General Methodology to Identify the Minimum Alphabet Size for Heteropolymer Design

*Chiara Cardelli, Francesca Nerattini, Luca Tubiana, Valentino Bianco, Christoph Dellago, Francesco Sciortino, and Ivan Coluzza**

Understanding how to design the structure of heteropolymers through their monomer sequence will have a significant impact on the creation of novel artificial materials. According to mean-field theories, the minimum number—or alphabet—of distinct monomers necessary to achieve such designability is directly related to the conformational entropy ω of compact polymer structures. Here, a computational strategy to calculate this conformational entropy is introduced and thus predict the minimum alphabet to achieve designability, for a generalized heteropolymer model. The comparison of the predictions with previous results proves the robustness of the approach. It is quantified for the first time how the number of directional interactions is critical for achieving the designability. The methodology that is introduced can be easily generalized to models representing specific polymers. A comparison between conventional polymers monomers are provided, and it is predicted that polyurea, polyamide, and polyurethane residues are optimal candidates to be functionalized for the experimental synthesis of designable heteropolymers. As such, our method can guide the engineering of new types of self-assembling modular polymers, that will open new possibilities for polymer-based materials with unmatched versatility and control.

1. Introduction

Self-assembling artificial[1–5] and natural materials[6–15] possess the ability to organize themselves into complex and heterogeneous structures. This ability is mainly driven by reversible interactions (e.g., van der Waals interactions, hydrogen bonds, dipolar interactions, depletion interactions).[16]

Considerable effort has been spent in understanding how the self-assembling behavior arises and how it can be controlled through careful design of the conformational and physicochemical properties of the primary components of the material.[17–20] A particularly successful strategy to control self-assembly is to join the components into hetero-chains, for example, proteins or small RNAs, following a specific sequence of the components along the chain. The sequence can be optimized to drive the system to fold into a unique target structure of the chain. The search for such optimized sequences for a specific target structure is commonly referred to as design,[11,21–24] and a system for which it is possible to find such a sequence for at least one target structure is called designable. The mean-field theory of the random energy model (REM)[21,25–27] predicts that for a system to be designable, it needs to respect the inequality $q > e^{\omega}$, where q is the number of different constituents (i.e., the alphabet size) and ω is the conformational entropy per monomer of the system, defined as the logarithm of the total number of accessible *compact* conformations per monomer.[26,27] This inequality can be used to identify the minimum alphabet necessary to design a heteropolymer, that is the smallest integer $q_{\min}$ greater than e^{ω}, $q_{\min} \equiv \lceil e^{\omega} \rceil$. From this, it follows that reducing ω also reduces the alphabet size $q_{\min}$. Such predictions have been tested only for lattice models.[22,27] A holy grail of the field of artificial polymer-based materials is the ability to control ω such that the alphabet of components employed in the polymer synthesis is

Dr. C. Cardelli, Dr. F. Nerattini, Dr. L. Tubiana, Prof. C. Dellago
Faculty of Physics
University of Vienna
Boltzmanngasse 5, 1090 Vienna, Austria

Dr. V. Bianco
Faculty of Chemistry
Chemical Physics Department
Universidad Complutense de Madrid, Plaza de las Ciencias, Ciudad Universitaria
Madrid 28040, Spain

Prof. F. Sciortino
Dipartimento di Fisica
Sapienza Università di Roma
Piazzale Aldo Moro 2, 00185 Rome, Italy

Prof. I. Coluzza
CIC biomaGUNE
Paseo Miramon 182, 20014 San Sebastian, Spain
E-mail: icoluzza@cicbiomagune.es

Prof. I. Coluzza
IKERBASQUE
Basque Foundation for Science
48013 Bilbao, Spain

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.201900031

DOI: 10.1002/adts.201900031

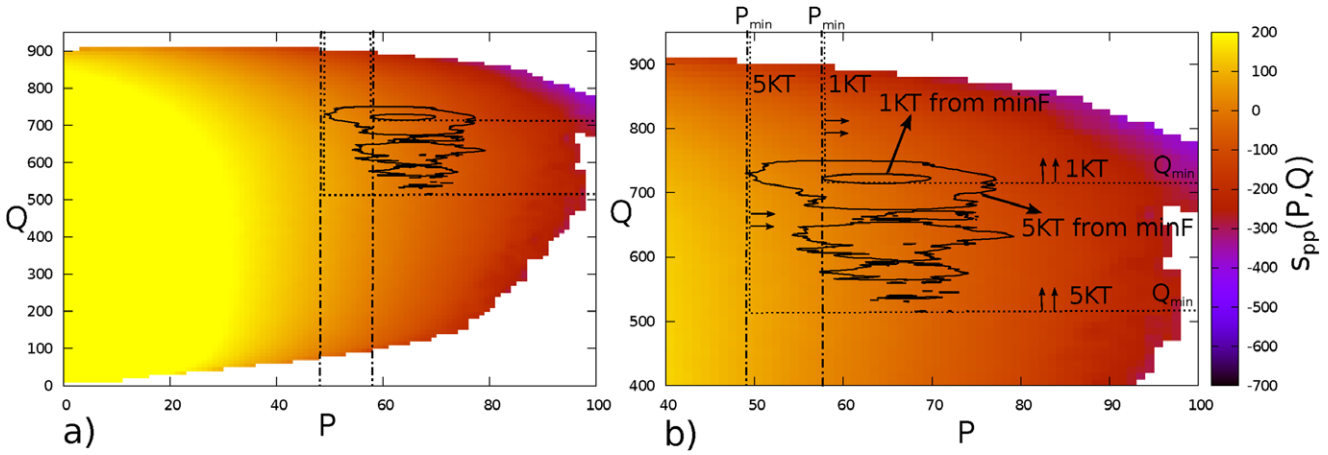

Figure 1. Shifted $s_{pp}(P, Q)$ for three patches, which indicated the different areas of $s_{pp}(P, Q)$ defined as compact conformations to calculate ω. a) Total surface. b) Zoom of the data shown in panel (a). The areas arise from the different definitions of compactness. The solid curves enclose the compact conformations within 1 k_bT (inner curve) and 5 k_bT (rough outer curve) from the free energy minimum $minF$. The vertical and horizontal lines represent, respectively, $P_{\min}$ and $Q_{\min}$ calculated as the lowest P and Q within 1 k_bT and 5 k_bT from $minF$. The dotted lines enclose the corresponding areas at their right and above, as the compactness has been defined as $P > P_{\min}$ and $Q > Q_{\min}$. For the definition of compactness based on the criterion $P > P_{\min}$, the whole areas to the right of the dashed-dotted lines have been considered.

reduced to a minimum, all while keeping a large variety of possible different target structures. It is important to stress that the *compact* polymer conformations are smaller in number than the total possible ones, hence $\omega < s$ where s is the polymer entropy. An operative definition of *compact* for off-lattice polymers is not given in the REM, making it difficult to establish a general methodology to estimate ω and in turn the designability of a heteropolymer.

Recently, Cardelli et al. have studied the self-assembling properties of a general heteropolymer consisting of patchy particles monomers (patchy-polymer) across the transition from $q < e^{\omega}$ to $q > e^{\omega}$.[28] Such a study has shown that the directional interactions are the key for designability, inducing the transition to designable polymers at $q > q_{\min} = \lceil e^{\omega} \rceil$. This suggests that the role of the directional interactions is that of decreasing ω by reducing the number of accessible compact conformations.

In this work, we verify this intuition by introducing a general methodology to estimate ω for any polymer model and apply it to patchy polymers with different numbers of patches. Since the total number of compact conformations (and thus ω) does not depend on the specific sequence,[29] we estimate ω employing a homopolymer version of the patchy polymer. Following a thermodynamic path from a reference system with known entropy, we compute the total number of conformations of the homopolymer employing advanced Monte Carlo (MC) methods.[30–32] From the total number, we then select different subsets of compact conformations, according to different operative definitions of compactness. The definitions we introduce here can be easily quantified for off-lattice polymer systems.

The methodology to estimate ω allows us to test the hypothesis and prediction of the REM on the ensemble of most designable conformations.[28,33] Having verified the validity of the theory, we can use it to predict the size of the alphabet necessary to enlarge the variety of target conformations.

2. Results and Discussion

Let us start by considering that we are projecting the whole conformational space over the number of directional contacts between the patches, P, and of the number of isotropic contacts between the beads, Q (see **Figure 1** and Figures S3– S7, Supporting Information). Significant variations of P and Q correspond to considerable changes in the chain conformations. Hence, if for the same alphabet a certain number of patches m allows for the design of a large portion of the P and Q space, we can guarantee that we will have a broad variety of different chain arrangements with high variability. This would translate into a versatile system for material science applications. Our data suggest that the m = 4,6 can access a large portion of the P and Q space (see Figures S5 and S6, Supporting Information). At the same time, the broad range still amounts to an overall number of target conformations small enough to guarantee designability with alphabets down to 2–3 letters (see **Figure 2**). Hence, m = 4,6 could be optimal for potential applications.

In this work, we use the freely rotating chain (FRC) patchy polymer model, introduced in ref. [28]. Each bead has a diameter σ and presents M patches (inset in **Figure 3**), whose arrangement on the bead surface is chosen according to the most symmetrical geometry (equidistant on the equator for $M = 3$, on a tetrahedron for $M = 4$, on an octahedron for M = 6 and see Supporting Information for $M = 10$).

The FRC is a general model for a designable heteropolymer, for which it has been demonstrated that the designability can be controlled by the number of patches $M = 0, 1, 3, 4, 6, 10$,[28] that is homogeneous along the chain for different monomer types. The alphabet of different monomers was instead defined according to the set of isotropic interaction terms used (see U_{bb} in Equation (7)). However, since ω does not depend on the sequence, from now on we employ a homopolymer version of the patchy polymer (see Equation (8)). We argue that such a hypothesis, although exact for the patchy polymer model, is valid

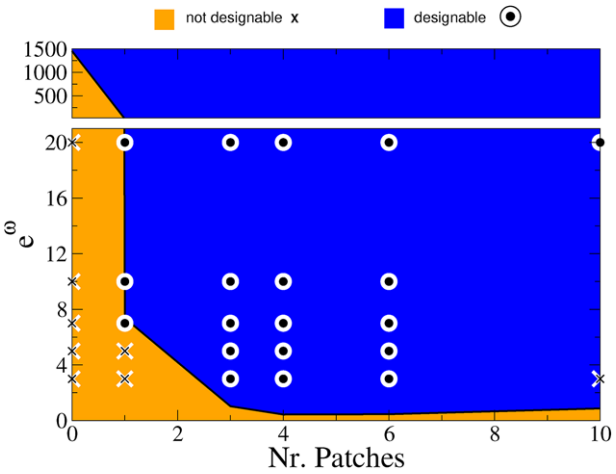

Figure 2. The broken line is exp(ω) calculated on the ensemble of compact conformations within 1 k_bT from the free energy minimum $minF$ (enclosed in the ellipse of 1 k_bT form $minF$ in Figure 1), for patches M=0,1,3,4,6,10. exp(ω) in the REM represents the alphabet q_{min} at which the transition between not designable and designable occurs. Accordingly, two areas are defined: yellow area (not designable) and blue area (designable). The line depends depends on the maximum resolution $a = \sigma/2$ we chose. Different values will shift the curve towards higher alphabets upon increasing resolution and viceversa lower one upon lowering the resolution. The points (circles and crosses) represent the alphabets verified in [28]. The circles are the designable cases, i.e. where the polymer designed with the indicated alphabet has been tested to fold into the target structure, while the crosses the ones where it does not (not designable). The broken line predicts that protein-like heteropolymers (2 patches per residue) can be designed with just 4 letters.

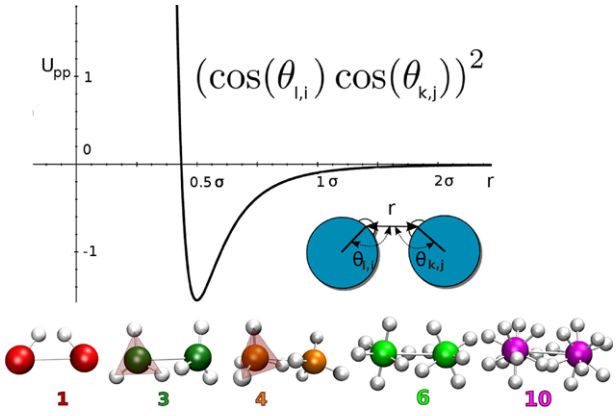

Figure 3. Patchy polymer model. Top: Directional potential of interaction between the patches: the radial Lennard-Jones contribution in the plot is multiplied by the directional contribution $(\cos\theta_{l,i}\cos\theta_{k,j})^2$, where $\theta_{l,i}$ and $\theta_{k,j}$ are the angles between the vector connecting each patch to the respective bead center and the vector that connects the two patches (right inset). Bottom: schematic illustration of two consecutive beads of the patchy polymers. The central bead is represented smaller than its size for the sake of visualization. The small white spheres indicate the patches that are placed on the surface of the bead.

also if we include the conformational entropy of the amino acids side chains that we estimated from the rotamers libraries (see Supporting Information for details). For this reason, decoupling the sequence and the backbone conformational entropy is a good approximation.

We consider the same patch arrangements of ref. [28], shown at the bottom of Figure 3. For these cases, we aim at computing the conformational entropy Ω as a function of the number of directional contacts between the patches, P, and of the number of isotropic contacts between the beads, Q (see Experimental Section). Ω, according to the REM,[27] is defined as the logarithm of the number of *compact* conformations $\mathcal{N}_c$ of a heteropolymer composed of N monomers. Hence, we have to firstly compute the total conformational entropy $s_{pp}(P, Q) = \ln(\mathcal{N}_{pp}(P, Q))$ of the patchy polymer, that is the logarithm of the total number of polymer conformations $\mathcal{N}_{pp}(P, Q)$. Secondly, from the total number $\mathcal{N}_{pp}(P, Q)$, we select the subset of the compact conformations.

The total conformational entropy $s_{pp}(P, Q)$ computed with Monte Carlo simulations is not an absolute entropy, it is defined up to an arbitrary constant. For this reason, as explained in the following and represented in **Figure 4**, at first we need to compute the absolute entropy of a self-avoiding polymer $s_{saw} = \ln(\mathcal{N}_{saw})$, where $\mathcal{N}_{saw}$ is the number of conformations of a self-avoiding chain. It is important to stress that the self-avoiding polymer model is equivalent to the FRC model when the bead–bead and patch–patch interactions are switched off (see Equation (9)).

To correctly compute s_{saw}, we need to know the number of conformations of a reference state. The chosen reference state is a trimer of self-avoiding bonded beads, whose entropy $s_{saw}|_{N=3}$ can be calculated analytically. The partition function is measured by the algorithm we employed and in principle can also be calculated analytically (e.g., Taylor[34]). We checked that the numerical ≈ 0.588 and analytical calculations $Z_2 = 4\pi \int_1^\infty r^2 \exp[-\beta(K(R_{i\,i+1} - 1)^2)] \approx 0.588$ of the partition function of a dimer are consistent. However, this volume is only proportional to the number of conformations, as well as the entropy is defined according to an arbitrary constant. Hence, we introduce a length scale a to discretize the continuum space. The physical meaning of such length scale is the resolution of discrimination power of the design process. The choice of the resolution is arbitrary and the following results will rescale according to the chosen resolution. In fact, higher resolution corresponds to smaller values of a and an increase in the number of conformations which in turn will require a larger alphabet. Viceversa lower resolution corresponds to larger values of a and smaller conformational space, and finally smaller alphabets. In what follows, we will use the same resolution introduced in the REM[27] $a \equiv \sigma/2$ as the radius of the tube following one chain so that all the other conformations that fall within it are considered equivalent. This definition implies that two beads with centers closer than $a \equiv \sigma/2$ correspond to indistinguishable conformations. Following this definition, we consider $a \equiv \sigma/2$ as the characteristic length, that distinguishes between different states.

Since the bond spring is strong, each bead can only be placed on the surface of the other one with tiny fluctuations $\approx 0.026\sigma$. Hence, we assume that the number of conformations of each bead is merely the number of ways we can place a sphere of diameter σ on the surface of an identical sphere.

The close packing of hard spheres would give 12 different arrangements, where each conformation is distant by σ (or more) from the other, along with the surface. By partitioning the total surface into sections of side length a, we obtain $12(\frac{\sigma}{a})^2 = 48$ different conformations. In fact, starting from the close packing conformations, each bead can be shifted of a in four directions

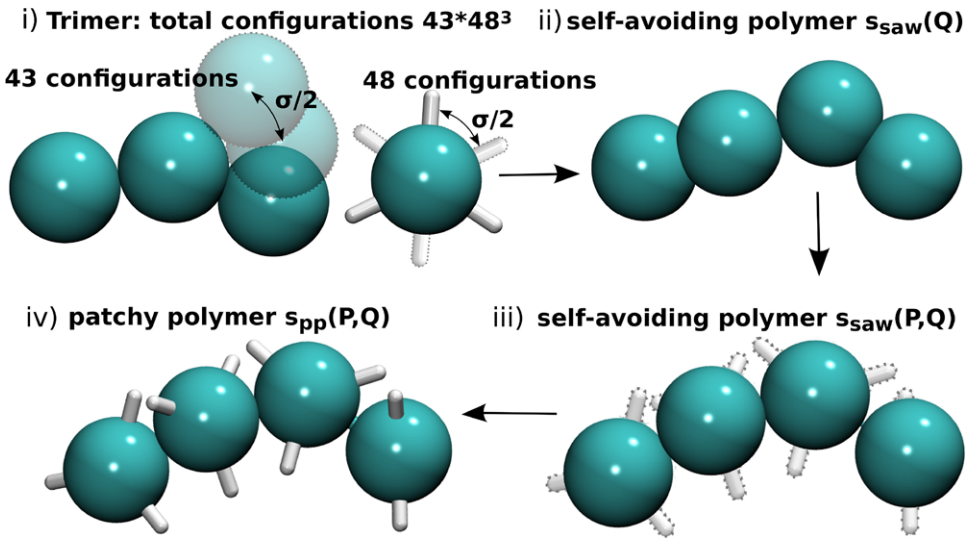

Figure 4. Scheme of the path used to calculate the total conformation entropy of the patchy homopolymer starting from an analytically known system (trimer). i) The absolute total entropy $s_{saw}|_{N=3}$ of a self-avoiding trimer is calculated analytically. Accordingly, the absolute total entropy for the self-avoiding polymer with $N = 50$ $s_{saw}|_{N=50}$ is calculated, employing the gran canonical method in refs. [30,31]. ii) With the same method, we calculate $s_{saw}(Q)$ for $N = 50$ as a function of the number of isotropic contacts Q, and shift it so that $\ln \int \mathcal{N}_{saw}(Q)dQ = s_{saw}|_{N=50}$ defined in Equation (2). iii) For $N = 50$, we calculate via the enhanced MC method in ref. [32] $s_{saw}(P, Q) = \ln \mathcal{N}_{saw}(P, Q)$ with the additional variable P, naming the number of patch–patch contacts. We shift $s_{saw}(P, Q)$ so as $\ln \int \mathcal{N}_{saw}(P, Q)dP$ overlaps with $s_{saw}(Q)$ (as in Figure S1, Supporting Information). iv) For $N = 50$, we calculate via the enhanced MC method in ref. [32]. $s_{pp}(P, Q)$ of the patchy homopolymer and we shift it upon $s_{saw}(P, Q)$.

along with the surface getting a different conformation. In the trimer, we fix the position of the first bead to avoid over-counting configurations related by translations of the entire chain. The position of the second bead with respect to the first one is also fixed, to avoid over-counting configurations related by rotations of the entire chain. Finally, we subtract the contribution from the torsional degrees of freedom around the dimer bond $\ln(12) \approx 2.5$ (see Supporting Information for details). From the total 48 conformations that we should consider for the third bead, we have to subtract $(\frac{\sigma}{a})^2 + 1 = 5$ because the third bead is not allowed to overlap (completely or partially) with the first one. On top of this, we have to count also the rotational degrees of freedom of all the beads because the patches break the rotational invariance symmetry. Following the same criteria, each bead can be rotated in $12(\frac{\sigma}{a})^2 = 48$ different conformations, leading to $[12(\frac{\sigma}{a})^2]^3 = 48^3$ rotational states for the trimer. Hence, the total trimer's conformational entropy $s_{saw}|_{N=3,a=\frac{\sigma}{2}}$ is[35]

$$s_{saw}|_{N=3} = \ln\left(11\left(\frac{\sigma}{a}\right)^2 - 1\right) + 3\ln\left(12\left(\frac{\sigma}{a}\right)^2\right) - \ln(12) \approx 12.9 \tag{1}$$

Starting from the trimer as the reference system, following the gran canonical method proposed in refs. [30,31], we can calculate the absolute value of the total entropy for a self-avoiding polymer of length $N = 50$ (see Figure S2, Supporting Information). The value calculated from the simulation $s_{saw}^{\text{simul}}|_{N=50}$ needs to be shifted considering the absolute entropy of the trimer, thus by the difference between the absolute value in Equation (1) and the value calculated by the simulation $s_{saw}^{\text{simul}}|_{N=3} + 3\ln(12(\frac{\sigma}{a})^2)$. Note that also in the latter, the need to add $3\ln(12(\frac{\sigma}{a})^2)$ is because the term $s_{saw}^{\text{simul}}|_{N=3}$ does not take into account the rotational degrees of freedom, because they are not included in the simulation method used. For the same reason, we also have to sum $50\ln(12(\frac{\sigma}{a})^2)$—the rotational degrees of freedom for $N = 50$—to the total absolute entropy for $N = 50$, that results as following

$$s_{saw}|_{N=50} = s_{saw}|_{N=3} - \left[s_{saw}^{\text{simul}}|_{N=3} + 3\ln\left(12\left(\frac{\sigma}{a}\right)^2\right)\right] + s_{saw}^{\text{simul}}|_{N=50} + 50\ln\left(12\left(\frac{\sigma}{a}\right)^2\right) \approx 368 \tag{2}$$

Hence, we have enumerated the number of distinct number of states according to our choice of the resolution a. It is important to stress that the number of states does not change significantly with the resolution. In fact, for $a = \sigma/10$ (typical refolding precision of the patchy polymer[28] that in protein would correspond to 0.4 Å resolution) $s_{saw}|_{N=50} = 377$, while for $a = 1.5\sigma$ (resolution that would not distinguish between folded and unfolded) $s_{saw}|_{N=50} = 359$, which correspond to a 2% (see Figure S2, Supporting Information). Then, we compute with the same method the curve $s_{saw}(Q)$ for $N = 50$: It is important to stress that such a method does not include the sampling over the patch–patch bond collective variable P and thus samples conformational entropy $s_{saw}(Q)$ only as a function of the number of isotropic contacts Q. We shift the curve $s_{saw}(Q)$ in such a way, that the total entropy is corresponding to the absolute value of $s_{saw}|_{N=50}$ in Equation (2)

$$\ln \int \mathcal{N}_{saw}(Q)dQ = s_{saw}|_{N=50} \tag{3}$$

Then, we compute through Monte Carlo simulations, enhanced via the virtual move parallel tempering (VMPT) algorithm,[32]

both $s_{saw}(P, Q)$ with the additional collective variable P and $s_{pp}(P, Q)$ for the patchy homopolymer (see Equation (8)). By firstly aligning the surfaces $s_{pp}(P, Q)$ and $s_{saw}(P, Q)$, and then aligning the curves $\ln \int \mathcal{N}_{saw}(P, Q) dP$ and $s_{saw}(Q)$, we get the absolute surface $s_{pp}(P, Q)$.

We observe a good agreement between the absolute conformational entropy of the self-avoiding polymer $s_{saw}(Q)$ calculated via the gran canonical method and $\ln \int \mathcal{N}_{saw}(P, Q) dP$, calculated via the VMPT MC simulation,[32] in the range of Q that they both sample (see Figure S1, Supporting Information). The VMPT method samples a much wider range of Q, thanks to its sampling enhancement. Figure 1b shows the conformational entropy for the patchy-polymer $s_{pp}(P, Q)$, shifted as described in Experimental Section. We observe that $s_{pp}(P, Q)$ has its maximum for low P and Q, while the free energy minimum is found in the area with high P and Q, enclosed in the ellipse in panel b) labelled as "1 k_bT from $minF$". This observation is the first indication that the number of accessible conformations at low temperatures is highly reduced by the presence of the directional interactions between the patches.

From the absolute value of $s_{pp}(P, Q)$, we now can calculate $q_{min} = \lceil e^{\omega} \rceil$ by counting the number of compact conformations ω. According to the REM, the compact off-lattice conformations are the ones with "small temporal and spatial density fluctuations,"[27] but no operative definition is given to identify them. To overcome this problem, here we firstly calculate e^{ω} as the sum of the number of conformations found in the basin within 1 k_BT from min F, as indicated by the inner ellipse in Figure 1. Such a definition is based on the operative definition of the designability given in the work of Cardelli et al.,[28] where the design is performed for the structures found within 1 k_BT of the global free energy minimum of the P–Q space. This definition of compactness includes a measure of the effect of the directionality on the conformational entropy ω.

In Figure 2, we show the results for number of compact conformations per monomer (e^{ω}) for different patch numbers, calculated with the methodology described above. e^{ω}, following the REM, predicts the minimum alphabet size q_{min} for the system to be designable, that is, it traces the transition line between designable and not designable alphabets for different patch numbers. In Figure 2, we compare this transition line with the designability categorization given in ref. [28], obtained by explicitly designing and refolding each target structure for different alphabet sizes. We observe considerable accordance between the two methods: the designable points verified in ref. [28] are within the designable area predicted by REM, except the ten patches case. Our calculations predict that to design heteropolymers with two directional interaction sites per monomer (e.g., proteins), the minimum alphabet necessary is composed of just four letters. This prediction is consistent with the experimental observation that just five letters are enough to encode structural information in proteins[36–41] and with a study performed with the Caterpillar protein model[42,43] that will be presented in an upcoming publication.

In **Figure 5**, we extend the predictions of the REM by testing different definitions of compactness that correspond to larger P–Q areas of possible target conformations, to predict how much would the minimum alphabet grow with a bigger variability of target conformations. The dotted lines in Figure 5 represent e^{ω}

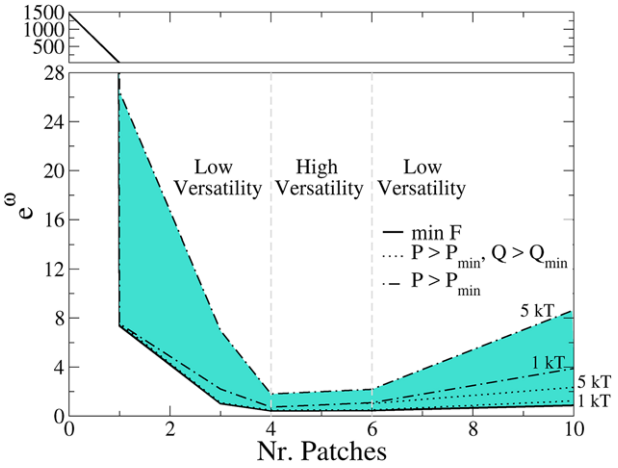

Figure 5. $e^{\omega} = q_{min}$ alphabet at which the transition between not designable and designable occurs versus the number of patches. The broken lines represent the calculated alphabet exp(ω) with different definitions of the ensemble of compact structures, as defined in Figure 1. The lowest solid line contains actually two overlapping not distinguishable lines: e^{ω} calculated on the two ensembles enclosed in the curve of 1 k_bT and 5 k_bT from $minF$, as in Figure 1. The dashed and dotted-dashed lines correspond to e^{ω} calculated within the ensembles defined by the corresponding vertical and horizontal lines in Figure 1. We observe that for the systems with one and ten patches, the minimum alphabet at which the system is designable increases considerably by increasing the variety of possible target conformations (different P and Q), while for the systems with four and six patches, it remains low.

calculated summing all compact conformations with $P > P_{min}$ and $Q > Q_{min}$ as delimited by the dotted lines in Figure 1, where P_{min} and Q_{min} are the lowest P and Q values found within 1 k_bT (or 5 k_bT) $minF$. The dashed-dotted lines in Figure 5 represent e^{ω} calculated by summing all the compact conformations with $P > P_{min}$ for 1 k_bT (or 5 k_bT), as delimited by the corresponding dotted lines in Figure 1. We observe that by enlarging the P–Q area, and thus the variety of possible target conformations, e^{ω} increases, meaning that the system requires a larger minimum alphabet in order to be designable. However, for all the scenario considered, the variation in the minimum alphabet is not very large, except for the $P > P_{min}$ scenario where we also included many conformations that started to be more open ($P < 60$ and $Q < 500$) hence with larger conformational entropy. We can draw two conclusions from this observations: i) the results are not very sensitive to the particular combination of P and Q (and implicitly the U_{bb} interactions) provided that there are enough of one of the two contacts, ii) the alphabet size is extremely sensitive to the variation of P.

For the systems with four and six patches, q_{min} does not increase significantly by increasing the variety of possible target structures, therefore with these number of patches one can choose different target structures that maintain designability with low alphabet sizes, thus allowing a higher variability of structures. Finally, the system with ten patches loses designability more easily by increasing the variety of target structures; this is explained by the loss in directionality that makes the monomer with M = 10 act as a bigger self-avoiding sphere, as shown in ref. [28].

Table 1. Below we list common polymers monomers and the number of potential hydrogen bonding per monomer m. According to our predictions, polymers with $m \geq 1$ on their backbones are good candidates for the scaffold of designable heteropolymers (indicated with a "Yes" or "No" in the last column). For DNA/RNA, we indicated only a lower bound for m because the Watson and Crick pairing is not the only directional interaction possible (see discussion in the Supporting Information). Among the synthetic compounds, polyurea, polyamide, and polyurethane monomers are optimal candidates for an experimentally designable heteropolymer. For the design, the polymers backbone should be heterogeneous and with control over the sequence of residues. The polymers in this table are generally modifiable (e.g., polyurea has groups that can be modified along the backbone) and the chain growth can be highly effective. A possible alternative strategy is to alternate along the chain beads with directional interactions with beads with just volumetric interactions.[51–54] The sequence would be applied only to the latter, while the directional beads are used to reduce the minimum alphabet.

Polymer	m	Designable
Proteins	2	Yes
DNA/RNA	> 1	Yes
Glycans[a)]	> 1	Yes
Polysaccaride[b)]	2	Yes
Polyurethane	2	Yes
Polyamide	4	Yes
Polyglycolide	0	No
Polylactic acid	0	No
Polycaprolactone	0	No
Polyhydroxyalkanoate	2	Yes
Polyhydroxybutyrate	2	Yes
Polybutylene succinate	0	No
Poly(3-hydroxybutyrate-co-3-hydroxyvalerate)	0	No
Polyethylene terephthalate	0	Yes
Polybutylene terephthalate	0	No
Polytrimethylene terephthalate	0	No
Polyethylene naphthalate	0	No
Vectran	0	No
Polyurea[c)]	4	Yes
Silicone	0	No
Polycarbonate	0	No
Polyethylene glycol	0	No
Polypropylene glycol	1	Yes
Paraformaldehyde	1	Yes
Polytetrahydrofuran	0	No

[a)] Depends on the chain structure; [b)] Depending on the solvent; [c)] $M = 4$ is in the high versatility region.

Based on our results, we can compile a table of potential monomers that could be used to synthesize designable heteropolymers backbone (see **Table 1**). The monomers with potential hydrogen bonding sites are the candidates that we identified could be used as a scaffold to synthesise folding heteropolymers. In particular, we think that monomers of polyurea is particularly attractive as a starting system as it offers at least two groups that can be functionalized. Moreover, polyurea has four potential hydrogen bonds per monomer, which fall within the 'high versatility" region shown in Figure 5. For similar reasons, polyamide and polyurethane are good alternatives. A possible way of synthesis would follow a similar protocol to the ones for protein syntheses where the amino acids chains are grown from a solid-state substrate, and the different monomers are added sequentially.[44,45] It is important to stress that the implications of our results extend beyond the specific patchy polymer model. The are two main arguments to support such a statement. The first important aspect is related to the role of the molecular backbone bonds to restrict the conformational space of the chain. In principle, the patchy polymer FRC that we tested has the minimum number of backbone constraints of realistic polymer chains. In a previous publication, we showed that by increasing the backbone constraints by simply moving the chain bond from the particle centers to the surface of them increased the designability.[28] Hence, a more realistic model that would include torsional and dihedral constraints atmost could only further reduce the conformational space, and not increase it, making our predictions an upper limit for the minimum alphabet. Interestingly, Chen et al.[46] recently made a related observation on the modelling of FiP35 where, with a bare C_α model, it was not possible to define a specific interaction matrix (in this case, a 35 letter alphabet) that would allow distinguishing between the folded and the unfolded state clearly. Upon adding the C_β atoms, expanding the alphabet to 70 letters, and introducing a directionality due to the excluded volume, the model chain was capable of correctly folding into the target structure. Hence, the result further supports the importance of the backbone directional interactions to reduce the frustration of the protein folding.

Secondly, the side chains add further directionality to the residue–residue interactions (e.g., they tend to spend most of the time on one side of the backbone) that necessarily would reduce further the conformational space, making our estimate for the minimum alphabet an overestimate. We believe that even an overestimate of the $q_{\min}$ can be extremely useful especially when the values are already within reach of experiments.

3. Conclusion

In conclusion, in this manuscript, we present the first accurate calculations of the REM conformational entropy ω for a patchy polymer as a function of the number of patches. The values of ω are absolute, as they were obtained via a thermodynamic path, starting from the analytically solvable model of a self-avoiding trimer all the way to the full patchy polymer. The calculation of the path required extensive simulations for each number of patches investigated.

The main conclusion that can be drawn from our work is that the REM inequality $q > e^{\omega}$, defining the minimal alphabet for heteropolymer design, is valid and accurately reproduces the designability phase diagram previously obtained by Cardelli et al.[28] Moreover, the inequality allows the prediction of the minimal alphabet necessary to increase the variety of target structures for the design. It is important to stress that the predicted minimal alphabets are upper limits, as in real systems there are more molecular constraints and particles are not necessarily free to rotate. In fact, we showed previously[28] that reducing the bead rotational degrees of freedom further reduces the minimal alphabets.

This knowledge combined with our methodology for the calculation of ω provides a powerful tool for heteropolymer

engineering. Finally, we could identify in the four and six patches geometries the optimal ones for heteropolymer design, since the minimal alphabet grows very slowly with the space of different conformations included in the calculations of ω. Our methodology can be directly transferred to any heteropolymer model to establish the minimal alphabet making the system designable. In particular, we think it will be particularly effective in cases where the conformational space is restricted by directional interactions such as hydrogen bonding (proteins, RNA, the building blocks in DNA origami) or dipolar interactions (dipolar or magnetic colloids). We ranked common polymers monomers according to our prediction on their designability, and we identified polyurea, polyamide, and polyurethane as optimal choices for the synthesis of designable heteropolymers. It is important to mention that the method can be applied to compare the designability of different protein models and estimate the minimum alphabet necessary to design them, that seems to be already of four letters for the Caterpillar coarse-grained model,[43] as suggested by an upcoming work of some of the present authors. As a future perspective, we are planning to measure the dependence of omega from the bending rigidity and for combinations of particles with different geometries. Additionally, we have preliminary results that show that a reduced conformational space can also be achieved using dipolar interactions. Dipoles could be enough to allow for designability as hinted by the work of Combe et al.[47] who used them to represent the hydrogen-bond interactions in a folding protein model.

4. Experimental Section

FRC Model Details: The full Hamiltonian of the FRC patchy polymer is given by

$$H_{FRC} = \sum_{i<j}^{N} U(R_{ij}) + 4 \sum_{i+2<j}^{N} \sum_{l,k=1}^{M} U_{pp}(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}) + \sum_{i=1}^{N-1} U_B(R_{i\,i+1}) + \sum_{i+2<j}^{N} U_{bb}(R_{ij}) \quad (4)$$

where N is the total number of beads, and R_{ij} is the distance between the beads i and j. The indexes i, j run over the N beads, while the indexes l and k run on the M patches of the beads i and j, respectively.

The first term represents a self-avoiding potential acting between all the beads and is defined as $U(R_{ij}) \equiv \infty$ if $R_{ij} \leq \sigma$ and $U(R_{ij}) = 0$ otherwise.

The second term represents the directional interaction between the patches of beads that are separated by at least two positions along the chain, and is given by the potential derived by Irbäck et al.,[48] commonly used to model hydrogen bonds

$$U_{pp}(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}) \equiv \epsilon_p(\cos\theta_{l,i}\cos\theta_{k,j})^{\nu}\left[5\left(\frac{\sigma}{2r_{lk}}\right)^{12} - 6\left(\frac{\sigma}{2r_{lk}}\right)^{10}\right] \quad (5)$$

where the $\theta_{l,i}$ and $\theta_{k,j}$ are the angles between the vector $\mathbf{r}_{lk} \equiv \mathbf{r}_{l,i} - \mathbf{r}_{k,j}$ and the vectors $\mathbf{r}_{l,i} - \mathbf{R}_i$ and $\mathbf{r}_{k,j} - \mathbf{R}_j$, connecting the patches l and k to the center of their bead, respectively (Figure 3). Accordingly, we have $r_{lk} \equiv |\mathbf{r}_{lk}|$. The potential U_{pp} tends to align the vectors $\mathbf{r}_{lk}$, $\mathbf{r}_{l,i} - \mathbf{R}_i$, and $\mathbf{r}_{k,j} - \mathbf{R}_j$ with minimum when the patches are at distance $r_{kl} = \sigma/4$, representing a conformation where the patches face each other, while it vanishes when $\mathbf{r}_{lk}$ is orthogonal to one of the vectors $\mathbf{r}_{l,i} - \mathbf{R}_i$ or $\mathbf{r}_{k,j} - \mathbf{R}_j$. When $\theta_{l,i} < \pi/2$ or $\theta_{k,j} < \pi/2$, we fix $U_{pp} = 0$ to avoid anti-parallel arrangements of the patches. U_{pp} is cut off at $R_{i,j} = 1.5\sigma$. For the values of the prefactor and exponent, we take $\epsilon_p = 3.1\,k_BT$ and $\nu = 2$.[48] Note that the term U_{pp} excludes the interactions between beads that are first and second neighbors along the backbone.

The third term in the Hamiltonian is a harmonic bonding potential, connecting consecutive beads along the chain

$$U_B(i,\,i+1) = K\,(R_{i\,i+1} - \sigma)^2 \quad (6)$$

where $K = 10\,k_BT$. The last term represents the isotropic bead–bead interaction and is modelled with the potential

$$U_{bb}(R_{ij}) = \epsilon_{ij}\left[\frac{1}{1.0 + e^{-2.5(3\sigma - R_{ij})}}\right] \quad (7)$$

where ϵ_{ij} is a sequence-dependent prefactor depending on the types of bead i and j. U_{bb} is essentially a sigmoidal function with constant value $U_{bb} \approx \epsilon_{ij}$ for $R_{ij} \leq 2.5\sigma$, decaying to $U_{bb} \approx 0$ for $R_{ij} \geq 3.5\sigma$, with the inflection point $U_{bb}(3\sigma) = \epsilon_{ij}/2$. The factor 4 in front of the term U_{pp} is fixed to balance the directional contributions U_{pp} with respect to the isotropic one U_{bb}.

The FRC homopolymer Hamiltonian is obtained by setting this last term $U_{bb} = 0$ in Equation (4)

$$H_{homo} = \sum_{i<j}^{N} U(R_{ij}) + s \sum_{i+2<j}^{N} \sum_{l,k=1}^{M} U_{pp}(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}) + \sum_{i=1}^{N-1} U_B(R_{i\,i+1}) \quad (8)$$

while the self-avoiding chain Hamiltonian is obtained by setting $U_{bb} = 0$ and $U_{pp} = 0$ in Equation (4)

$$H_{saw} = \sum_{i<j}^{N} U(R_{ij}) + \sum_{i=1}^{N-1} U_B(R_{i\,i+1}) \quad (9)$$

Note that in all models, the only interactions of nearest neighbor beads are the harmonic spring potential (third term) and the hard core repulsion (first term).

For the sampling of the entropy, we employ Monte Carlo simulations to compute the total conformational entropy, that is , the logarithm of the total number of polymer conformations s_{saw} and s_{pp} of the self-avoiding polymer (see Equation (9)) and a patchy homopolymer (see Equation (8)), respectively. Both are homopolymer models, since s_{saw} and s_{pp} (and Ω) do not depend on the specific sequence, as already discussed in the main text. The entropy of the patchy polymer s_{pp} is computed via canonical Monte Carlo (MC) simulations where only the polymer conformations are sampled via a set of conformational moves. We perform single particle translation and rotations as well as pivot and crankshaft moves.[49] The conformational entropy is projected onto both the number of contacts Q and the number of patch–patch contacts P. Similar algorithms in the past showed the ability to sample extensively the conformational space.[28,50]

The number of bead–bead isotropic contacts Q is associated to the U_{bb} term

$$Q = \sum_{i+2<j}^{N}\left[\frac{1}{1.0 + e^{-2.5(3\sigma - R_{ij})}}\right] \quad (10)$$

and the number of directional contacts P is connected instead to the U_{pp} term. P is given by the number of i–j pairs of patches that face each other by satisfying the conditions: $r_{ij} < 0.625\,\sigma$, $\theta_{l,i} > 0.8\,\pi$, and $\theta_{k,j} > 0.8\,\pi$ (as shown in Figure 3).

We enhance the sampling with the VMPT algorithm,[32] iteratively accumulating the biasing potential over Q and P, and performing each simulation at 32 different temperatures in the set [3.0 2.75 2.5 2.0 1.9 1.7 1.6 1.5 1.45 1.4 1.35 1.3 1.25 1.2 1.15 1.05 1.025 1.0 0.95 0.925 0.9 0.8 0.75 0.7 0.65 0.6 0.590 0.575 0.55 0.5 0.45 0.4]. For each patch geometry, we perform ten independent simulations run in subsequent simulation blocks of 10^9 MC steps. We repeat the iterations until we observe that the conformational entropy landscape does not show an appreciable difference from the previous run and among the ten independent simulations.

The entropy of the self-avoiding chain s_{saw} is computed both with the VMPT scheme described above ($s_{saw}(P, Q)$) and with the gran canonical scheme introduced in refs. [30,31] ($s_{saw}(Q)$). The latter method allows computing the ratio between the partition functions of polymers with N and $N-1$ monomers via a gran canonical Monte Carlo simulation, where the chain grows by adding one bead at a time and sampling all its possible conformation. Hence, starting from a trimer, it is possible to compute the entropy difference between $N = 50$ and $N = 3$, which is analytically known from Equation (1).

To pass from the curve $s_{saw}(Q)$ to the surface $s_{saw}(P, Q)$ for a self-avoiding polymer, we perform new Monte Carlo simulations of the FRC self-avoiding patchy polymer (defined in Equation (9)) for each value of M. We sample its conformational entropy $s_{saw}(P, Q)$ with the additional collective variable P. By integrating over the order parameter P, $\ln \int \mathcal{N}_{saw}(P, Q) dP$, we obtain the same curve $s_{saw}(Q)$ up to an additive constant. Thus, by aligning $\ln \int \mathcal{N}_{saw}(P, Q) dP$ with the absolute entropy $s_{saw}(Q)$, we compute the proper shift for the surface $s_{saw}(P, Q)$. A key difference between the two simulations we employed to compute $s_{saw}(Q)$ and $s_{saw}(P, Q)$ is that in the latter also single particle rotational moves are performed, allowing us to create a path between the bare self-avoiding polymer and the full patchy polymer. Moreover, it is worth noticing that we could not directly compute the absolute value of $s_{saw}(P, Q)$ with the gran canonical method of refs. [30,31] because it does not sample properly the polymer conformations corresponding to large values of P.

The last step consists of a Monte Carlo simulation of the patchy polymer with the patches interaction enabled, allowing us to sample the conformational entropy $s_{pp}(P, Q)$ for the different numbers of patches. The computed surface $s_{pp}(P, Q)$ is then aligned with $s_{saw}(P, Q)$ to get the absolute entropy for the patchy polymer.

All the shifts are obtained by minimizing the sum of square distances of the points of the conformational entropy profiles.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

All simulations presented in this paper were carried out on the Vienna Scientific Cluster (VSC). The authors acknowledge support from the Austrian Science Fund (FWF) project 26253-N27. V.B. acknowledges support from the FWF, Grant No. M 2150-N36.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

bionic protein, heteropolymer, protein design, protein folding

Received: February 13, 2019
Revised: April 2, 2019
Published online:

[1] Y. Umena, K. Kawakami, J.-R. Shen, N. Kamiya, *Nature* **2011**, *473*, 55.
[2] D. G. Nocera, *Acc. Chem. Res.* **2012**, *45*, 767.
[3] E. A. Kamenetzky, L. G. Magliocco, H. P. Panzer, *Science* **1994**, *263*, 207.
[4] S. Furumi, *J. Mater. Chem. C* **2013**, *1*, 6003.
[5] Y. Yang, D. Bolikal, M. L. Becker, J. Kohn, D. N. Zeiger, C. G. Simon, *Adv. Mater.* **2008**, *20*, 2037.
[6] W. M. Shih, J. D. Quispe, G. F. Joyce, *Nature* **2004**, *427*, 618.
[7] P. W. K. Rothemund, *Nature* **2006**, *440*, 297.
[8] G. Zanchetta, T. Bellini, M. Nakata, N. A. Clark, *J. Am. Chem. Soc.* **2008**, *130*, 12864.
[9] J. Conde, N. Oliva, M. Atilano, H. S. Song, N. Artzi, *Nat. Mater.* **2016**, *15*, 353.
[10] W. F. DeGrado, C. M. Summa, V. Pavone, F. Nastri, A. Lombardi, *Annu. Rev. Biochem.* **1999**, *68*, 779.
[11] S. F. Betz, D. P. Raleigh, W. F. DeGrado, *Curr. Opin. Struct. Biol.* **1993**, *3*, 601.
[12] M. Chino, O. Maglio, F. Nastri, V. Pavone, W. F. Degrado, A. Lombardi, *Eur. J. Inorg. Chem.* **2015**, *2015*, 3371.
[13] T. A. Whitehead, A. Chevalier, Y. Song, C. Dreyfus, S. J. Fleishman, C. De Mattos, C. A. Myers, H. Kamisetty, P. Blair, I. A. Wilson, D. Baker, *Nat. Biotechnol.* **2012**, *30*, 543.
[14] N. P. King, W. Sheffler, M. R. Sawaya, B. S. Vollmar, J. P. Sumida, I. André, T. Gonen, T. O. Yeates, D. Baker, *Science* **2012**, *336*, 1171.
[15] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, *Science* **2003**, *302*, 1364.
[16] D. S. Lawrence, T. Jiang, M. Levett, *Chem. Rev.* **1995**, *95*, 2229.
[17] E. Bianchi, R. Blaak, C. N. Likos, *Phys. Chem. Chem. Phys.* **2011**, *13*, 6397.
[18] P. F. Damasceno, M. Engel, S. C. Glotzer, *Science (80-.)* **2012**, *337*, 453. arXiv:1202.2177 (cond-mat.soft).
[19] W. B. Rogers, W. M. Shih, V. N. Manoharan, *Nat. Rev. Mater.* **2016**, *1*, 16008.
[20] E. Bianchi, B. Capone, I. Coluzza, L. Rovigatti, P. D. J. van Oostrum, *Phys. Chem. Chem. Phys.* **2017**, *19*, 19847. arXiv:1705.04383.
[21] A. M. Gutin, E. I. Shakhnovich, *J. Chem. Phys.* **1993**, *98*, 8174.
[22] M. R. Betancourt, J. N. Onuchic, *J. Chem. Phys.* **1995**, *103*, 773.
[23] I. Coluzza, *J. Phys. Condens. Matter* **2017**, *29*, 143001.
[24] V. Bianco, G. Franzese, C. Dellago, I. Coluzza, *Phys. Rev. X* **2017**, *7*, 21047.
[25] B. Derrida, *Phys. Rev. B* **1981**, *24*, 2613.
[26] V. S. Pande, A. Y. Grosberg, T. Tanaka, *Biophys. J.* **1997**, *73*, 3192.
[27] V. S. Pande, A. Y. Grosberg, T. Tanaka, *Rev. Mod. Phys.* **2000**, *72*, 259.
[28] C. Cardelli, V. Bianco, L. Rovigatti, F. Nerattini, L. Tubiana, C. Dellago, I. Coluzza, *Sci. Rep.* **2017**, *7*, 4986.
[29] For each sequence, some conformations are more likely than others, but this does not prevent the polymer to explore them.
[30] T. Vissers, F. Smallenburg, G. Munaò, Z. Preisler, F. Sciortino, *J. Chem. Phys.* **2014**, *140*, 144902.
[31] M. Ronti, L. Rovigatti, J. M. Tavares, A. O. Ivanov, S. S. Kantorovich, F. Sciortino, *Soft Matter* **2017**, *13*, 7870. https://doi.org/10.1039/C7SM01692A.
[32] I. Coluzza, D. Frenkel, *Chemphyschem* **2005**, *6*, 1779.
[33] See Supporting Information for details on the scaling properties of a 50 bead chain.
[34] M. P. Taylor, *J. Chem. Phys.* **2003**, *118*, 883.
[35] Alternatively, distributing the sphere randomly, gives an estimate of $(2+\sqrt{3})/4 * 48 = 24 + 12\sqrt{3} \approx 44$ very close to the one in Equation 1.

[36] K. W. Plaxco, D. S. Riddle, V. Grantcharova, D. Baker, *Curr. Opin. Struct. Biol.* **1998**, *8*, 80.
[37] H. S. Chan, *Nat. Struct. Biol.* **1999**, *6*, 994.
[38] J. Wang, W. Wang, *Nat. Struct. Biol.* **1999**, *6*, 1033.
[39] L. R. Murphy, A. Wallqvist, R. M. Levy, *Protein Eng. Des. Sel.* **2000**, *13*, 149.
[40] M. T. Reetz, S. Wu, *Chem. Commun.* **2008**, 5499.
[41] A. D. Solis, *Proteins Struct. Funct. Bioinforma.* **2015**, *83*, 2198.
[42] I. Coluzza, *PLoS One* **2011**, *6*, e20853.
[43] I. Coluzza, *PLoS One* **2014**, *9*, e112852.
[44] T. Durek, C. F. W. Becker, *Biomol. Eng.* **2005**, *22*, 153.
[45] D. Olschewski, C. F. W. Becker, *Mol. Biosyst.* **2008**, *4*, 733.
[46] J. Chen, J. Chen, G. Pinamonti, C. Clementi, *J. Chem. Theory Comput.* **2018**, *14*, 3849.
[47] N. Combe, D. Frenkel, *Mol. Phys.* **2007**, *105*, 375.
[48] A. Irbäck, F. Sjunnesson, S. Wallin, *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 13614.
[49] D. Frenkel, B. Smit, *Understanding Molecular Simulations*, Academic Press, San Diego, London **2002**.
[50] E. D. Nelson, J. N. Onuchic, *Proc. Natl. Acad. Sci.* **1998**, *95*, 10682.
[51] L. E. Matolyak, J. K. Keum, K. M. Van De Voorde, L. T. Korley, *Org. Biomol. Chem.* **2017**, *15*, 7607.
[52] L. Feng, J. O. Iroh, *Eur. Polym. J.* **2013**, *49*, 1811.
[53] D. J. Buckwalter, M. Zhang, D. L. Inglefield, R. B. Moore, T. E. Long, *Polymer* **2013**, *54*, 4849.
[54] Y. Pang, X. Li, S. Wang, X. Qiu, D. Yang, H. Lou, *React. Funct. Polym.* **2018**, *123*, 115.