landau and the canonical approach levels allow us to identify the optimal conditions of applicability of both methods. We then proceed to study a step by step generalization introduced by the OP, PSC and the combined OPSC models, using Wang-Landau model and comparing with previous results, when available, obtained using the replica exchange canonical approach. In all cases, we obtain the complete phase diagram in the temperature-interaction range plane. Interestingly,the phase diagrams in the relevant range show a single coil-helix or a double coil-hybrid-helix transitions, in close analogy with the single coil-crystal or double coil-globule-crystal transitions in the polymer model.

As a by-product of our analysis, we will re-obtain some of the results of Ref. [19], with particular emphasis on the role of the interaction range that is argued to be crucial for the phase diagram. In all the above cases, we will use condition realistic for protein systems.

The paper is organized as follows. Section II describes the four different models used in the present study, followed by Section III describing the connection to thermodynamics in both the microcanonical and canonical representations. Section IV illustrated the computational Monte Carlo techniques, and Section V the corresponding obtained results. The paper will be closed by Section VII reporting some concluding remarks and perspectives.

II. FROM POLYMERS TO PROTEINS: FOUR DIFFERENT MODELS

Following the standard approach [11, 12], we model a polymer as a flexible chain formed by a sequence of N monomers, located at positions $\{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$, each having diameter σ. Consecutive monomers are connected by a tethering potential keeping the $N-1$ consecutive monomers at a fixed bond length equal to b. Non-consecutive monomers interact *via* a square-well (SW) potential

$$\phi(r) = \begin{cases} +\infty, & r < \sigma \\ -\epsilon, & \sigma < r < R_c \equiv \lambda b \\ 0, & r > \lambda b \end{cases} \tag{1}$$

where $r_{ij} = |\mathbf{r}_{ij}| = |\mathbf{r}_j - \mathbf{r}_i|$, and $\lambda - 1$ is the well width in units of b and defines the range of interaction $R_c = \lambda b$. Here ϵ defines the well depth and thus sets the energy scale (see Fig.1). The model has a discrete spectrum given by $E_n = -\epsilon n$, where n is the number of SW overlaps [12] that, in turn, depends upon λ.

In addition to the N backbone beads, we will also consider $(N-2)$ sidechain beads of hardcore diameter σ_s. For each of non-terminal backbone beads, one defines a tangent and a normal vector as

$$\widehat{\mathbf{T}}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_{i-1}}{|\mathbf{r}_{i+1} - \mathbf{r}_{i-1}|} \tag{2}$$

$$\widehat{\mathbf{N}}_i = \frac{\mathbf{r}_{i+1} - 2\mathbf{r}_i + \mathbf{r}_{i-1}}{|\mathbf{r}_{i+1} - 2\mathbf{r}_i + \mathbf{r}_{i-1}|} \tag{3}$$

where $i = 2, \ldots, N-1$. Consequently, one can also define a binormal vector as:

$$\widehat{\mathbf{B}}_i = \widehat{\mathbf{T}}_i \times \widehat{\mathbf{N}}_i. \tag{4}$$

Note that $\{\widehat{\mathbf{B}}_i, \widehat{\mathbf{T}}_i, \widehat{\mathbf{N}}_i\}$, $(i = 2, \ldots, N-1)$ are the discretized version of the Frenet-Serret local coordinate frame description [32] that is frequently used in the continuum approach of the polymer theory [33–35].

To each non-terminal backbone bead is attached a sidechain bead in the anti-normal direction with the position given by:

$$\mathbf{r}_i^{(2)} = \mathbf{r}_i - \mathbf{n}_i(\sigma + \sigma_s)/2. \tag{5}$$

The potentials involving sidechain beads are just hardcore repulsions that vanish when $\sigma_s \to 0$.

We will then consider four different cases, schematically illustrated in Fig. 2. In the simplest case for the Polymer model denoted as (P) in Fig. 2, $b/\sigma = 1$ and there are no sidechains ($\sigma_s = 0$). In the Overlapping Polymer ((OP) model in Fig. 2), consecutive backbone beads are allowed to overlap ($b/\sigma < 1$) but there are still no sidechains ($\sigma_s = 0$). The third intermediate case, that we dubbed Polymer with Side Chains (PSC) model, $b/\sigma = 1$ contains

sidechains ($\sigma_s > 0$). Finally, in the Overlapping Polymer with Side Chains (OPSC) model, backbone beads are allowed to overlap ($b/\sigma < 1$) and sidechains are present ($\sigma_s = 0$).

Henceforth, we will assume b and ϵ as the units of lengths and energies respectively, and investigate different ranges of the backbone bead sizes σ/b and the sidechain bead sizes σ_s/b, as well as the effect of the interaction range λ, on the phase diagram of the system. Eventually, connections will be established with realistic values in the case of the proteins.

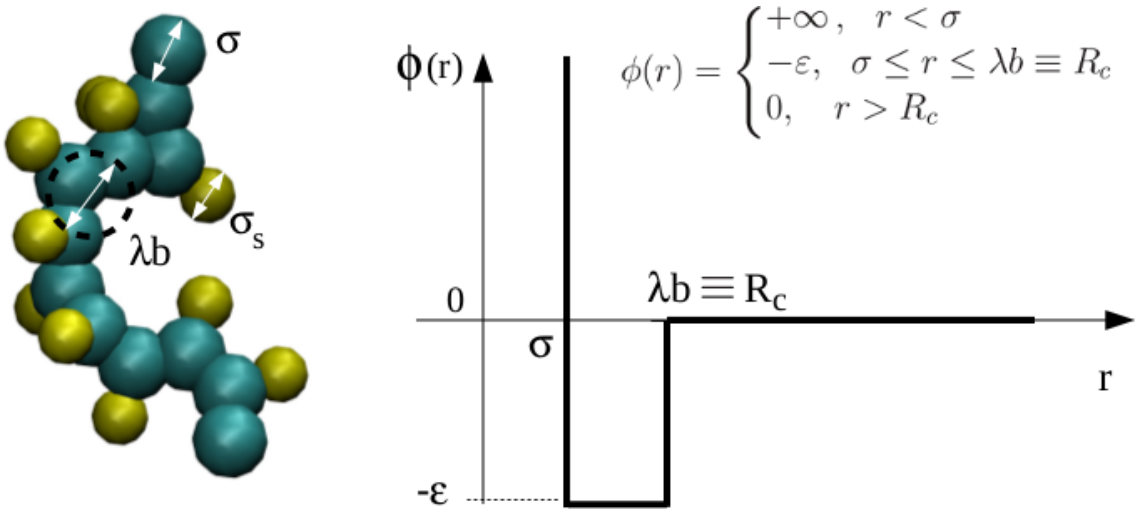

FIG. 1. The square-well potential used in the models

III. THERMODYNAMICS IN THE WANG-LANDAU APPROACH

In the micro-canonical approach, the central role is played by the density of states (DOS) $g(E)$ that is related to the micro-canonical entropy as

$$S(E) = k_B \ln g(E), \tag{6}$$

(k_B is the Boltzmann constant) and thence to the whole thermodynamics. Canonical averages can be also computed using the partition function

$$Z(T) = \sum_E g(E)\, e^{-E/(k_B T)}, \tag{7}$$

and the probability function

$$P(E,T) = \frac{1}{Z(T)} g(E)\, e^{-E/(k_B T)}, \tag{8}$$

to find the system in a conformation with the energy E and temperature T. Helmholtz and internal energies can then be obtained as

$$F(T) = -k_B T \ln Z(T) \tag{9}$$

$$U(T) = \langle E \rangle = \sum_E E P(E,T) \tag{10}$$

where the average $\langle \ldots \rangle$ is over the probability (8).

As we shall see, two important probes of the properties of polymers and the transitions between different ordered states are given by the heat capacity

$$C(T) = \frac{dU(T)}{dT} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2} \tag{11}$$

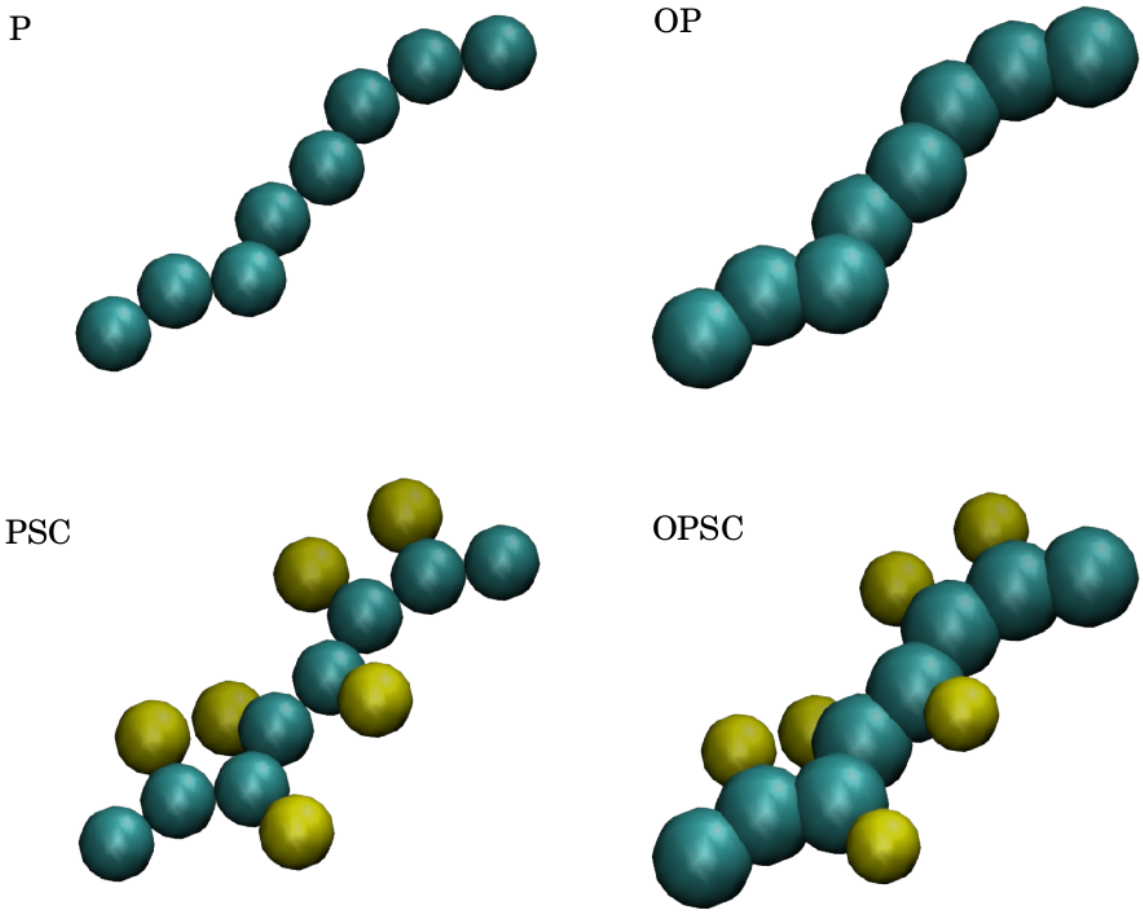

FIG. 2. The four different models considered in the present study. (P) has tangent backbone beads $b/\sigma = 1$ and no sidechains $\sigma_s = 0$. (OP) has no sidechains but its backbone is formed by overlapping beads $b/\sigma < 1$. (PSC) has sidechains $\sigma_s > 0$ but no overlap $b/\sigma = 1$. (OPSC) has sidechains $\sigma_s > 0$ and overlapping beads in the backbone $b/\sigma < 1$.

and by an averaged value of the radius of gyration [36] $R_g^2 = \sum_{i<j}^{N} (\mathbf{r}_i - \mathbf{r}_j)^2 / N^2$. This can obtained in microcanonical ensemble by constructing the distribution $P(Rg^2, E)$ of R_g at a given E to get the micro-canonical average $\langle \ldots \rangle_E$ over this distribution

$$\left\langle R_g^2 \right\rangle_E = \sum_{R_g} P\left(R_g^2, E\right) R_g^2 \tag{12}$$

as well as in the canonical ensemble using

$$\left\langle R_g^2(T) \right\rangle = \sum_{E} \left\langle R_g^2 \right\rangle_E g\left(E\right) e^{-E/(k_B T)}. \tag{13}$$

It is important to remark that the derivation of thermodynamics from the $g(E)$ is quite general, and does not depend on the specific method of simulation. The DOS is therefore the optimal tool to compare different methods and assess the pros and cons of each of them, which is also one of the aims of the present work.

IV. MONTE CARLO SIMULATIONS

In this section, we briefly recall the main points of the Wang-Landau technique as applied to polymers, as well as the main differences with its canonical counterpart, with the aim of emphasizing the strengths and weaknesses in a comparative perspective.

A. Micro-canonical approach: Wang-Landau method

Following general established computational protocols [29, 30], and refs. [11, 12] for the specificities related to the polymers, we use Wang-Landau (WL) method [31] to sample polymer conformations according to micro-canonical distribution, by generating a sequence of chain conformations $a \to b$, and accepting new configuration b with the micro-canonical acceptance probability

$$P_{acc}(a \to b) = \min\left(1, \frac{w_b g(E_a)}{w_a g(E_b)}\right), \tag{14}$$

where w_a and w_b are weight factors ensuring the microscopic reversibility of the moves.

A sequence of chain conformations is generated using a set of Monte Carlo moves, which are accepted or rejected according to Eq. 14. The set of Monte Carlo moves employed consists of a local-type moves such as single-bead crankshaft moves, reptation moves and end-point moves; as well as of a non-local-type moves among which pivot moves, bond-bridging moves and back-bite moves. All move types are chosen at random and one sweep of the Wang-Landau algorithm is considered completed as soon as at least N monomer beads have been displaced.

The bond-bridging moves and back-bite types of the moves are found to be important in order to sample correctly and efficiently the low-energy states. Both of them consist of choosing one chain end (bead 1 or N, old end bead in the following) and attempt to connect it to an interior bead i randomly chosen in its neighbourhood within 2σ range. In the bond-bridging version of the move [12, 37] this attempt is realised by removal of the bead next to the chosen one in the direction of the chosen end (bead $i-1$ or $i+1$) and its re-insertion between the bead i and the old-end bead via a crankshaft type of the move. The factor that in this case ensures the microscopic reversibility of the move reads $w_b/w_a = n_a\, r_b/n_b\, r_a$, where n_a and n_b are the numbers of neighbours of the old and the ne chain and within the 2σ range, respectively; while the r_a and r_b are the distances between (old and new) chain end and the bead i. In the back-bite version of the move [38] the old chain end (1 or N) is attempts to be directly connected with the randomly selected bead i in its neighbourhood and the correction of the acceptance probability in this case is simply $w_b/w_a = n_a/n_b$. The weight factors for other types of Monte Carlo moves are all equal to unity.

B. Canonical approach: replica exchange

The parallel tempering (or replica exchange) technique [39, 40] is a powerful method for sampling in systems with rugged energy landscape. It allows the system to rapidly equilibrate and artificially cross energy barriers at low temperatures. Furthermore, the method can be easily implemented on a parallel computer. Parallel tempering technique can be used with both Monte Carlo and Molecular Dynamics simulations, but in this study, we apply this technique specifically to Monte Carlo simulation. The method entails monitoring M canonical simulations in parallel at M different temperatures, T_i, $i = 1, 2, \ldots, M$. Each simulation corresponds to a replica, or a copy of the system in thermal equilibrium. In individual Monte Carlo simulations, new moves are accepted with standard acceptance probabilities given by the Metropolis method [41]. The replica exchange technique allows the swapping of replicas at different temperatures without affecting the equilibrium condition at each temperature. Specifically, for two replicas, Γ_i being at T_i and Γ_j being at T_j, the swap move leads to a new state, in which Γ_i is at T_j and Γ_j is at T_i. The acceptance probability of such a move can be derived based on the detailed balance and is given by:

$$P_{\text{swap}} = \min\left(1, \exp\left[\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)(E_i - E_j)\right]\right). \tag{15}$$

The choice of replicas to perform an exchange can be arbitrary, but for a pair of temperatures, for which replicas are exchanged, the number of swap move trails must be large enough to ensure good statistics. The efficiency of a parallel tempering scheme depends on the number of replicas, the set of temperatures to run the simulations, how frequent the swap moves are attempted, and is still a matter of debate. It has been suggested that for the best performance, the acceptance rate of swap moves must be about 20% [42].

In our parallel tempering scheme we consider 20 replicas and the temperatures are chosen such that they decrease geometrically: $T_{i+1} = \alpha T_i$, where $\alpha = 0.8$. The highest reduced temperature, $k_B T_1/\epsilon = 10$, at which the polymer is well poised in the swollen phase. We allow replica exchange only between neighboring temperatures and for each replica a swap move is attempted every 50 Monte Carlo steps. Standard pivot and crank-shaft move sets [43] are used in Monte Carlo simulations. A typical length of the simulations is 10^9 steps per replica. Results from parallel tempering simulations are the equilibrium data and are convenient for analysis using the weighted histogram analysis method [44]. The latter allows one to estimate the density of states as well as to calculate the thermodynamic averages from simulation data at various equilibrium conditions.

V. RESULTS

A. Model P

Before tackling the more complex systems, we will test our WL and replica exchange approach using the simpler, and yet interesting P model ($\sigma/b = 1$ and $\sigma_s = 0$) for which several studies [11, 12] have been previously performed and can be contrasted with.

The aim of the calculation is the computation of the DOS $g(E)$. In the WL method, $g(E)$ is constructed iteratively, with smaller scale refinements made at each level of iteration, controlled by the flatness of energy histogram. We typically consider an iteration to reach convergence after 30 levels of iteration, corresponding to a multiplicative factor values of $f = 10^{-9}$. This choice is neither unique, nor universally accepted [15, 45–47], but have checked this to be sufficient for our purposes by comparing results obtained with different values. An additional crucial step in WL algorithm hinges in the selection of ground state energy. As this must be defined at the outset, and is known to drastically affect the low-energy behavior of the system [15, 17], care must be exercised in its selection. At the present time, however, there is no universally accepted procedure for off-lattice Wang-Landau simulations, and in the present paper we will be following the procedure suggested in Refs. [11, 12], that has been reported to be reliable in most of the cases. A preliminary run with no low-energy cutoff is carried out for a sufficient number of MC steps ($10^8 N$ in our case). This provides an estimate of the minimal ground state energy.

We have explicitly performed this procedure for chains ranging from $N = 4$ to $N = 128$, as depicted in Fig. 3 where the reduced minimum energy per monomer $-E_{min}/(N\epsilon)$ is plotted against N. The results are in perfect agreement with those obtained by our replica exchange canonical method and with those performed in the extensive simulations by Taylor *et al* [11, 12].

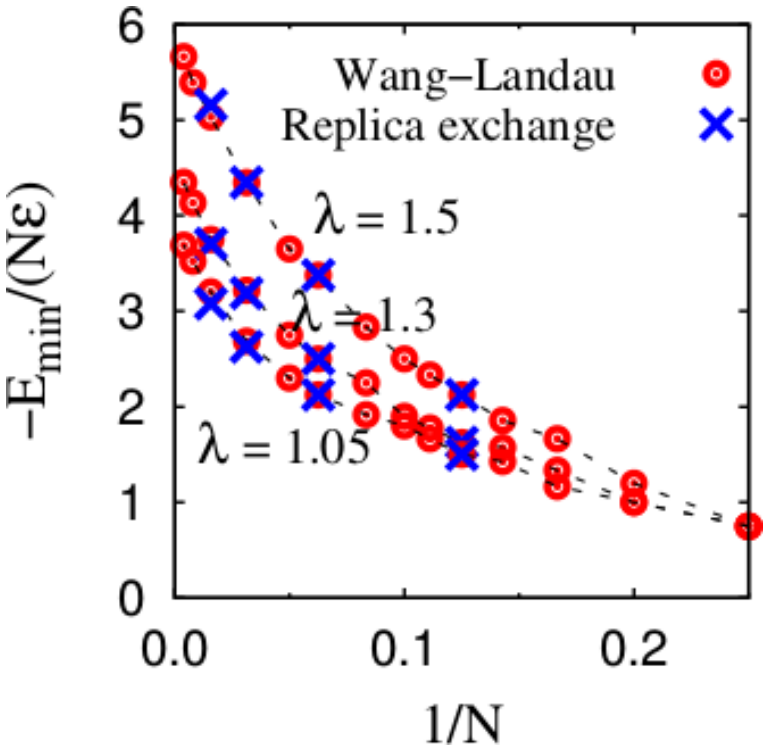

FIG. 3. Plot of the reduced minimum energy per monomer $-E_{min}/(N\epsilon)$ as a function of $1/N$ for $\lambda = 1.5$, using both canonical (crosses) and Wang-Landau (circles) approaches.

As an additional preliminary step, we have tested our code against exact analytical results valid for small $N = 4$ (tetramers) and $N = 5$ (pentamers) [9, 10, 48] at different vaues of the interaction range λ, finding again perfect agreement.

Further support stems from a comparison with the Molecular Dynamics (MD) and canonical MonteCarlo (MC) results by Zhou and Karplus [8], as seen in Fig. 4 (top) for $8 \leq N \leq 64$, and can be contrasted against Figure 3 of Ref. [8], again with extremely good agreement. Note that, in addition to the well-expressed peak at low temperatures, our findings indicate the presence of a less distinct peak at higher temperatures. It is also worth

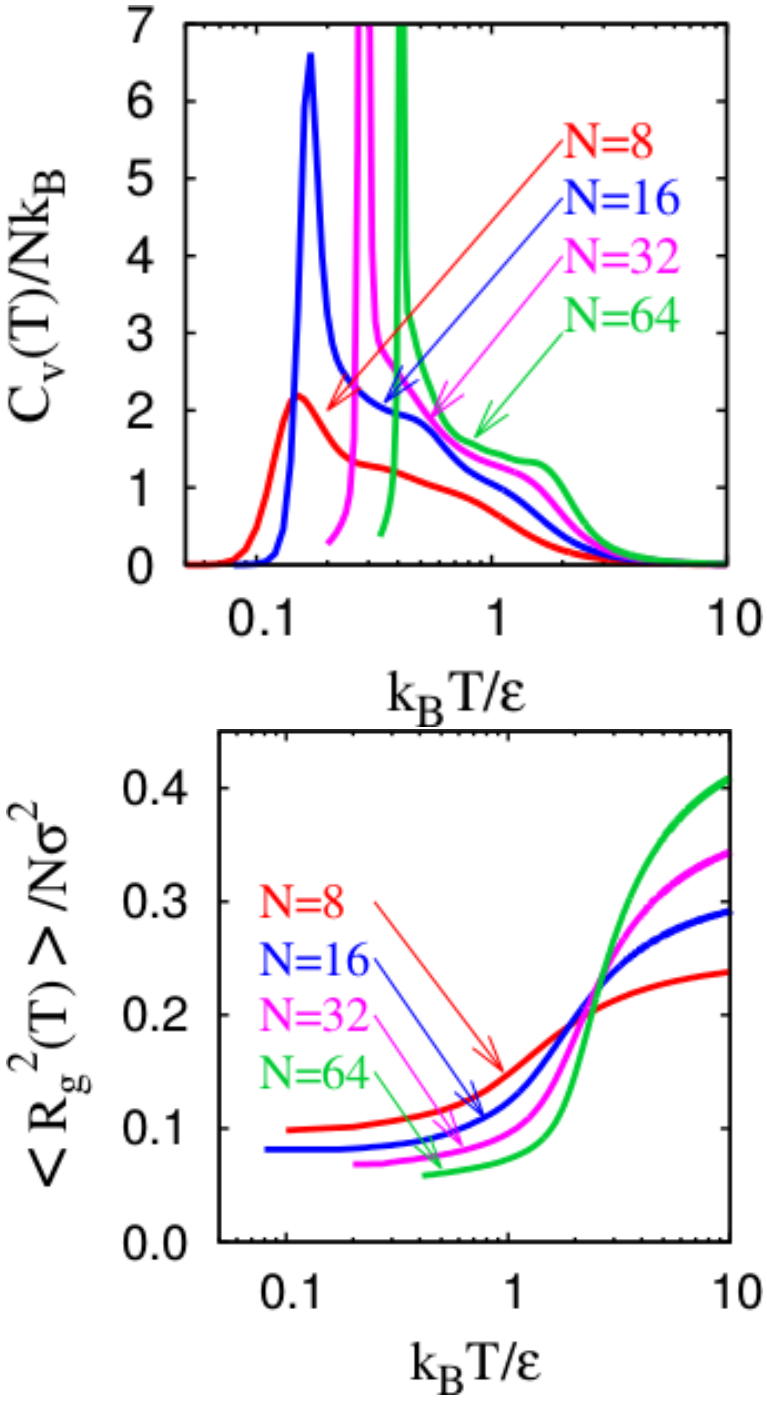

FIG. 4. Reduced heat capacity per monomer $C(T)/(Nk_B)$ (top), and the mean square radius of gyration per monomer $\langle R_g(T)\rangle/(N\sigma^2)$ (bottom) for $N = 8, 16, 32, 64$ as a function of k_BT/ϵ, at $\lambda = 1.5$. These findings are in very good agreement with Zhou and Karplus (see Figures 1 and 3 in Ref. [8]).

stressing that our calculation is able to probe lower temperatures and clearly shows the peak that remained blurred in Ref.[8]. This is because, results from Ref.[8] are affected by very large errors at low temperatures (as commonly found in canonical simulations) even for moderate chain lengths. For instance, the case $N = 16$ shows the onset of a large error below $k_BT/\epsilon = 0.3$, so lower temperatures are in fact cut out (see again Figure 3 in Ref.[8]). Conversely, in the WL calculation, low and high temperatures are equally well sampled and because of this, we have managed to demonstrate the appearance of a peak in the heat capacity also at low temperatures, indicating a possible structural transition (see Fig. 4).

WL approach is therefore extremely useful in this respect. Similarly good agreement can be obtained for the radius of gyration, that is usually regarded as another good probe for identifying structural changes. This is reported again in Fig. 4 (bottom) and can be contrasted with results from Figure 1 in Ref.[8], where the mean square radius of gyration per monomer $\langle R_g^2\rangle/(N\sigma^2)$ is plotted as a function of the reduced temperature k_BT/ϵ for the same chain lengths as above. It is worth noticing that each polymer significantly shrinks in the region of temperatures between 1.0 and 3.0, in the same interval where a very weak high-temperature peak/plateau is observed in the heat capacity curve (see Fig. 4). No noticeable changes occur on further cooling down, as opposed to the heat capacity, that shows additional, well pronounced peaks at low temperatures. This can be probably traced back to the fact that the second transition is expected to be related to a globule-to-crystal phase where no appreaciable size change is usually expected.

B. Model OPSC

Having discussed the P model as a suitable model for polymers, we now add two additional ingredients in the strive of modelling a protein. While the P model is perfectly able to describe the correct low temperature behaviour of polymers, it cannot be used for proteins not even at the minimal level of description. There are a number of reasons for that. Previous results [11, 12] briefly recalled in Section V A, unambiguously showed that a model of a chain formed by tethered spherical beads can only have a high temperature coil phase, an intermediate globular phase – provided the range of interactions to be sufficiently long – and a low temperature crystal phase. Due to its inherent isotropy, the P model cannot capture the essential ingredients for the formation of secondary structures in proteins.

One possibility to cope with this is to replace beads with disks, thus breaking the spherical symmetry. In a continuum description, this gives rise to the so-called tube (or thick polymer) model [20, 21]. This effect alone was shown to give rise to formation of helices and planar conformations [33]. A similar effect can be achieved by allowing partial overlapping of consecutive spherical monomers, and this gives rise to model OP considered here. In the OP model consecutive spheres overlap ($\sigma/b < 1$) but sidechains are absent ($\sigma_s = 0$). Another ingredient that is known to play a fundamental role in the low temperature behavior of proteins is the steric hindrance of the side chains. This can be seen, for instance, from the Ramachandran plots in real proteins showing that both alpha helices and beta sheets are constrained into well defined regions of the dihedral angles [24]. One can then envisage the possibility of introducing additional spherical beads, attached to each backbone bead at the appropriate location and representing the van der Waals spheres of the side chains. This leads to the PSC model that is complementary to the OP model in the sense that consecutive spheres do not overlap ($\sigma/b = 1$) but sidechains are present ($\sigma_s > 0$). Finally, one can consider both effects, thus obtaining the OPSC model. Note that both backbone and side chain spheres are considered aspecific, unlike in real proteins where both have their own specific character. The effect of this ingredient has been considered in the past and therefore will be neglected for simplicity in the present work, although it could be in principle accounted for without any difficulties.

The OPSC model was already considered in Ref. [19] using the replica-exchange canonical simulations, both for short ($N = 16$) and long ($N = 50$) chains. The necessary effort for reaching the correct ground state was considerable for the longer chain case, due to the appearance of an increasing number of metastable configurations that might hamper the correct low temperature sampling. In the present study, we will rederive all results obtained in Ref.[19] by a Wang-Landau method, plus additional results that will allow us to put these models into a wider perspective framework. More specifically, we will highlight the fundamental role of the range of interaction on the formation of the secondary structures. In OPSC consecutive spheres overlap ($\sigma/b < 1$) and sidechains are present ($\sigma_s > 0$). As in Ref.[19], we will use $N = 16, 50$, $\sigma = 3$Å and $\sigma_s = 2.5$Å and $b = 3.8$Å. The range of interactions here is $R_c = 7.5$Å. The WL calculation was performed by using different ground state energies from $E_{gs}/\epsilon = -55$ to $E_{gs}/\epsilon = -60$, and Fig. 5 reports the specific heat per monomer $C_V/(Nk_BT)$. Here $N = 16$ as in Ref. [19]. While the first peak at higher temperatures, associated with the coil-to-globule transition, is unaffected by the actual value of the ground state, the different behavior at low temperatures is clearly visible. When $E_{gs}/\epsilon = -55$ the low temperature second peak is absent, indicating that the ground state is the double helix structure. Note that the double helix is stabilized by consecutive non-local contacts, in the same way as protein's β-sheets, but lacks the planar symmetry of the latter. As the energy decreases toward the correct ground state $E_{gs}/\epsilon = -60$, a second low temperatures peak emerges indicating an additional structural transition to a helical state, that is then the ground state of the system under these conditions. Note that the results for $E_{gs}/\epsilon = -60$ are in full agreement with those obtained in Fig.3a of Ref.[19] *via* replica-exchange canonical simulations.

Even when the actual low energy limit is properly accounted for, the corresponding ground state is strongly dependent upon the chosen range of the interactions R_c. This is not surprising on physical grounds. As $R_c/b \approx 1$, the range of attraction is limited to the nearest-neighbor monomers, and the chain will have the tendency to form a coil configuration at high temperatures and a suitable secondary structure maximizing the number of favorable contacts at low temperatures. In the opposite limit $R_c/b \gg 1$, each monomer can interact with essentially any other monomer

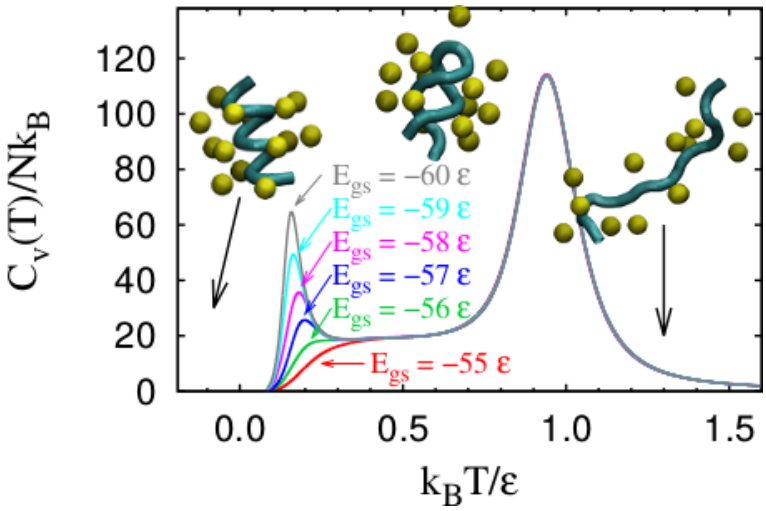

FIG. 5. Specific heat per monomer $C_V/(Nk_B)$ as a function of the reduced temperature k_BT/ϵ in the case $N = 16$ with different ground state energies $-60 \geq E_{gs}/\epsilon \geq -55$ These results for $E_{gs}/\epsilon = -60$ are compatible with Fig.3a of Ref.[19]. The insets include snapshots of representing configurations. Here and below, a tube representation of the backbone is employed for a better visual reproduction.

in the chain, and at low temperature will tend to form a globular structure that that will have the lowest possible energy. On the basis of results given in Fig. 5 for $N = 16$, one could reasonably expect a formation of a stable helix for some intermediate value of the R_c/b ratio. This is confirmed in Fig. 6, where we report the specific heat per monomer $C_V/(Nk_BT)$ as a function of the reduced temperature k_BT/ϵ in the case $N = 16$ (top panel) and $N = 50$ (bottom panel), at different ranges of interactions R_c. In the former case ($N = 16$) a single well-developed peak is clearly visible for $R_c = 6.5$Å, corresponding to the coil-helix transition. As R_c increases, this single peak gradually evolves first into two, around $R_c \approx 7.5$Å, and eventually into three at $R_c \approx 9.0$Å, accounting for an intermediate double helix configuration at $k_BT/\epsilon \approx 0.8$, and for a globular configuration at the lowest k_BT/ϵ. The $N = 16$ result is particularly useful because it allows the presence also of larger values of R_c that typically require long computational time.

We note that this result is in perfect agreement with the results presented in Fig.S2(a) in the supplementary material of Ref.[19] for the same parameters that was obtained by replica-exchange canonical simulations. In both cases the ground state energy was found identical (dependent on the choice of R_c)

The case $N = 50$ allows for a better statistics at the expense of an increased computational effort, see Fig. 6 (bottom panel). A striking feature of Fig. 6 is again the presence of a single strong peak indicating the ground state of the system corresponds to a helical organization, as illustrated in the insert snapshot, along with a secondary peak at slightly higher temperatures. The second peak increases, and in fact becomes the dominant one, upon increasing the interaction range R_c, indicating the progressive stabilization of the double helix phase found at intermediate temperatures. This is in line with previous results for $N = 16$ with the proviso that the globular phase is missing here due to the fact that the computational effort to investigate larger R_c becomes prohibitive above $R_c \gg 9$ Å, a region where this phase is expected. Compatible results (not shown) can be obtained also by considering the gyration radius, in analogy to what done in the P model case.

Guided by the above findings, we can draw the phase diagram in the temperature k_BT/ϵ - interaction range R_c plane. This is reported in Figure 7 in the case $N = 16$ (top panel) and $N = 50$ (bottom panel). Several interesting features are apparent. Starting from a coil high temperature configuration, one finds a direct transition to a helix conformation upon cooling for sufficiently short interaction ranges ($R_c \lesssim 7$ Å). Above this value, there is first a transition to a double helix structure followed by a successive transition to helix structure. The two transitions are expected to merge at large interaction range, where the low temperature ground state is presumed to be a globule, irrespective of R_c. Interestingly, this increases the already rich polymorphism present in the polymer model, with additional structural changes that will be further addressed in Section VI

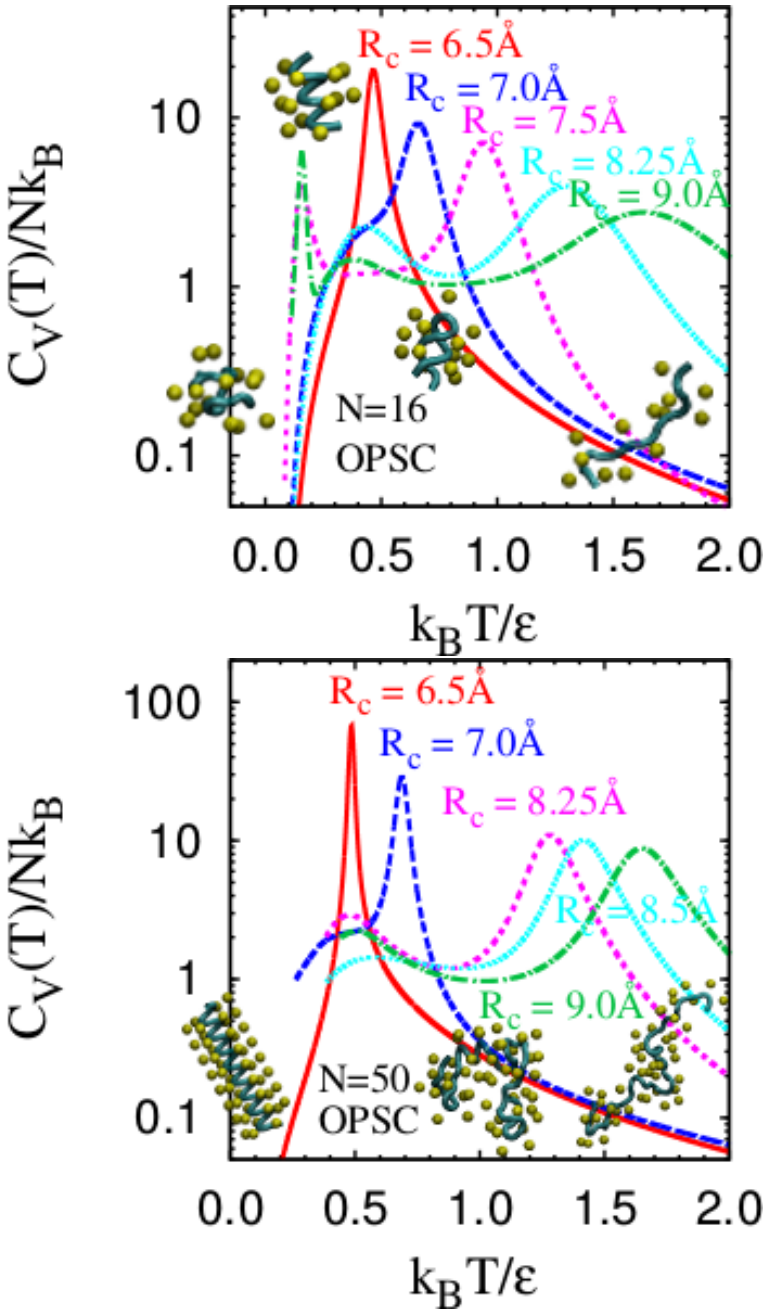

FIG. 6. (Top) Specific heat per monomer $C_V/(Nk_B)$ as a function of the reduced temperature k_BT/ϵ in the case $N = 16$ with different interaction ranges R_c. (Bottom) Same as above in the case $N = 50$. In this case, the results for $R_c = 6.5$Å are compatible with Fig.S2(a)in the supplementary material of Ref.[19].

C. Model OP

In the OP model, consecutive beads overlap, as in the OPSC model, but there the side chains are missing, so that the steric hindrance of the side chain beads is also absent. On the other hand, partial interpenetration of consecutive beads breaks the spherical symmetry of the P model, along the backbone direction. Effectively, this provides a restriction on the number of possible local conformations that the chain can achieve in a way similar to having a collection of disk-like rather than sphere-like monomers. A similar effect can be obtained in the tube model by enforcing a finite thickness of the chain via a three-body potential [21]. In the $N = 16$ case, for high and low temperatures the coil and globule conformations are realized respectively, with intermediate temperature phases typically as the double helical

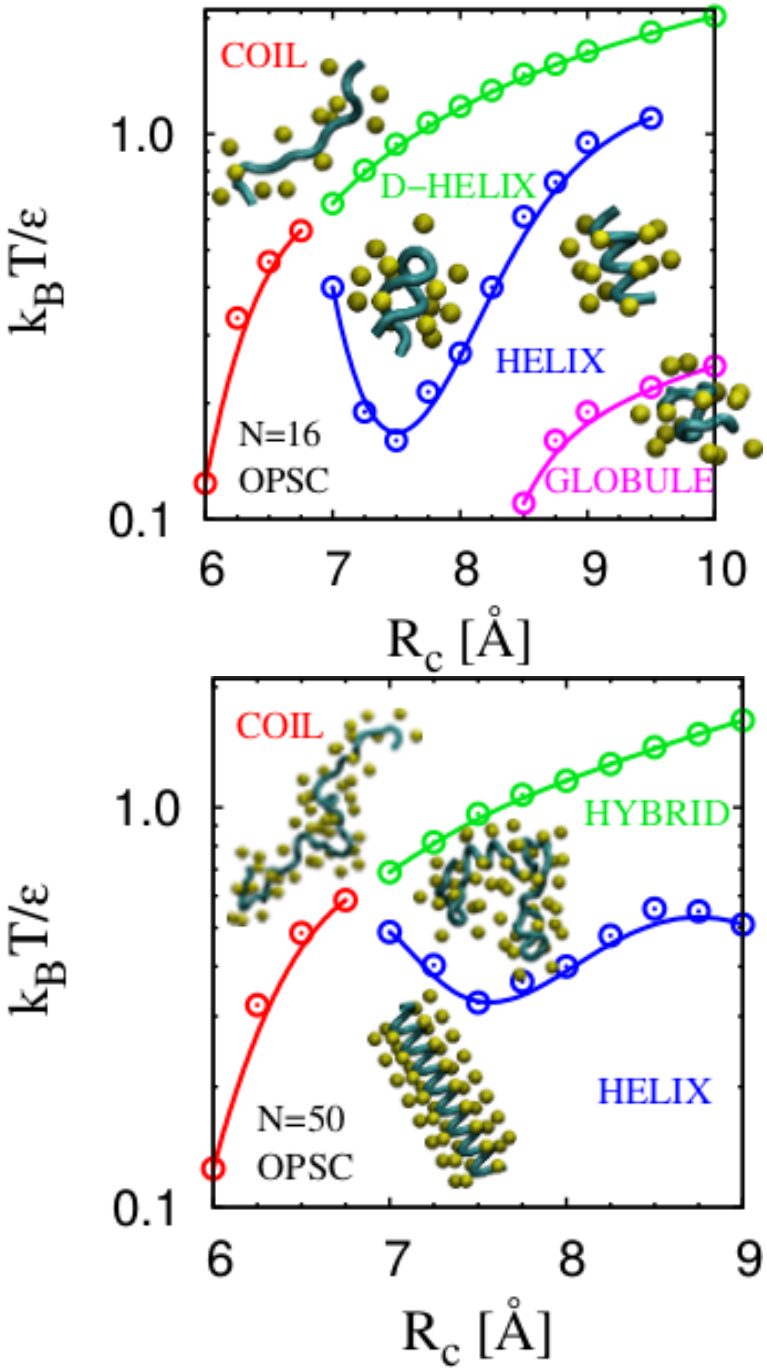

FIG. 7. Phase diagram of the OPSC model in the temperature-interaction range plane.(Top) $N = 16$ case. (Bottom) $N = 50$ case.

conformation at low R_c and a helicoidal conformation at larger R_c. This agrees with findings from the tube model, in spite of the different driving forces in the two systems [33]. It also agrees with the replica exchange simulations by Magee and collaborators [48, 49] who observed helix formation in the OP model for sufficiently high interpenetration with $N = 20$.

However, we find these intermediate double helical and helicoidal configurations to become less and less stable as the length of the polymer increases, so that in the case of $N = 50$ only a stable coil and globular configurations, with some intermediate planar-like structures, are observed. In Figure 8, we display the obtained phase diagram in the case $N = 50$. Interestingly, below $R_c \approx 6$ Å, we expect a direct coil-globule transition.

A final point is in order here. While it is intuitive that the OP model shares some similarities with the tube model

discussed in Ref. [21, 33], since both lead to a breaking of the spherical symmetry with the appearance of a priviledged axis along the chain backbone, we believe that some of the specificities of the tree-body potential involved in the tube model cannot be represented by a two-body potential as discussed here.

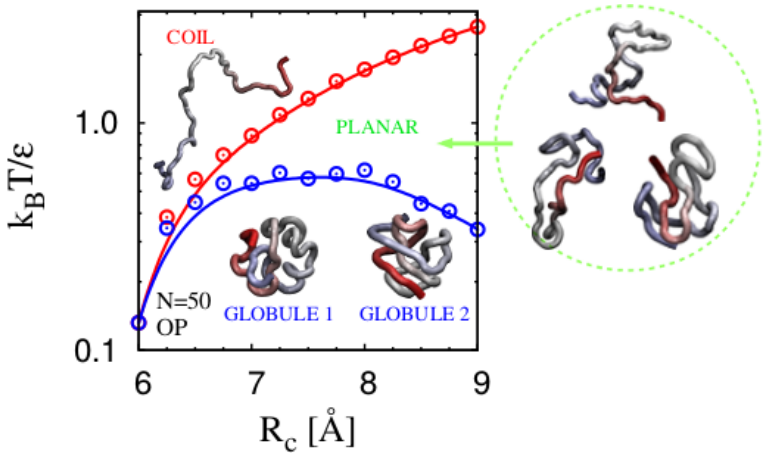

FIG. 8. Phase diagram of the OP model in the temperature-interaction range plane in the $N = 50$ case.

D. Model PSC

The OPSC model becomes the PSC model in the absence of overlapping. This was already partially considered in Ref.[19], where a phase diagram similar to that of the OPSC model was found, although with more unstable secondary structure configurations. Our results support this view. All in all, the punchline is that *both* the interpenetrability *and* the side chains are crucial ingredients to observe stable secondary structures.

VI. CORRELATION FUNCTION AND THE FISHER–WIDOM LINE

The Landau-Peierls theorem [50, 51] forbids the presence of a thermodynamic phase transition in a strictly one dimensional system with only short range interactions, and a similar results holds true for liquid systems [52].

In fluid systems and in the presence of competing attractive/repulsive interactions, such as those occurring here, a Fisher–Widom line (not to be confused with the Widom line [53]) is separating two different regimes in the radial distribution function [54], characterized by an oscillatory (when repulsions are dominating) and an exponentially decaying (when attractions are dominating) behavior. The rationale behind the FW line is that on approaching the critical points where attractions become more and more effective, the behavior of correlation functions must switch from oscillatory (characteristic of repulsive interactions) to exponential with a well defined correlation length. The FW line was also observed in one-dimensional fluid of penetrable spheres [55, 56], where Landau-Peierls theorem is no longer valid. In practice, the FW line is associated with an abrupt discontinuity in the structural behaviour of the system (as signalled by the correlation function) that is not, however, related to any discontinuity in the thermodynamical behaviour.

As shown below, a similar feature occurs in the coil-helix transition line that can then be seen as the FW line within this framework. Consider the tangent-tangent correlation function between two consecutive beads $\langle \widehat{\mathbf{T}}_{i+1} \cdot \widehat{\mathbf{T}}_i \rangle$, where the average is the canonical one. In the coil state, the conformation of the chain is random and hence there is no correlation between the orientations of consecutive backbone beads. As a result, the tangent-tangent correlation is an exponentially decaying function of the sequence separation. In the helix state, however, the tangent-tangent correlation is an oscillating function of the sequence, with periodicity of about 4 units. The transition between the two regimes is an abrupt one, associated with the coil-to-helix transformation. Figure 9 confirms this expectation. Here the tangent-tangent correlation is plotted as a function of the sequence separation at two different reduced temperatures, $k_BT/\epsilon = 1.0$ above the FW line and $k_BT/\epsilon = 0.1$ below it. The transition from an exponential to an oscillating behavior upon cooling below the FW line is thus evident. Note that this transition can be experimentally probed by e.g. circular dichroism, and some straightforward procedures have been devised to interpret the results

of these measurements in terms of simple models [57, 58]. Similar transitions also occur for the normal-normal and binormal-binormal correlations.

Our findings are in perfect agreement with those by Banavar *et al* [19] obtained *via* the replica-exchange method, thus supporting the physical interpretation given there.

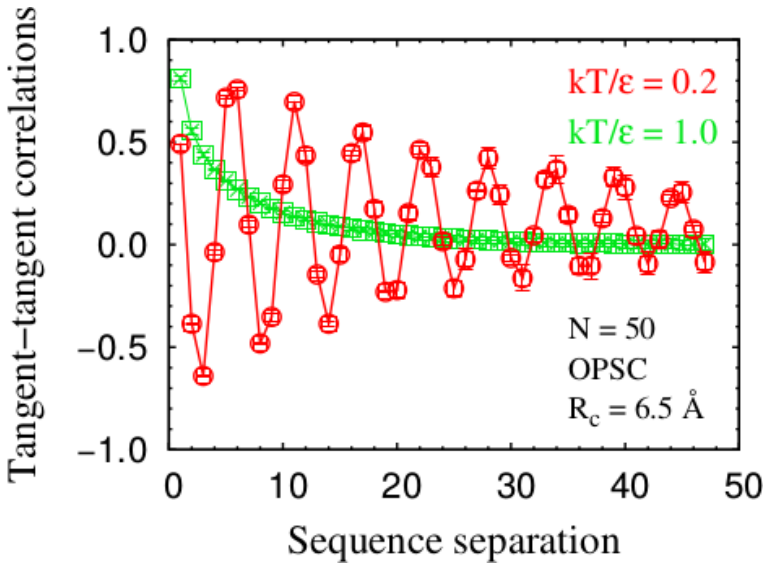

FIG. 9. Plot of the tangent-tangent correlation functions as a function of the sequence separation at two different reduced temperatures k_BT/ϵ above and below the transition line.

VII. CONCLUSIONS

In this paper we have studied the equilibrium statistics of a homopolymer formed by a sequence of tangent identical monomers represented by impenetrable hard spheres. A tethering potential keeps the consecutive backbone beads linked, whereas non-consecutive ones interact via a square-well potential, that then drives the collapse of the chain at sufficiently low temperatures. Three different variants of the above P model were considered as a gradual step toward a more realistic model for proteins. By allowing consecutive backbone spheres to interpenetrate, one breaks the rotational symmetry of the P model, thus opening up the possibility of having additional intermediate transitions. This model, dubbed the OP model, is similar in spirit to the tube model [20, 21], where the symmetry breaking was enforced by replacing spherical with disk-like beads. Another variant, denoted as PSC model, keeps the backbone beads tangent, but allows the presence of additional side-chain beads located at the position of the side chain center. This allows the side chains to play an important role in the formation of secondary structures [59]. The last variant, the OPSC demol, combines and generalizes all of the above effect.

We first reproduced some of the Wang-Landau results obtained by Taylor *et al* [11, 12] for model P, with long chains (with up to $N = 256$ monomers), and re-obtained via replica-exchange canonical approach. This comparison allows a comparative assessment of the pros and cons of each approach, as well as a comparison with previous results obtained in the canonical ensemble [8].

Having done that, we have then used the same Wang-Landau method to tackle OP, PSC, and OPSC models. The latter can be regarded as a minimal model for the formation of secondary structures in proteins and was studied by replica-exchange techniques in Ref. [19]. In particular, we were able to underpin the fundamental role played by the interaction range in selecting the secondary structure. For very short range square-well attractions, the lowest energy is achieved by a double helix, in agreement with our intuition and with previous findings on the tube model [33]. At intermediate interaction range, the lowest energy is associated with the formation of a well-defined helical structure, whereas at interaction ranges much longer of the bead sizes, the ground state is a globule, as it should be in view of the mean-field like character of this interaction. The possibility of tuning a single coil-to-globule or a double coil-to-helix and helix-to-globule transition, can be regarded as the OSCP model counterpart of the results found in the P model by Taylor *et al* [11, 12], where the double transition coil-to-globule and globule-to-crystal transitions found at moderate attractive range, are replaced by a direct coil-to-crystal transition below a critical interaction range. Note that, in

turn, this is effect is similar to what found in square-well fluids, where a double gas-liquid and liquid-solid transitions is preempted by a direct gas-solid transition below a critical range of interactions.

In addition to to the OPSC model, we have also studied the simpler OP and the PSC models, where side chains are absent (in the OP) or the overlapping is absent (in the PSC). We found that, while OP model the polymer displays a phase diagram similar to the full OPSC model, the ground state of the PSC model is a globule. This indicates that the side chains alone are not sufficient ingredients to drive the formation of secondary structures, whereas the overlapping is, a finding consistent with those obtained via the tube model [33] that is considered to be a close relative of the OP model.

Finally, we have also discussed the transition in the behavior of the tangent-tangent correlation functions, as well as the the normal-normal and the binormal-binormal correlation functions, as a function of the temperature. We found that the transition from monotonic exponental behaviour in the separation dependence, found in the coil configuration at high temperatures, is replaced by an oscillatory behaviour at lower temperatures, indicating a helical organization. We proposed that this transition can be regarded as the analogue of the Widom-Fisher line found in fluid systems.

The formation of secondary structures is an important intermediate step for the general process of the folding of a protein. Our results are consistent with previous studies suggesting that the introduction of a cylindrical (as opposed to spherical) symmetry is in fact sufficient to remove the typical glassy structure of the low energy state in polymers, thus highlighting the formation of secondary structures due to the presence of side chains, making those structures more stable. The rigidity of the chan mostly ascribed to the presence of intrachain hydrogen bonds is known to be another important ingredient for protein folding, leading to the general issue on the existence of minimal model for a correct description of protein folding. Work along these lines is currently underway and will be reported in a future study.

ACKNOWLEDGMENTS

We thank Amos Maritan and Jay Banavar for many enlightening discussions on the subject. This work was also supported by MIUR PRIN-COFIN2010-2011 (contract 2010LKE4CC). The use of the SCSCF multiprocessor cluster at the Università Ca' Foscari Venezia is greatfully acknowledged. T.X.H. acknowledges support from Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant No. 103.01-2013.16. R.P. acknowledges support from grant P1-0055 of the Slovene Research Agency.

[1] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic, New York, 1986).
[2] B. J. Alder and T. E. Wainwright, *J Chem Phys* **31**, 459 (1959).
[3] J.A. Barker and D. Henderson, Rev. Mod. Phys. **48**, 587 (1976)
[4] D.L. Pagan, and J.D. Gunton, J. Chem. Phys. **122**, 184515 (2005)
[5] H. Liu, S. Garde, and S. Kumar, J. Chem. Phys. **123**, 174505 (2005)
[6] C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids, Vol. 1: Fundamentals* (Clarendon, Oxford, 1984).
[7] J. Lyklema, *Fundamentals of Interface and Colloid Science, Vol. I: Fundamentals* (Academic, London, 1991).
[8] Y. Zhou, M. Karplus, J.M. Wichert and C.K. Hall, *J Chem Phys* **107**, 24 (1997).
[9] M.P. Taylor, *Molecular Physics* **86**, 73 (1995).
[10] M.P. Taylor, *J Chem Phys* **118**, 883 (2003).
[11] M.P. Taylor, W. Paul and K. Binder, *J Chem Phys* **131**, 114907 (2009).
[12] M.P. Taylor, W. Paul and K. Binder, *Phys Rev E* **79**, 050801(R) (2009).
[13] A. Grosberg and A. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).
[14] K. Yue, and K.A. Dill, Proc. Natl. Acad. Sci. U.S.A.**89**, 4163 (1992)
[15] D.T. Seaton, T. Wst and D.P. Landau, *Comp Phys Comm* **180** 587 (2009).
[16] T. Wst,Y.W. Li and D.P. Landau, *J Stat Phys* **144** 638 (2011).
[17] T.Wst and D.P. Landau, *Comp Phys Comm* **179** 124 (2008).
[18] I. Coluzza, PLOS One **6** e20853 (2011)
[19] J.R. Banavar, M. Cieplak, T.X. Hoang, and A. Maritan, Proc. Natl. Acad. Sci.U.S.A. **106**, 6900 (2009)
[20] J.R. Banavar and A. Maritan, Rev. Mod. Phys. **75**, 23 (2003)
[21] A. Maritan, C. Micheletti, A. Trovato and J. R. Banavar, Nature **406**, 287 (2000)
[22] B. Park and M. Levitt, Phys.Rev. E J. Mol. Biol. **258**, 367 (1996)
[23] J.R. Banavar, M. Cieplak, A. Flammini, T.X. Hoang, R.D. Kamien, T. Lezon, D. Marenduzzo, A. Maritan, F. Seno, Y. Snir, and A. Trovato, Phys. Rev. E **73**, 031921 (2006)
[24] A. V. Finkelstein and O. B. Ptitsyn *Protein Physics* (Academic Press 2002)
[25] H. Taketomi, Y.Ueda and N.Go, *Int J Pept Protein Res* **7**, 445 (1975).

[26] C. Clementi, H. Nymeyer, and J.N. Onuchic, *J Mol Biol* **298** 937 (2000).
[27] N. Koga and S. Takada, *J Mol Biol* **313** 171 (2000).
[28] A. Badasyan, Z. Liu and Hue Sun Chan, *J Mol Biol* **384** 512 (2008).
[29] M.P. Allen and D. J. Tildesley, *Computer Simulations of Liquids* (Clarendon, Oxford 1987)
[30] B. Smith and D. Frenkel, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego, 2002)
[31] F. Wang and D. P. Landau, *Phys Rev Lett* **86** 2050 (2001)
[32] H.S.M. Coexter *Introduction to Geometry* (Wiley 1989)
[33] C. Poletto, A. Giacometti, A. Trovato, J.R. Banavar, and A. Maritan, Phys. Rev. E **77**, 061804 (2008)
[34] R. Kamien, Rev. Mod. Phys. **74**, 953 (2002)
[35] S.M. Bhattacharjee, A. Giacometti, and A. Maritan, J. Phys.: Condens. Matter, **25**, 503101-503132 (2013)
[36] J. L. Barrat, and J. P. Hansen, *Basic Concepts for Simple and Complex Liquids* (Cambridge University Press, Cambridge) (2003).
[37] C. Leitold and C. Dellago, J. Chem. Phys. **141**, 134901 (2014).
[38] D. Reith and P. Virnau, Comp. Phys. Comm. **181**, 800 (2010).
[39] R. H. Swendsen and J.-S. Wang, *Phys. Rev. Lett.* **57**, 2607 (1986).
[40] C. J. Geyer, In Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, p. 156, New York, 1991. American Statistical Association.
[41] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
[42] N. Rathore, M. Chopra, and J. J. de Pablo, *J. Chem. Phys.* **122** 122, 024111 (2005).
[43] A. D. Sokal, *Nucl. Phys. B* **47**, 172 (1996).
[44] A. M. Ferrenberg, & R. H. Swendsen, Optimized Monte Carlo data analysis, *Phys. Rev. Lett.* **63**, 1195-1198 (1989).
[45] C. Zhou and R.N. Bhatt, *Phys Rev E* **72** 025701(R) (2005).
[46] A.D. Swetnam, M.P. Allen, *J Comp Chem* **32** 816 (2011).
[47] R.E. Belardinelli and V.D. Pereyra, *Phys Rev E* **75** 046701 (2007).
[48] J.E. Magee, L. Lue, and R.A. Curtis, *Phys Rev E* **78** 031803 (2008).
[49] J. E. Magee, Z. Song, R.A.Curtis, and L. Lue, J. Chem. Phys. **126**, 144911 (2007)
[50] Landau L. D. and Lifshitz E. M., *Statistical physics*, Vol. 5 of Course of Theoretical Physics (Butterworth-Heinemann) 1980.
[51] Thouless D. J., Phys. Rev., **187**, 732 (1969).
[52] L. Van Hove, Physica (Amsterdam) **16**, 137 (1950).
[53] G. Franzese, H. E. Stanley, J. Phys.: Condens. Matter **19**, 205126 (2007)
[54] M.E. Fisher and B. Widom, J. Chem. Phys. **50**, 3756 (1969)
[55] R. Fantoni, A. Giacometti, Al. Malijewski, and A. Santos, J. Chem. Phys. **133**, 024101 (2010)
[56] R. Fantoni, A. Giacometti, Al. Malijevský, and A. Santos, J. Chem. Phys. **131**, 124106 (2009).
[57] A. Badasyan, S.Tonoyan, A. Giacometti, R. Podgornik, and V.A. Parsegian, Phys. Rev. Lett. **109**, 068101 (2012)
[58] A. Badasyan, A. Giacometti, Y. Sh. Mamasakhlisov, V.F. Morozov and A.S. Benight Phys.Rev.E, **81**, 021921 2010.
[59] S. Yasuda, T. Yoshidome, H. Oshima, R. Kodama, Y. Harano, and M. Kinoshita, J. Chem. Phys. **132**, 065105 (2010)