Manuscript submitted to **Biophysical** *Journal*
Article

Protein design under competition for amino acids availability

F. Nerattini[1], L. Tubiana[1], C. Cardelli[1], V. Bianco[1], C. Dellago[1], and I. Coluzza[2,*]

ABSTRACT Understanding the origin of the 20 letter alphabet of proteins is a long-lasting biophysical problem. In particular, studies focused extensively on the effect of a reduced alphabet size on the folding properties. However, the natural alphabet is a compromise between versatility and optimisation of the available resources.

Here, for the first time, we include the additional impact of the relative availability of the amino acids. We present a protein design scheme that involves the competition for resources between a protein and a potential interaction partner that, additionally, gives us the chance to investigate the effect of the reduced alphabet on protein-protein interactions. We identify the optimal reduced set of letters for the design of the protein, and we observe that even alphabets reduced down to 4 letters allow for single protein folding. However, it is only with 6 letters that we achieve optimal folding, thus recovering experimental observations.

Additionally, we notice that the binding between the protein and a potential interaction partner could not be avoided with the investigated reduced alphabets. Therefore, we suggest that aggregation could have been a driving force for the evolution of the large protein alphabet.

INTRODUCTION

The amino acid alphabet encoding the protein function is common to all living organisms and is the result of millions of years of evolution. It is composed of 20 letters, in contrast to the ones of other biopolymers, such as DNA and RNA, which possess 4 letters only. Such a large alphabet gives to proteins the vast variety of configurations and functions that we know so far.

The advent of artificial protein evolution (also known as protein design) (1–16) opens the possibility to address fundamental questions about the nature of the amino acid alphabet (17–20). One of the questions that mostly attracted the attention of the scientific community was about the universality of the 20 letters. Why 20 and not less? Could it be possible to design proteins to fold using a reduced alphabet? The early work on protein design with alphabets of different sizes was carried out for protein lattice models in which the protein chain is constrained to be on a cubic lattice. With such models it was possible to design heteropolymers with a large variety of alphabets defined by the amino acid interactions (21–30). It became rapidly apparent that even in such simplified systems it is necessary to have a minimum number of residue types to encode the target configurations (31). Moreover, such simple models allowed to explore the related question on how the alphabet size influences protein-protein interactions (32–35). Finally, works done on realistic models, offer substantial evidence that protein design with a minimalistic alphabet is possible (36–40). In particular, statistical analysis of protein databases demonstrated that a considerable fraction of the information encoded in natural proteins could be packed into smaller efficient alphabets of just 5 residue types (36, 38, 41–44). However, all the mentioned studies completely neglected the possibility that a competition for the availability of amino acids may have played a role in the evolution of the protein alphabet size.

In this work, we devised a design strategy which not only include such a competition, but also check the effect of a reduced alphabet on protein folding and protein-protein interaction. We consider systems composed of the natural protein G (PDB ID: 1PGB, already successfully redesigned with several protein models (3, 7)) and a potential binding partner (a mould of a part of protein G, that mimics with a surface-like shape a potential binding site of a larger protein). Both protein and binding partner are represented with the Caterpillar coarse-grain model, which has been successfully tested to design and refold natural and artificial proteins (7, 9). We simultaneously design the sequences of the protein-partner system according to different optimisation pathways, namely optimal folding for the protein G and just optimal interaction with the solvent for the binding partner.

The adopted design scheme leads to a competition for the available amino acids, since the protein and the surface cannot use all the 20 amino acids simultaneously for the sequences optimisation. In practice, although the whole alphabet accessible

by the binary system is still composed of 20 amino acids, the condition that we impose to the design procedure are such that the protein will have a smaller alphabet available to optimise the folding: the larger the binding surface, the stronger is its effect on the segregation.

Such a procedure allow us to control the strength of the competition as a function of the size and geometry of the binding partner. We obtain the optimal protein alphabet with the minimum number of letters, without the need of imposing the composition of the protein. Additionally, by having a protein surface system, we can explore the effect of the alphabet segregation on the aggregation in different protein-surface binding scenarios. The results show that for the folding of a small protein the minimum number of amino acid types needed is just 4. However, such a small alphabet compromises the heterogeneity of the protein-protein interactions (21, 29, 33–35) and binding cannot be avoided.

This result has interesting implications towards the understanding of the evolution of protein sequences and structure when the amino acid availability is taken into account. In fact, living systems are under constant pressure for using the smallest variety of amino acids as possible, e. g. to limit the resources needed to construct specialised tRNA molecules necessary for the translation process (45). Hence, it is reasonable to assume that during the early stages of life, the protein capable of being designed with a smaller alphabet could have been advantageous. If protein aggregation was not crucial at that stage, then our results demonstrate that protein-based life could have started with alphabets size compatible with the one of DNA and RNA. On the other hand, the simple condition of avoiding protein aggregation could be a strong driving force against alphabet segregation.

The structure of the article is the following: firstly we discuss the modelling procedure to construct a test system for protein-protein interactions that allow us to perform a protein design under competing conditions for amino acid availability. We then describe the computational method for studying design and folding of such a test system. In the central part of the article, we show the results regarding the reduced alphabets, the folding properties of the protein alone and in the presence of a binding partner. In the last part, we highlight the main conclusions of our investigation.

MATERIALS AND METHODS

Protein model

The Caterpillar protein model reduces the amino acid structure and represent it by the backbone atoms: C, O, C_α, N, H (Fig. 1(a)). The intramolecular energy for a protein of length N has the form:

$$E_{intra-P} = \sum_{i=1}^{N} \sum_{j>(i+2)}^{N} E_{hb}(\theta_1, \theta_2, r_{O_i H_j}) + E_{hb}(\theta_1, \theta_2, r_{O_j H_i}) + \sum_{i=1}^{N} \sum_{j>(i+2)}^{N} \alpha E_{sc}(r_{C\alpha_i C\alpha_j}) + \\ + \sum_{i=1}^{N} \alpha E_{HOH} E_{sol}(\Omega - \Omega_i) + \sum_{i=1}^{N} E_{bond}(r_{C_i N_i}) \tag{1}$$

where α and E_{HOH} are added to balance the relative weight of different energy terms. $E_{hb}(\theta_1, \theta_2, r_{OH})$ is a $10-12$ Lennard-Jones potential commonly used to represent hydrogen bonds (46):

$$E_{hb}(\theta_1, \theta_2, r_{OH}) = -\epsilon_H (\cos(\theta_1)\cos(\theta_2))^\nu \left[5 \left(\frac{\sigma}{r_{OH}} \right)^{12} - 6 \left(\frac{\sigma}{r_{OH}} \right)^{10} \right], \tag{2}$$

being r_{OH} the distance between the hydrogen atom of the amide group and the oxygen atom of the carboxyl group of the main chain; θ_1, θ_2 the angles between the atoms COH and OHN respectively (Fig. 1(a)), and account for the hydrogen bonds directionality; $\nu = 2$; $\sigma = 2$ Å and $\epsilon_H = 3.1 K_B T$.

$E_{sc}(r_{C\alpha_i C\alpha_j})$ mimics the side chain-side chain interaction via a smoothed square-well-like isotropic potential:

$$E_{sc}(r_{C\alpha_i C\alpha_j}) = \epsilon_{C\alpha_i C\alpha_j} \left[1 - \frac{1}{1 + \exp^{2.5(r_h - r_{C\alpha_i C\alpha_j})}} \right], \tag{3}$$

where $r_{C\alpha_i C\alpha_j}$ is the distance between the $C\alpha$ atoms and $r_h = 12$ Å. Eq. 3 is a sigmoid function with a flex at $r_h = r_{C\alpha_i C\alpha_j}$, where $E_{sc}(r_{C\alpha_i C\alpha_j}) = \epsilon_{C\alpha_i C\alpha_j}/2$. The terms $\epsilon_{C\alpha_i C\alpha_j}$ are the residue-residue interaction parameters of the interaction matrix and their value is taken from Tab. S1 of Ref. (9).

$E_{sol}(\Omega - \Omega_i)$ is an implicit solvent energy term that acts as an energy penalty if a hydrophobic (hydrophilic) amino acid is

exposed (buried), and has the form:

$$E_{sol}(\Omega - \Omega_i) = \begin{cases} \epsilon^i_{sol}[\Omega - \Omega_i] & \Omega_i \leq \Omega \\ 0 & \Omega_i > \Omega \end{cases} \quad , \Omega_i = \sum_{j=1}^{N} E_{sc}(r_{C\alpha_i C\alpha_j}), \tag{4}$$

where $\Omega = 21$ is a threshold for the number of contacts in the native structure above which the amino acid is considered to be fully buried, and ϵ^i_{sol} is the Dolittle hydrophobicity index (47).

$E_{bond}(r_{CN}) = k(r_{CN} - r_{CN_{ref}})^2$ is a harmonic bonding term with elastic constant $k = 20\ K_B T \text{Å}^{-2}$, that keeps fixed the distance r_{CN}, along with the $C\hat{C_\alpha}N$ backbone angle.
For more details about the model see Ref. (7, 9). For the binding site amino acids, instead, we use a single atom representation and make use of the E_{sol} and E_{sc} terms only.

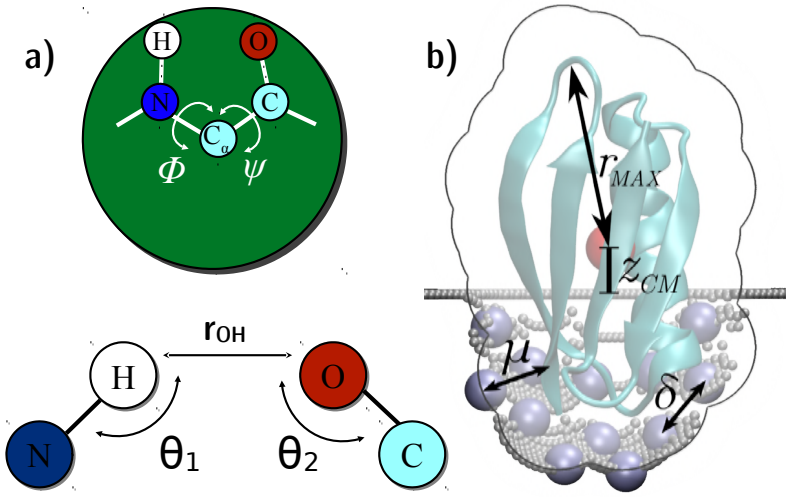


Figure 1: **a)** Caterpillar protein model: the green circle represents the amino acid self-avoidance volume, which has a radius of 2 Å and is centred on the position of the C_α atom. Each amino acid is represented through backbone atoms only. Side chain-side chain interactions are represented via a square-well-like potential (Eq. 3).For the hydrogen bonds, the model include a $10-12$ Lennard-Jones potential, that is a function of the distance r_{OH}, and tunes it with a multiplicative factor that involves the angles θ_1 and θ_2, so to account for the directionality of the bond. **b)** Artificial binding site parameters. Gray dots are self-avoiding beads; blue spheres are C_α atoms of the binding site. The red spot represents the centre of mass (CM) of the protein, z_{CM} is its height with respect to the $z = 0$ plane and r_{MAX} is the maximum CM-C_α distance. μ is the minimum C_α protein-C_α surface distance, i.e. the value one has to use to inflate the protein for the binding site shaping. δ is the minimum distance between two activated C_α of the binding site.

Binding site model

The binding site is modelled as a fixed layer of amino acids, idealizing a small pocket on the surface of a much larger globular protein. To represent the artificial binding site, we generate a mould by pushing the protein on a mesh of self-avoiding beads initially lying flat on the $z = 0$ plane, as illustrated in Fig. 1(b). The high density of beads in the mesh prevents the protein from passing through it. A finite number of C_αs, the type of which is assigned during the design procedure, are distributed on the surface binding site, to model its protein nature. We refer to the latter atoms as $C_{\texttt{Surf}}$.

There are three important parameters in the modelling procedure: ζ, the reduced centre of mass (CM) height with respect to the $z = 0$ plane; μ, the minimum C_α protein-C_α surface distance; and δ, the gap between two C_αs of the binding site. To define the reduced height ζ, we first identify r_{MAX}, i.e. the maximum protein radius with respect to the CM ($r_{MAX} \sim 16$ Å for protein G), and then use it to normalise the height z_{CM} of the CM from the $z = 0$ plane: $\zeta = \frac{z_{CM}}{r_{MAX}}$. Hence, we push the protein into the $z = 0$ plane until we reach the desired ζ value. The mesh is pushed downwards accordingly, keeping every point at the minimum distance between all atoms of the protein and the binding site at $\mu = 13$ Å (see Fig. 1(b)). The value $\mu = 13$ Å ensures a low influence of the binding site on the protein design regarding the protein-binding site energy E_{sc}

(Eq. 3). We construct systems with different ζ values, and for each one of them, we perform a rotational analysis by keeping the position of the CM fixed and by shaping the mesh for various orientations of the protein until we reach the maximum surface area of the binding site. It is important to notice that the surface area decreases with the increase of ζ.

Having obtained the geometrical shape of the pocket, we proceed to "activate" a subset of its beads by assigning to them a C_α nature. These beads constitute the C_{surf} set. The activated beads are homogeneously distributed in the pocket and are always separated by a distance of $\delta = 5$ Å from each other. The value of δ is derived from the typical nearest distance between two residues in natural proteins, as explained in Fig. S8 of the Supplementary Information (SI).

Since the amino acids in the $C_{\mathtt{Surf}}$ set are represented by C_α atoms, the protein-binding site interaction energy includes the Caterpillar E_{sc} and E_{sol} terms only (see Eqs. 3 and 4). Given that our binding site representation is limited to the surface residues, such an approximation affects the correct evaluation of the binding site solvent exposure term E_{sol}. Since in turns E_{sol} will influence the protein binding properties, we add an offset to the number of contacts for each amino acid of the binding site (~ 6), to correctly account for the solvent exposure term E_{sol} of the binding site amino acids. The offset was calculated in such a way that the maximal amino acid exposure is compatible to the one of natural proteins (e.g. the protein G itself described in Fig. S10 of the SI). Finally, since the binding site conformation is fixed, we neglect all the interactions between all the pairs of the C_{surf} set.

For more information about the algorithms used in the binding site modelling procedure, see the section *Binding site modelling* of the SI.

Design

To investigate the sequence space, we perform Virtual Move Parallel Tempering (48) (VMPT) Monte Carlo simulations with swap and single point amino acid mutation moves along the sequence, a procedure that has been already successfully employed in protein design (7, 9, 16). We simulate the same system at different temperatures, and swap sequences between the replicas on the fly, thus enhancing the overcoming of energy barriers and, therefore, improving the sampling. Moreover, at each temperature, we collect statistics using the information coming from all other replicas, according to the virtual move scheme described in Ref. (48). In our implementation we simulate 16 replicas with a set of temperatures (10.000; 5.000; 2.000; 1.000; 0.500; 0.333; 0.250; 0.200; 0.167; 0.143; 0.125; 0.111; 0.100; 0.091; 0.083; 0.077) in units of K_B. We also consider the protein-binding site as a single object, frozen in the target conformation generated via the binding site modelling described above. The sequences of the protein and the binding site are optimised jointly. The best candidate sequences for the folding are the ones which minimise the energy of the target structure and maximise the number of permutations N_p, given by:

$$N_p = \frac{N!}{\prod_{i=1}^{q} n_i!}, \tag{5}$$

to increase the composition heterogeneity along the joint protein-binding site sequence. In Eq. (5) $q = 18$ is the alphabet size (proline and cysteine are not included in the design, due to their peculiar role in protein structure, which is beyond the scope of our model); N is the total number of monomers in the protein-binding site system; n_i is the number of monomers of type i in the joint sequence.

We enhance the sampling by introducing an adaptive bias potential $W[E, \ln N_p, T]$ over two collective variables E and N_p. $W[E, \ln N_p, T]$ is used to bias the acceptance probability of each Monte Carlo mutation move. Therefore, each Monte Carlo step is accepted with a probability

$$P_{acc}^{rep} = min\{1, \exp\left[\Delta W - (\Delta E - E_p \ln \frac{N_p^{new}}{N_p^{old}})/K_B T\right]\} \tag{6}$$

or

$$P_{acc}^{swap} = min\{1, \exp\left[\Delta W - \Delta E/K_B T\right]\}, \tag{7}$$

associated to the amino acid replacement and amino acids swap moves, respectively. In the latter equations ΔW, ΔE and $\ln \frac{N_p^{new}}{N_p^{old}}$ refer to the differences of the bias potential, the energy and the number of permutation between the new configuration and the old one; $E_p = 20\ K_B T$ is a scaling factor set to a value high enough to generate highly heterogeneous sequences.

Folding

We adapted the Monte Carlo folding procedure for a single protein, described in Ref. (7, 9), to handle a system composed of interacting partners. The simulation is performed in a cubic box (as shown in Fig. S11) containing two replicas of the binding

site (box side $\sim$ 360 Å), that are the mirror image of the other with respect to the xy plane passing through the centre of the box.

The binding sites are frozen in the box, and the protein starts the folding from a fully stretched conformation. An impenetrable slab region is defined between the binding sites to prevent the protein to approach them from the convex side instead of the concave one generated with the moulding procedure.

We sample the protein conformations using crankshaft, pivot, rotation, translation and mirroring moves, to let the protein reorient, diffuse in the box and switch from right to left handed conformations. The simulation is performed in parallel at 32 different temperatures and the replicas exchange information through the VMPT bias scheme (48). The set of reduced temperatures (8.5; 7.8; 7.2; 6.6; 6.0; 5.4; 4.9; 4.6; 4.3; 3.9; 3.5; 3.1; 2.8; 2.55; 2.35; 2.2; 2.05; 1.9; 1.75; 1.6; 1.45; 1.3; 1.1; 1.0; 0.98; 0.95; 0.92; 0.88; 0.85; 0.82; 0.79; 0.76) is chosen so to observe the protein repeatedly detaching from the binding site.

Each replica sampling is enhanced by a bias potential depending on two collective variables, namely the distance root mean square displacement within the protein $DRMSD_{intra}$, and between protein and binding site, $DRMSD_{inter}$. The $DRMSD$ is a measure of the distance of a configuration from a target structure:

$$DRMSD = \sqrt{\frac{1}{C}\sum_{ij}(|\Delta\vec{r_{ij}}| - |\Delta\vec{r_{ij}}^{\,T}|)^2}, \tag{8}$$

where the sum runs over the ij pairs in contact in the target structure (namely the native contacts, identified using a cut off of 17 Å according to previous studies (7, 9)), $\Delta\vec{r_{ij}}$ is the distance between the residues i and j belonging to the same ($DRMSD_{intra}$) or different ($DRMSD_{inter}$) proteins, and $\Delta\vec{r_{ij}}^{\,T}$ is the same distance calculated over the target structures. Accordingly, the $DRMSD_{intra}$ is computed adopting the native and isolated protein conformation as a target structure, while the $DRMSD_{inter}$ is computed using as a target structure the native protein bound to the binding site. The normalisation factor C is the number of native contacts for $DRMSD_{inter}$ and twice the number of native contacts for $DRMSD_{intra}$. $\Delta\vec{r_{ij}}$ for $DRMSD_{inter}$ is evaluated with respect to the binding site closest to the protein for each ij pair.

RESULTS AND DISCUSSION

The purpose of the present work is to design proteins with the optimum reduced amino acid alphabet, to check both their folding abilities and their aggregation behaviour in the presence of a possible binding site. To this aim, we chose a value of $\mu = 13$ Å that minimises the optimisation of the protein-binding site interaction (see section *Binding site model*), thus leading to the design problem described in the following.

The globular structure of the protein leads to a complex design optimisation problem since every residue is in contact with many others and the water exposure profile varies from the surface to the buried residues. On the other hand, the residues of the binding site possess few contacts; these weakly interacting residues are mainly optimised for the exposure to the solvent. Since the coupling between these two optimisation procedures is only through Eq. (5), i.e. the condition of maximizing the composition heterogeneity, the design leads to the spontaneous segregation of a reduced alphabet on the protein sequence, keeping the variability high on the one of the binding site.

We investigate four systems differing in terms of ζ=(0.20, 0.40, 0.60, 0.80), thus leading to a surface area=(4717.5, 3842.2, 3051.5, 2320.5) Å^2 and $C_{\mathtt{Surf}}$=(158, 127, 100, 78) residues respectively. For all scenarios, we generated a large number of sequences that we group in solution basins. In Fig. 2 we plot the distribution of the Hamming distances calculated for all the possible pairs between the basins. The latter is measured using the Hamming distance that determines the difference between two sequences of equal length by counting the number of substitutions needed to make them coincide. It is often more convenient to normalise the Hamming distance by dividing by the protein length N. For all scenario we observe large number of solutions that differ for more than 50% of the residues. However, ζ= 0.60 and 0.80 have also a significant number of related solutions with sequences differing for 20% only (see Fig. 2(f)). To check that nevertheless the ζ= 0.60 and ζ= 0.80 basins are distinct, we also calculated the self-overlap distributions for ζ= 0.60 and ζ= 0.80 (see Fig. S9 in the SI) and we obtained profiles that are different from each other. More importantly, they differ from the one plotted in Fig. 2(f), demonstrating that although there are similarities, the two basins are distinct.

A crucial feature of the distributions in Fig. 2, are the peaks at high Hamming distances. They demonstrate that the protein-binding site coupling induces the design process to explore distinct solution basins. Given that protein and binding site are energetically very weakly coupled (evident from the low value of E_{sc}-inter that ranges from -1.4 to -2.7 K_BT), the influence of the binding site can be attributed to the maximisation of the letter permutations expressed by Eq. (5). Such coupling enforces an increasing competition between binding site and protein for the alphabet available while increasing the binding site size.

ζ	**0.20**	**0.40**	**0.60**	**0.80**
0.20	0	61	87	86
0.40		0	84	86
0.60			0	27
0.80				0

Table 1: Hamming distance [%] between designed protein sequences.

ζ **0.20**	SGGGGGGRKKKKKYYYVVVVVVGSSGGVKKKKKGGNYYYYYYYYVYVVKKKKGGGR
ζ **0.40**	FGGGGGGKKRRKVYYVKKKRRRGFFYYYYKKRRWRNFYYYYVVKKKKVYYGGGGGG
ζ **0.60**	FFFFGDGGYYYYYKKKKHHHHIRRRRRRFYYYYGGGRKKKKKKKYHHHRFFFGGGG
ζ **0.80**	FFFFFGGGYYYYVKKKHHHHHHKIIRRRRRFYYYGGGTKKLKKKYHHHRRRFFGGGG

Table 2: Protein G designed sequences for the investigated systems.

In Fig.3 we plot the occurrence frequency of each amino acid type in four sequences selected with highest permutation number and lowest energy among the ones in the solution basins (shown in Tab. 2 and Tab. 1).

Given that the Caterpillar alphabet is composed by 18 letters, the frequency of each amino acid in a maximally heterogeneous sequence would be $1/18 \sim 6\%$. Therefore, we set a threshold at a slightly higher value=10%, to safely discern the contribution of dominating amino acids from the random occurrences. From the plot it becomes apparent that the protein designed sequence has a restricted composition of amino acids leading to a segregated effective protein alphabet. We observe that the effective alphabet grows from 4 to 6 letters going from larger (ζ=0.20 and 0.40) to reduced binding sites (ζ=0.60 and 0.80) respectively. It is interesting to notice that the alphabets are made of amino acids with an average attractive pair-interaction energy and high variability in terms of the residue-solvent interactions (see Table S1 in Ref (9)). Moreover, the alphabets differ from each other, and for each scenario, the protein amino acids are not present in the corresponding binding sites sequence (see Fig. 3). The latter finding shows that the design process indeed mimics a process under competition for available amino acids.

To test the selected sequences, we first examine the folding stability of the protein itself, therefore performing a folding simulation in an empty box starting from a fully stretched configuration. The set of reduced temperatures for these simulations (2.0; 1.8; 1.6; 1.4; 1.3; 1.2; 1.1; 1.0; 0.9; 0.8; 0.7; 0.65; 0.6; 0.55; 0.5; 0.45) is chosen in order to observe repeatedly folding events. Fig. 4 shows the free energy profiles as a function of $DRMSD$. From previous works (7, 9), the criterion for assessing a stable fold is to observe a funnel shape of the free energy profile and a global free energy minimum for $DRMSD \leq 2$ Å. Using this criterion, we can say that all protein sequences fold back into the target configuration, although with different precision. Sequences with a larger effective alphabet fold with higher precision, as can be seen from the $DRMSD$ value of the configurations corresponding to the global free energy minimum for each system (right hand side of Fig. 4): $DRMSD = 2.1$ Å for $\zeta = 0.20$; $DRMSD = 1.9$ Å for $\zeta = 0.40$; $DRMSD = 1.3$ Å for $\zeta = 0.60$ and $DRMSD = 1.5$ Å $\zeta = 0.80$. [1] The sequence optimised for the binding site $\zeta = 0.40$ shows a secondary minimum in the free energy, corresponding to misfolded compact structures, therefore being the system less stable for the folding in the bulk.

From the described scenario, we can draw two important conclusions: firstly, design with a limited alphabet of 4 letters can produce a funnel-like folding free energy landscape; secondly, with 6 letters we recover the folding precision of previous Caterpillar designs made with 20 letters (9). Our results are consistent with the experimental observation that 6 letters are a minimal set necessary to maintain protein structure and function (36, 38, 41–44).

From the Random Energy Model (49–51) we know that for a heteropolymer to be designable it has to satisfy the relation $q > \exp(\omega)$, where q is the alphabet size and ω the conformational entropy per residue. Hence, a 4 letters alphabet gives an upper bound to the conformational entropy ω of the Caterpillar backbone and therefore of the more restricted natural protein backbones. Such a result is compatible with the recent observations of Cardelli *et al.* (52) who mapped the designability phase space for a general heteropolymer decorated with directional interactions similar to the hydrogen bonds present along the protein backbone. For polymers with two directional interactions per particle the minimum alphabet measured was three, i.e. close to the one presented here.

To test the effect of the alphabet restriction on protein-protein interactions, we perform folding simulations in the presence of the binding site. In Fig. 5 we plot the free energy landscape as a function of $DRMSD_{intra}$ using the native protein G as target configuration, and $DRMSD_{inter}$ using the folded bound configuration as a target. This choice allows us to monitor the folding and binding properties of the system independently. Conformations that are folded and bound can be found in the

[1] The DRMSD values correspond to 4.9; 5.5; 2.4 and 2.7 Å in $RMSD$ respectively.

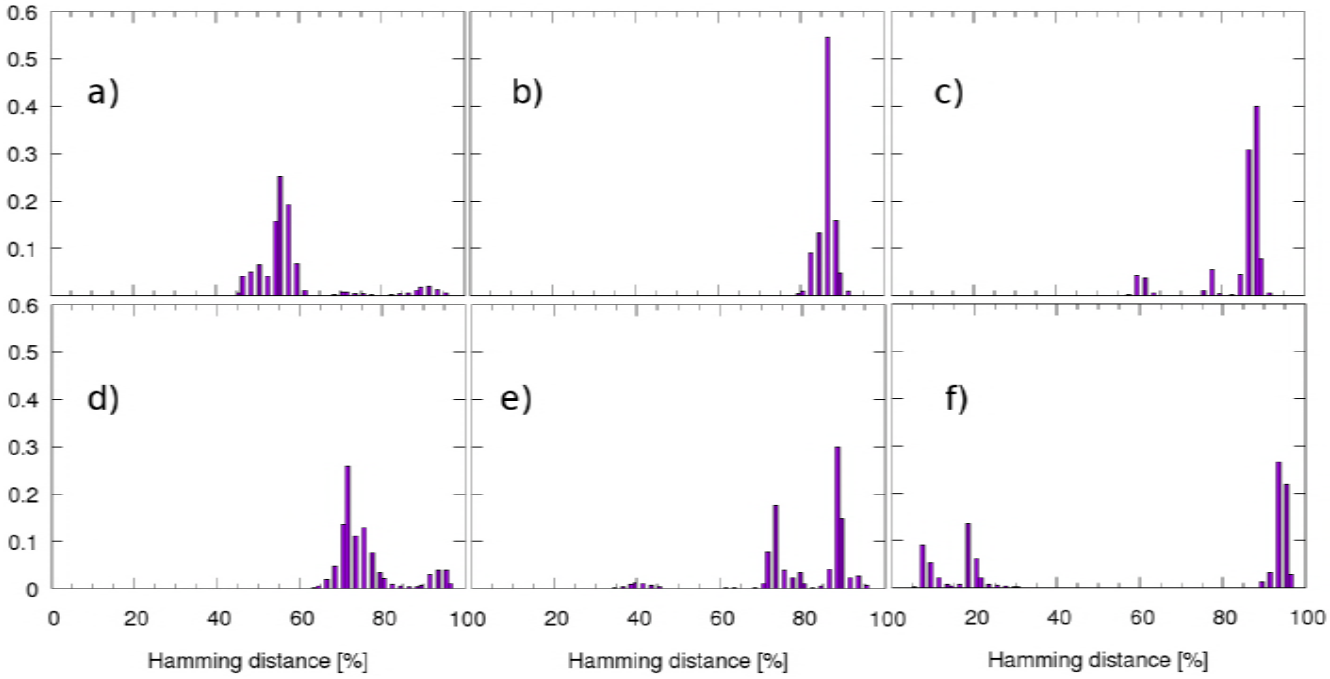


Figure 2: Histograms obtained by evaluating the Hamming distance (% relative to the chain length) between all possible pairs of sequences chosen selecting 200000 solutions around the design free energy minimum of systems *A* and *B*, corresponding to the ζ values specified in the following. a) *A*: $\zeta = 0.20$; *B*: $\zeta = 0.40$ b) *A*: $\zeta = 0.20$; *B*: $\zeta = 0.60$ c) *A*: $\zeta = 0.20$; *B*: $\zeta = 0.80$ d) *A*: $\zeta = 0.40$; *B*: $\zeta = 0.60$ e) *A*: $\zeta = 0.40$; *B*: $\zeta = 0.80$ f) *A*: $\zeta = 0.60$; *B*: $\zeta = 0.80$

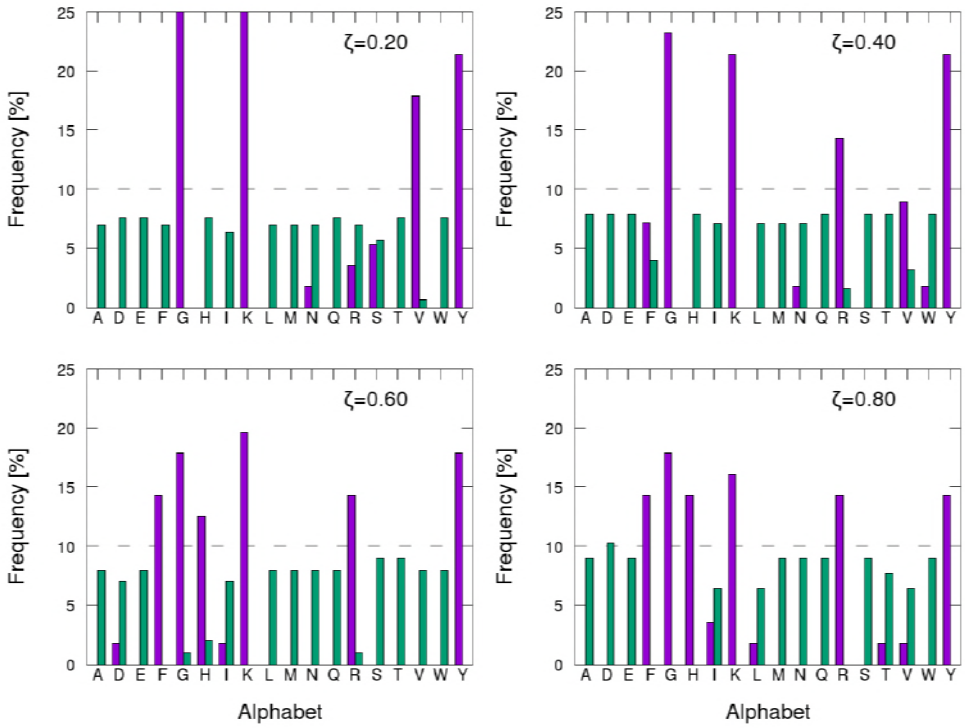


Figure 3: Frequency of amino acids in the designed sequences. The grey dotted line is the value used to consider one amino acid in the effective alphabet. Purple bars refer to the sequence of protein G, while green bars to the sequences of the binding site. Cysteine (C) and Proline (P) are not included in the design alphabet.

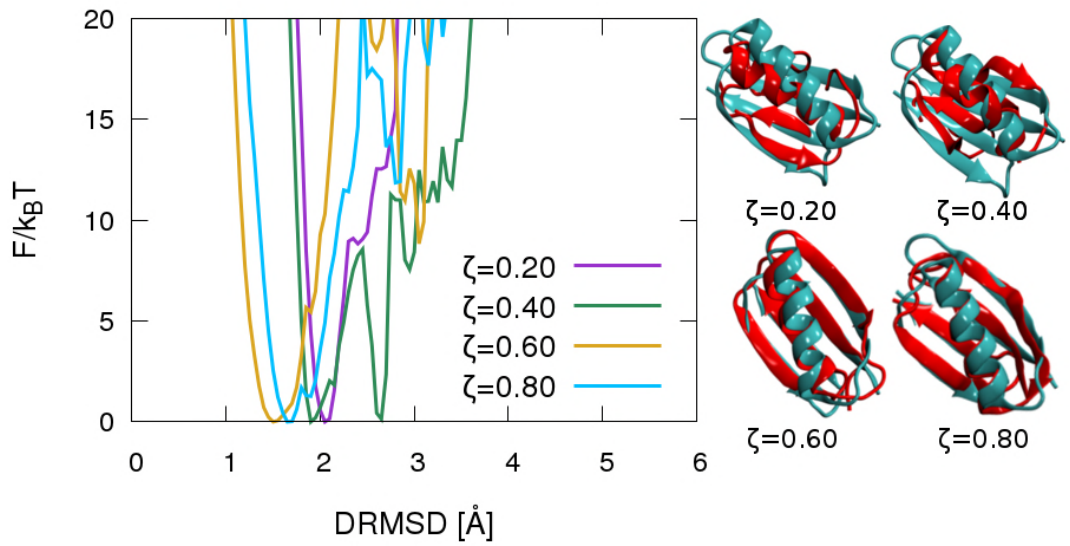


Figure 4: Folding free energy profiles F/k_BT of single protein (no binding site) at reduced temperature 0.55 as a function of DRMSD from the native target structure (Protein G, PDB ID: 1PGB). Different colours represent the folding free energy of the protein sequence obtained via the design procedure in the presence of the binding site characterised by the ζ value specified in the key. Configurations corresponding to the free energy minimum for each system are represented in red, compared to the native protein G (in green).

bottom left corner, while folded unbound ones in the top left one.

Additionally, we also check the free energy profiles as a function of $DRMSD_{intra}$ for bound conformations (see Fig. 6) and in bulk solution where no protein-site contacts are possible (see Fig. 7. For a sketch of the definition of interacting and bulk solution configurations see Fig. S11 in the SI. We have also verified that the free energy profiles of configurations in the latter region correctly fold into the target structure (Fig. 7), consistently to what we observe in the isolated protein folding simulations.

For all the scenarios, upon binding to the site, we observe a significant enhancement of the stability of the misfolded configurations with respect to the bulk solution (compare Fig.s 5,6(a) and 7). In particular, there is a considerable shift in the equilibrium towards states at $DRMSD \sim 3$ Å that are now comparable in free energy to the target configurations. This effect is due to the small protein alphabet that cannot now encode heterogeneous patterns to reduce the binding interaction with the large binding sites. It should be noted that natural binding sites expose much smaller surface areas then the one modelled here. Hence, it might be that the effect is mitigated when real protein binding sited are considered.

Analysing the behaviour of the binding process as a function of temperature we find that the random binding is so strong that the van't Hoff curve (53, 54) shows an exothermic process above the folding temperature with positive binding affinity (Supplementary Fig. S12; for details about the evaluation of the association constant see Fig. S11). At the same time the equilibrium shifts from partially-misfolded to fully-misfolded, indicating that the unfolding process takes place at the surface while the protein remains bound (see Fig. 6(b)). Considering the large surface area and the small alphabet employed it is not surprising that specificity of the protein-protein interactions could not be achieved. This is an essential factor that could explain the size of natural alphabets, i.e. the 20 amino acids. The study of the effect of the size of the binding site on the specificity is the object of upcoming work.

CONCLUSION

The design procedure employed in our work has a significant segregation effect on the alphabet letters used in the protein sequence. The larger the number of residues on the binding site, the smaller is the effective alphabet available for the protein sequence. On the one side, the design is capable of selecting a subset of letters that still allows the folding of the protein in the bulk solution even for the smallest effective alphabet (4 letters). The precision of the folding increases with the effective alphabet size. Interestingly, the experimentally determined alphabet size of 6 letters is also what we observed recovering the design accuracy commonly obtained with a 20 letter alphabet. Finally, the identified alphabets in this work are automatically optimised by the design and are composed of the most attractive residues while maintaining the broadest heterogeneity of the

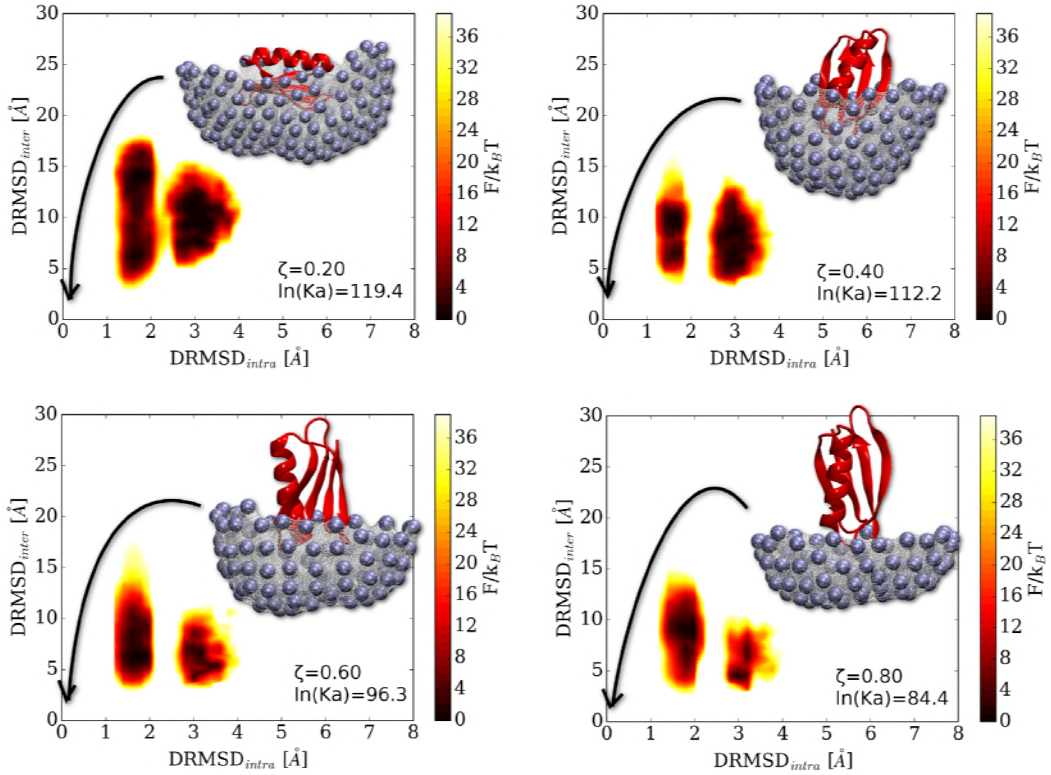
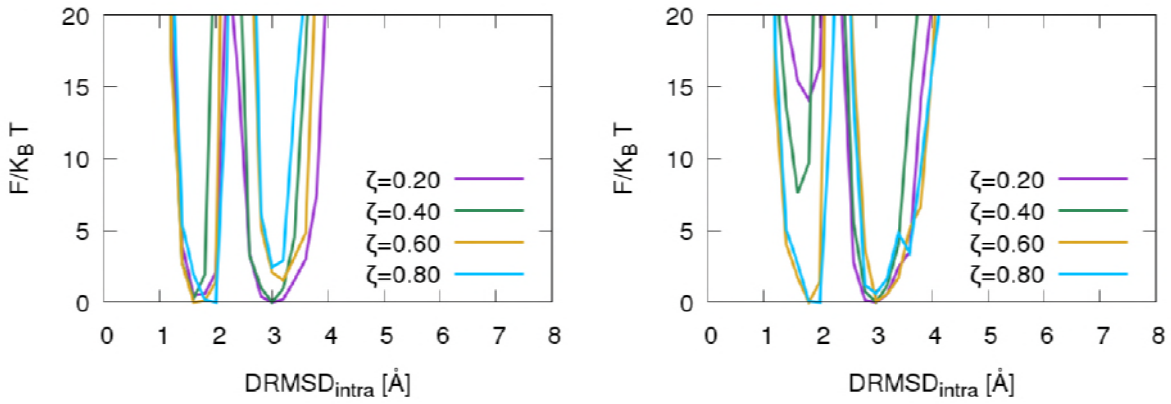


Figure 5: Folding free energy landscapes F/k_BT at reduced temperature 0.76 as a function of $DRMSD_{intra}$ protein from the native protein G as target and $DRMSD_{inter}$ protein-binding site from the folded protein bound to the binding site (configurations depicted in the panels). The binding affinity decreases along with the binding site size, as shown by the value of the association constants K_a in the plot key.



(a) (b)

Figure 6: Folding free energy profiles F/k_BT at reduced temperature 0.76 (left panel) and 1.00 (right panel) as a function of DRMSD intra protein from the native target structure. Simulations of protein G in presence of binding site at $\zeta = (0.20, 0.40, 0.60, 0.80)$. The curves have been evaluated on bound configurations, i.e. where the protein is directly interacting with the binding site. Left panel: The presence of two equivalent minima suggests that the protein binds in the target configuration as well as in a misfolded state with the same probability. Right panel: upon increasing the temperature, the correctly folded state is destabilised for large binding sites.

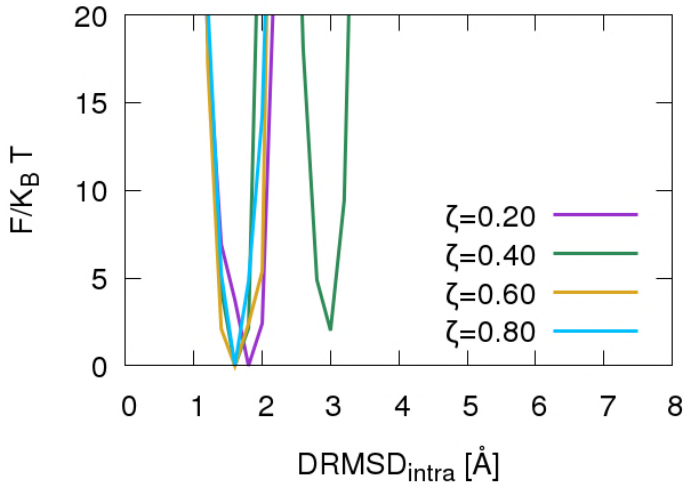


Figure 7: Folding free energy profiles F/k_BT at reduced temperature 0.76 as a function of $DRMSD_{intra}$ protein from the native target structure. Simulations of protein G in presence of binding site at $\zeta = (0.20, 0.40, 0.60, 0.80)$. The curves have been evaluated on configurations in the bulk solution, where the protein-binding site distance is larger than the interaction one. The presence of a minimum below $DRMSD$ =2 Å shows that the protein is folded when bound.

solvent interactions.

Our results have far-reaching implications both in the field of protein design and for the understanding of protein evolution. In protein design, the possibility of using a reduced alphabet would considerably accelerate the search of the sequence space for good folders. In the field of protein evolution instead, the understanding of the smallest alphabet necessary for accurate proteins design is still an open question. To the best of our knowledge, this study represents the first successful design of a full

natural protein structure with a reduced alphabet of just 4 letters.

AUTHOR CONTRIBUTIONS

I.C. and C.D. designed the research; F.N. and L.T. developed the simulation tools; F.N. performed the simulations and the data analysis. All the authors discussed the research and participated in the writing and the revision of the article.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Gutin, A. M., and E. Shakhnovich, 1993. Ground-state of random copolymers and the discrete Random Energy-model. *The Journal of Chemical Physics* 98:8174–8177. papers2://publication/uuid/AEF0E5B1-DB7F-4FEB-9367-BFD43B38854D.

2. Dahiyat, B. I., and S. Mayo, 1997. De Novo Protein Design: Fully Automated Sequence Selection. *Science* 278:82–87. http://www.sciencemag.org/cgi/doi/10.1126/science.278.5335.82.

3. Koehl, P., and M. Levitt, 1999. De novo protein design. I. In search of stability and specificity. *Journal of molecular biology* 293:1161–81. http://www.ncbi.nlm.nih.gov/pubmed/10547293.

4. Kortemme, T., and D. Baker, 2004. Computational design of protein-protein interactions. *Current Opinion in Chemical Biology* 8:91–97.

5. Fung, H. K., W. J. Welsh, and C. A. Floudas, 2008. Computational de novo peptide and protein design: Rigid templates versus flexible templates. *Industrial and Engineering Chemistry Research* 47:993–1001. http://pubs.acs.org/doi/abs/10.1021/ie071286k.

6. Samish, I., C. Macdermaid, J. Perez-Aguilar, and J. Saven, 2011. Theoretical and computational protein design. *Annual Review of Physical Chemistry* 62:129–149. papers://ae875177-834e-4ba8-8523-120292c79891/Paper/p4643.

7. Coluzza, I., 2011. A Coarse-Grained approach to protein design: Learning from design to understand folding. *PLoS ONE* 6:e20853. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3128589&tool=pmcentrez&rendertype=abstract.

8. Koga, N., R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, and D. Baker, 2012. Principles for designing ideal protein structures. *Nature* 491:222–227. http://dx.doi.org/10.1038/nature11600%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3705962&tool=pmcentrez&rendertype=abstracthttp://dx.doi.org/10.1038/nature11600http://www.nature.com/doifinder/10.1038/nature11600http://www.ncbi.nlm.nih.

9. Coluzza, I., 2014. Transferable Coarse-Grained Potential for De Novo Protein Folding and Design. *PLoS ONE* 9:e112852. http://dx.plos.org/10.1371/journal.pone.0112852http://www.ncbi.nlm.nih.gov/pubmed/25436908.

10. Thomson, A. R., A. R. Thomson, C. W. Wood, A. J. Burton, G. J. Bartlett, R. B. Sessions, R. L. Brady, and D. N. Woolfson, 2014. Computational design of water-soluble alpha-helical barrels. *Science* 346:485–488. http://www.sciencemag.org/cgi/doi/10.1126/science.1257452.

11. Sevy, A. M., T. M. Jacobs, J. E. Crowe, and J. Meiler, 2015. Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLoS Computational Biology* 11.

12. Pelay-Gimeno, M., A. Glas, O. Koch, and T. N. Grossmann, 2015. Structure-Based Design of Inhibitors of Protein-Protein Interactions: Mimicking Peptide Binding Epitopes. *Angewandte Chemie - International Edition* 54:8896–8927.

13. Chevalier, A., D.-A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K.-H. Lam, G. Yao, C. D. Bahl, S.-I. Miyashita, I. Goreshnik, J. T. Fuller, M. T. Koday, C. M. Jenkins, T. Colvin, L. Carter, A. Bohn, C. M. Bryan, D. A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I. A. Wilson, D. H. Fuller, and D. Baker, 2017. Massively parallel de novo protein design for targeted therapeutics. *Nature* 550:74–79. http://www.nature.com/doifinder/10.1038/nature23912.

14. Marcos, E., B. Basanta, T. M. Chidyausiku, Y. Tang, G. Oberdorfer, G. Liu, G. V. T. Swapna, R. Guan, D.-A. Silva, J. Dou, J. H. Pereira, R. Xiao, B. Sankaran, P. H. Zwart, G. T. Montelione, and D. Baker, 2017. Principles for designing proteins with cavities formed by curved β sheets. *Science* 355:201–206. http://www.sciencemag.org/lookup/doi/10.1126/science.aah7389.

15. Coluzza, I., 2017. Computational protein design: a review. *Journal of Physics: Condensed Matter* 29:143001. http://iopscience.iop.org/10.1088/1361-648X/aa5c76http://iopscience.iop.org/article/10.1088/1361-648X/aa5c76http://stacks.iop.org/0953-8984/29/i=14/a=143001?key=crossref.6b75a4256c5bfe8a8e20e6ef3834f61e.

16. Bianco, V., G. Franzese, C. Dellago, and I. Coluzza, 2017. Role of Water in the Selection of Stable Proteins at Ambient and Extreme Thermodynamic Conditions. *Physical Review X* 7:21047. http://link.aps.org/doi/10.1103/PhysRevX.7.021047.

17. Davidson, A. R., and R. T. Sauer, 1994. Folded proteins occur frequently in libraries of random amino acid sequences. *Proceedings of the National Academy of Sciences* 91:2146–2150. http://www.pnas.org/cgi/doi/10.1073/pnas.91.6.2146.

18. Riddle, D. S., J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker, 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nature Structural Biology* 4:805–809.

19. Cordes, M. H. J., A. R. Davidson, and R. T. Sauer, 1996. Sequence space, folding and protein design. *Current Opinion in Structural Biology* 6:3–10.

20. Davidson, A. R., K. J. Lumb, and R. T. Sauer, 1995. Cooperatively folded proteins in random sequence libraries. *Nature Structural Biology* 2:856. http://dx.doi.org/10.1038/nsb1095-856http://10.0.4.14/nsb1095-856.

21. Coluzza, I., and D. Frenkel, 2004. Designing specificity of protein-substrate interactions. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 70:51917. papers2://publication/uuid/1793B4D4-B30F-462A-8338-FFCCE44D45FBhttp://www.ncbi.nlm.nih.gov/pubmed/15600666.

22. Coluzza, I., H. G. Muller, and D. Frenkel, 2003. Designing refoldable model molecules. *Physical Review E* 68:46703. http://www.ncbi.nlm.nih.gov/pubmed/14683075.

23. Salvi, G., S. Mölbert, and P. De Los Rios, 2002. Design of lattice proteins with explicit solvent. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 66:061911. http://link.aps.org/doi/10.1103/PhysRevE.66.061911.

24. Wang, T. R., J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, 2000. Symmetry and designability for lattice protein models. *Journal of Chemical Physics* 113:8329–8336. papers2://publication/uuid/649C601D-29E5-4E4E-A9C6-4D8DD727B59D.

25. Deutsch, J. M., and T. Kurosky, 1995. A New Algorithm for Protein Design. *Physical Review Letters* 76:10. papers2://publication/uuid/7A4B228D-87BB-4761-BF89-4DA9434F9875%5Cnhttp://arxiv.org/abs/cond-mat/9508127.

26. Shakhnovich, E. I., and a. M. Gutin, 1993. Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences of the United States of America* 90:7195–7199. papers2://publication/uuid/E46F4571-68E5-4675-835F-37D95A099CBBhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=47103&tool=pmcentrez&rendertype=abstract.

27. Yue, K., and K. a. Dill, 1992. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences of the United States of America* 89:4163–4167. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=525653&tool=pmcentrez&rendertype=abstract.

28. Bryngelson, J. D. D., and P. G. G. Wolynes, 1987. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 84:7524–7528. http://www.pnas.org/content/84/21/7524.abstract%5Cnhttp://www.pnas.org/content/84/21/7524.shorthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=299331&tool=pmcentrez&rendertype=abstract.

29. Coluzza, I., and D. Frenkel, 2007. Monte Carlo study of substrate-induced folding and refolding of lattice proteins. *Biophysical journal* 92:1150–6. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1783883&tool=pmcentrez&rendertype=abstract.

30. Abeln, S., and D. Frenkel, 2008. Disordered flanks prevent peptide aggregation. *PLoS Computational Biology* 4:e1000241. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2588114&tool=pmcentrez&rendertype=abstract.

31. Chan, H. S., and K. A. Dill, 1996. Comparing folding codes for proteins and polymers. *Proteins: Structure, Function and Genetics* 24:335–344.

32. Sear, R. P., and J. a. Cuesta, 2003. Instabilities in Complex Mixtures with a Large Number of Components. *Physical Review Letters* 91:245701. http://link.aps.org/doi/10.1103/PhysRevLett.91.245701.

33. Sear, R. P., 2004. Specific proteinprotein binding in many-component mixtures of proteins. *Physical Biology* 1:53–60. http://www.ncbi.nlm.nih.gov/pubmed/16204822http://stacks.iop.org/1478-3975/1/i=2/a=001?key=crossref.d96ef7403afd02b733604b01480fd55f.

34. Sear, R. P., 2004. Highly specific protein-protein interactions, evolution and negative design. *Physical Biology* 1:166–172. papers2://publication/doi/10.1088/1478-3967/1/3/004http://stacks.iop.org/1478-3975/1/i=3/a=004?key=crossref.b6b2c5fab9a06a926f2683d352e1125a.

35. Madge, J., and M. A. Miller, 2015. Design strategies for self-assembly of discrete targets. *The Journal of Chemical Physics* 143:044905. http://aip.scitation.org/doi/10.1063/1.4927671.

36. Plaxco, K. W., D. S. Riddle, V. Grantcharova, and D. Baker, 1998. Simplified proteins: Minimalist solutions to the 'protein folding problem'. *Current Opinion in Structural Biology* 8:80–85.

37. Walter, K. U., K. Vamvaca, and D. Hilvert, 2005. An active enzyme constructed from a 9-amino acid alphabet. *Journal of Biological Chemistry* 280:37742–37746.

38. Reetz, M. T., and S. Wu, 2008. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chemical Communications* 5499. http://xlink.rsc.org/?DOI=b813388c.

39. Liu, B., J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, and K. C. Chou, 2014. IDNA-Prot|dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS ONE* 9.

40. Sun, Z., R. Lonsdale, X. D. Kong, J. H. Xu, J. Zhou, and M. T. Reetz, 2015. Reshaping an Enzyme Binding Pocket for Enhanced and Inverted Stereoselectivity: Use of Smallest Amino Acid Alphabets in Directed Evolution. *Angewandte Chemie - International Edition* 54:12410–12415.

41. Chan, H. S., 1999. Folding alphabets. *Nature Structural Biology* 6:994–996.

42. Wang, J., and W. Wang, 1999. A computational approach to simplifying the protein folding alphabet. *Nature Structural Biology* 6:1033–1038.

43. Murphy, L. R., A. Wallqvist, and R. M. Levy, 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering, Design and Selection* 13:149–152. https://academic.oup.com/peds/article-lookup/doi/10.1093/protein/13.3.149.

44. Solis, A. D., 2015. Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins. *Proteins: Structure, Function and Bioinformatics* 83:2198–2216.

45. Alberts et al, B., 2007. Molecular Biology of the Cell. Garland Science. http://en.wikipedia.org/wiki/Molecular_Biology_of_the_Cell_(textbook).

46. Irbäck, A., F. Sjunnesson, and S. Wallin, 2000. Three-helix-bundle protein in a Ramachandran model. *Proceedings of the National Academy of Sciences of the United States of America* 97:13614–13618. http://www.pnas.org/content/97/25/13614.abstract%5Cnpapers2://publication/uuid/4E7F0DC0-5043-462B-A318-3A961CB2CF5E%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=17624&tool=pmcentrez&rendertype=abstract.

47. Garnier, J., and B. Robson, 1989. Prediction of Protein Structure and the Principles of Protein Conformation. Springer US, Boston, MA. http://books.google.com/books?id=wdb_JfCDzZsC&pgis=1http://link.springer.com/10.1007/978-1-4613-1571-1.

48. Coluzza, I., and D. Frenkel, 2005. Virtual-move parallel tempering. *ChemPhysChem* 6:1779–1783. http://www3.interscience.wiley.com/journal/111081506/abstract?CRETRY=1&SRETRY=0http://www.ncbi.nlm.nih.gov/pubmed/16110517.

49. Derrida, B., 1981. PHENOMENOLOGICAL RENORMALIZATION OF THE SELF AVOIDING WALK IN 2 DIMENSIONS. *Journal Of Physics A-Mathematical And General* 14:L5–L9. papers2://publication/uuid/195D7FED-13BC-4E44-AF7E-7C5C7284E70E.

50. Pande, V. S., A. Y. Grosberg, and T. Tanaka, 2000. Heteropolymer freezing and design: Towards physical models of protein folding. *Reviews of Modern Physics* 72:259–314. papers2://publication/uuid/C216088C-3A62-4BB4-8DA3-A6F70F0A9759%5Cnhttp://link.aps.org/doi/10.1103/RevModPhys.72.259.

51. Pande, V. S., A. Y. Grosberg, and T. Tanaka, 1997. Statistical mechanics of simple models of protein folding and design. *Biophysical journal* 73:3192–3210.

52. Cardelli, C., V. Bianco, L. Rovigatti, F. Nerattini, L. Tubiana, C. Dellago, and I. Coluzza, 2017. The role of directional interactions in the designability of generalized heteropolymers. *Scientific Reports* 7:4986. http://dx.doi.org/10.1038/s41598-017-04720-7http://www.nature.com/articles/s41598-017-04720-7.

53. Lim, C. W., and T. W. Kim, 2012. Dynamic [2]Catenation of Pd(II) Self-assembled Macrocycles in Water. *Chemistry Letters* 41:70–72. http://www.journal.csj.jp/doi/10.1246/cl.2012.70.

54. Hino, S., T. Ichikawa, and Y. Kojima, 2010. Thermodynamic properties of metal amides determined by ammonia pressure-composition isotherms. *Journal of Chemical Thermodynamics* 42:140–143. http://dx.doi.org/10.1016/j.jct.2009.07.024.

SUPPLEMENTARY INFORMATIONS

Binding site modelling

The artificial binding site is a mould obtained by pushing the folded protein conformation on a planar mesh of self avoiding beads (self avoiding radius = 2 Å).

To obtain the mould, we initially generate a dense mesh in the $z = 0$ plane by placing the beads on a 2D square lattice with step 0.5 Å.

We then identify the centre of mass of the protein and use it to measure the height of the protein with respect to the plane, which is one of our main control parameters. Specifically, we consider the CM height z_{CM} normalised over the maximum distance,r_{MAX}, of a C_α from the CM of the protein, so that $\zeta = \frac{z_{CM}}{r_{MAX}}$ ranges between 0 (when the CM lies on the $z = 0$ plane) and 1.

We consider several values of ζ and for each of them we identify a protein orientation which maximise the contact surface with the binding site according to the following procedure. We begin by aligning the minor inertia axis of the protein with the z axis and then perform a discrete set of rotations around the x and y axes. For each of these rotations, we recompute the mould, map it to a triangular mesh, and calculate the surface of the latter. The area of the surface is obtained by summing the area of each triangle formed by all the triplets of the mesh points at $z \neq 0$.

The binding site is obtained by setting a minimum protein-mesh bead distance $\mu = 13$ Å and using it to inflate the radius of the protein C_αs, thus generating a mould by pushing downwards all mesh beads within the inflated radius. To avoid large gaps

within the binding site, we perform an iterative smoothing procedure. For each bead of the pocket we compute its distance to all its neighbouring beads on the square lattice, and if this is larger than the self-avoiding radius, we shift the vertical position of the bead so to fill the gap.

In order to confer a protein nature to the artificial binding site, a homogeneously distributed set of mesh points are "activated", i.e. switched from pure self avoiding beads to $C\alpha$ atoms. In order to do so, we firstly switch all the beads of the mesh to C_α, then we set a minimum binding site $C_\alpha - C_\alpha$ distance $\delta = 5$ Å and loop over all bead points starting from a corner of the mesh. For each activated atom in the loop we deactivate all points within a sphere of radius δ. Clearly, after the first atom the loop will incur both in activated and de-activated beads. The latter are simply skipped by the procedure. Finally, we deactivate all the beads at $z = 0$, so that only the pocket contains C_α beads.

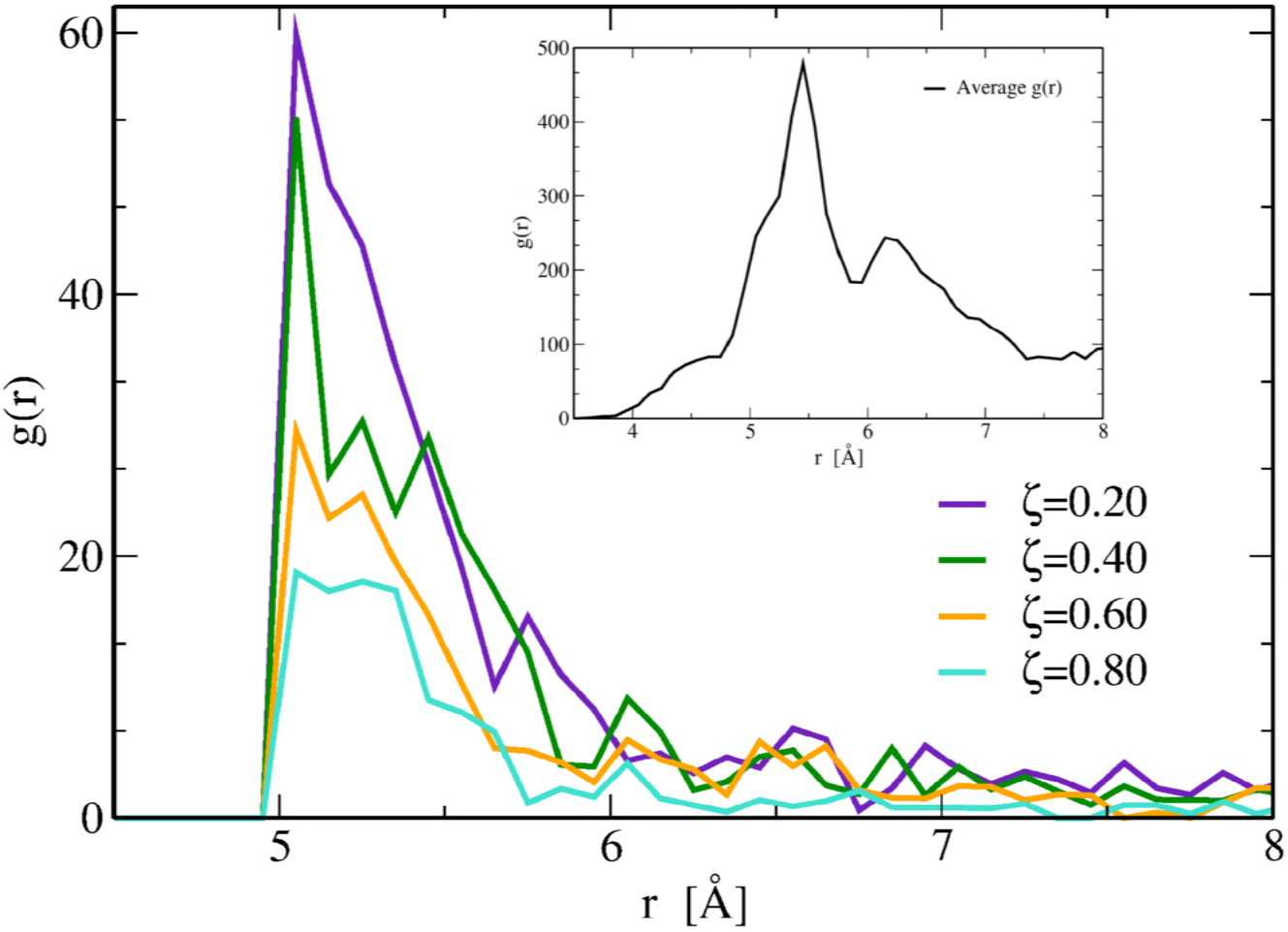


Figure S8: Radial distribution function $g(r)$ calculated over the active residues $C_{\mathtt{Surf}}$ of the binding site placed at the minimum distance $\delta = 5$ Å from each other. The latter was determined by identifying the value of δ that yields the mean nearest distance $\int_0^{7.5} \rho 4\pi r^2 g(r) r \, dr$ closest to 5.7 Å obtained by averaging over 145 protein structures taken from the PDB (black line in the inset). The integral interval was chosen visually taking a region large enough to fully contain the first peak in the $g(r)$. It is important to stress that in the protein $g(r)$ the selected distribution includes the contribution from all secondary structure elements (52).

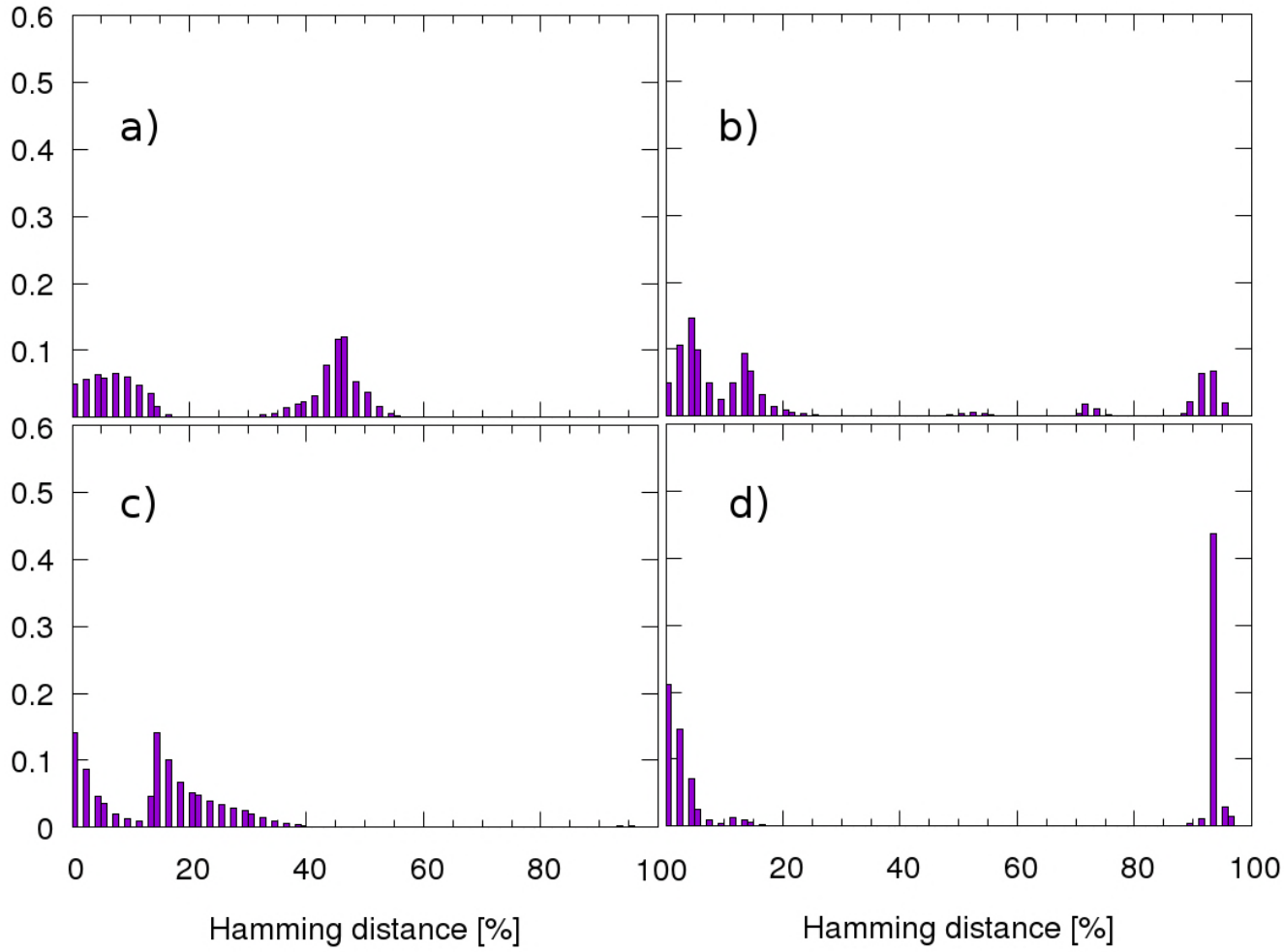


Figure S9: Histograms obtained by evaluating the Hamming distance (% relative to the chain length) between all possible pairs of sequences chosen selecting 200000 solutions around the design free energy minimum of systems A and B, corresponding to the ζ values specified in the following. All panels refer to self comparison of sequences belonging to the same basin. a) A: $\zeta = 0.20$; B: $\zeta = 0.20$ b) A: $\zeta = 0.40$; B: $\zeta = 0.40$ c) A: $\zeta = 0.60$; B: $\zeta = 0.60$ d) A: $\zeta = 0.80$; B: $\zeta = 0.80$

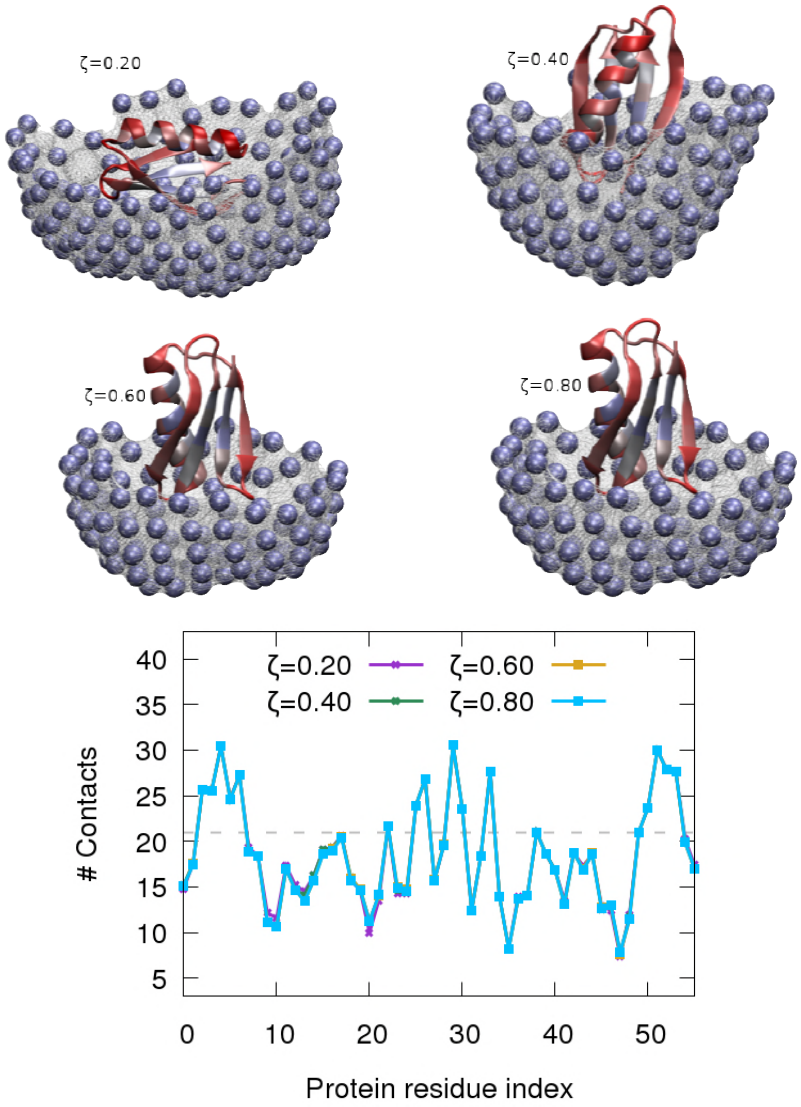


Figure S10: Water exposure profiles of protein G in presence of binding sites constructed with different ζ values. The colour scheme of the protein is related to the number of total contacts: solvent exposed regions in red, buried in blue. The plot shows the number of contacts for each protein residue. The grey dashed line defines the threshold $\Omega = 21$ on the number of neighbours above which an amino acid is considered buried in the protein core. We used the information relative to the exposure of protein G to set the radius of the probe sphere used in the evaluation of the number of contacts for each active C_α of the binding site (the blue spheres in the figure). We counted the offset for the number of neighbours for each of the binding site C_α by multiplying the fraction of the volume of a sphere of radius 6 Å that lies under the mesh surface by the average amino acid density in globular proteins ($\rho = 0.011687\ aa/\text{Å}^3$, evaluated over a set of 145 globular proteins). The radius of the sphere was optimized to reach an average exposure of each residue similar to the one of the most exposed residues of protein G (corresponding to the minima in the plot).

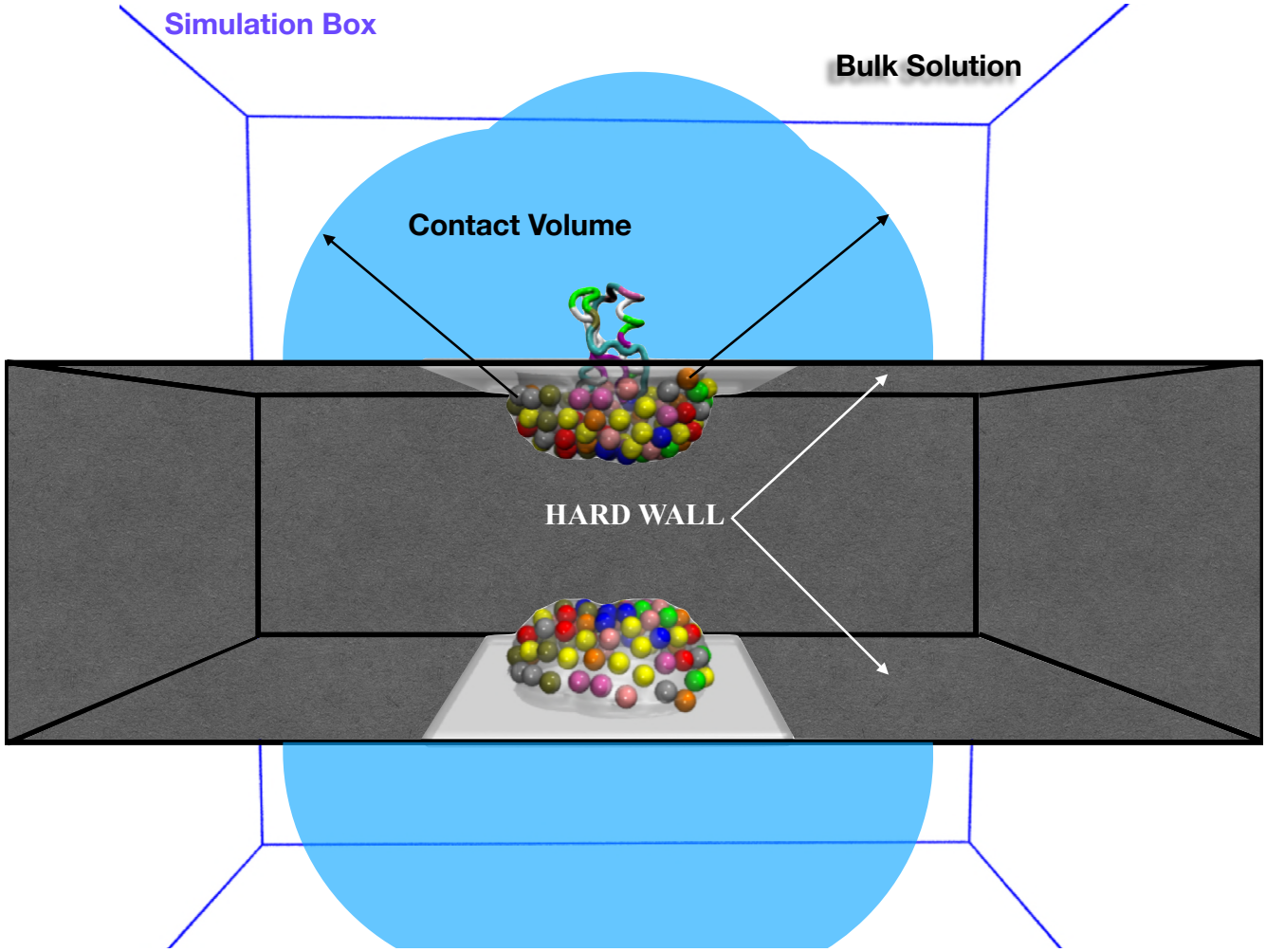


Figure S11: Simulation box for protein folding in presence of two copies of a potential partner. The partner is depicted as a binding site surface decorated with $C\alpha$ atoms, i.e. the coloured spheres in the image. Each colour, both on the protein backbone and on the $C\alpha$ of the binding sites, represents a different amino acid type. The protein sequence and the identity of the binding site amino acids are designed simultaneously with the procedure described in the *Design* sub-section. The protein is a flexible polymer diffusing in the box under Periodic Boundary Conditions. A mirroring move provides a swap between opposite chiralities, since the Caterpillar model does not take into account the amino acid chirality. The cubic box contains two binding sites, one the mirror image of the other, the position of which is fixed during the simulation. The binding sites are far apart enough to prevent the protein to interact with both at the same time. A hard wall between the binding sites prevents the protein to approach the surfaces from the convex side. For the evaluation of the association constant $K_a = \frac{Q_b}{Q_f}\frac{V_{box}}{n}$, we need to assess the number of binding sites in the box $n = 2$, define the accessible simulation volume V_{box} and discern bulk solution configurations (contributing to Q_f, the partition function of the free protein) from the ones where the protein is directly in contact with the binding site (contributing to Q_b, the partition function of the bound configurations). The accessible volume is easily obtained by subtracting the hard wall contribution to the box volume. The bound configurations are the ones characterised by a protein-binding site interaction energy $E_{inter} \neq 0$. The configurations contributing to Q_f, instead, are the ones for which the central amino acid of the protein is at a distance $r > H_{lenght} + R_{int}$ with respect to all the amino acids of the surface (being H_{lenght} half of the stretched protein length; R_{int} the interaction radius) and, therefore, where the interaction with the binding site is not possible. We refer to the latter region as the bulk solution.

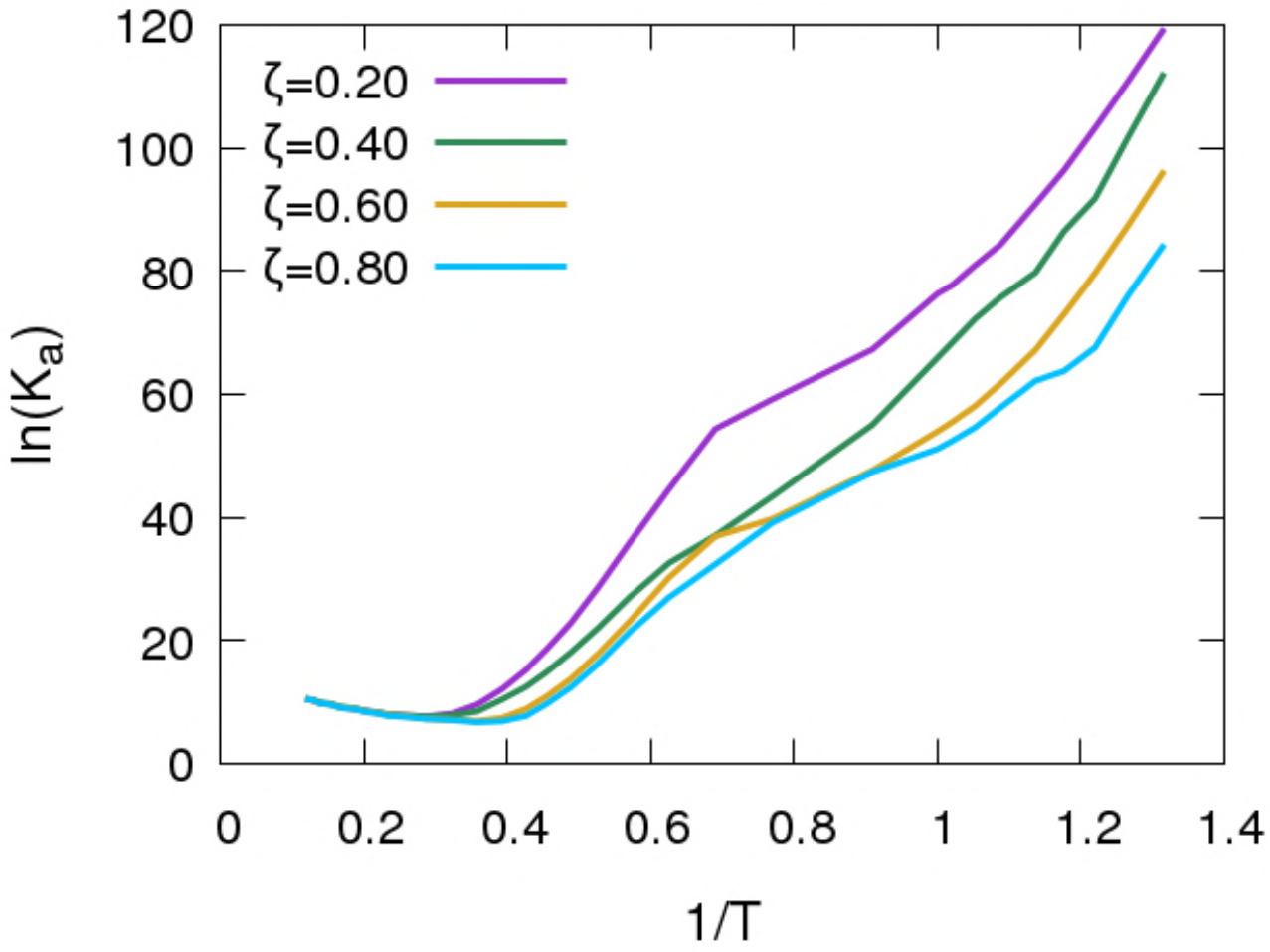


Figure S12: Van't Hoff plot of $\ln K_a$ as a function of the inverse reduced temperature $1/T$ for the investigated systems. The association constant $[l/mol]$ is computed as $K_a = \exp(-\Delta F/k_B T)V_{box}/n$, where V_{box} is the accessible volume of our simulation box, $n = 2$ is the number of binding sites and $\Delta F = -k_B T \ln(Q_b/Q_f)$ is the binding free energy (29). The partition functions Q_b and Q_f refer to all protein conformations bound to the binding site and free in the bulk solution, respectively.