SUPPLEMENTARY EXPERIMENTAL PROCEDURES

Determination of the reduced Structural Alphabet rSA

The 25 letters of the M32K25 SA were clustered using the ‘gromos’ clustering method implemented in Gromacs. The cutoff value was progressively increased from 0.35 to 1.1 Å by 0.05-Å steps and the resulting clustering quality was assessed by calculating the cluster silhouette widths. These indices are comprised between -1 and 1 and they measure the fitness of each object to its cluster. An optimal cutoff value of 0.65 Å was then selected, corresponding to the maximum average silhouette width value (Table S9). An identical cutoff has been used in the past to derive a completely independent SA, suggesting that this value is not specific to the M32K25 SA but it provides a measure of the intrinsic variability within groups of similar local states in experimental protein structures. The resulting 6 clusters, named after the letter of their representative (middle structures of the cluster), describe groups of fragment states (macrostates) that differ mainly for the pseudo-torsion \( \theta \) around the two central C\(^\alpha\) atoms (Table S1). Cluster A contains fragments in an extended conformation (\( \theta \sim 180^\circ \)), while fragments in clusters K (\( \theta \sim -110^\circ \)) and R (\( \theta \sim 120^\circ \)) are most frequently found in loops. Cluster U represents right-handed (\( \theta \sim 50^\circ \)) pseudo-helical conformations, while the two remaining singleton clusters contain the turn-shaped fragments N (\( \theta \sim 20^\circ \)) and Y (\( \theta \sim -50^\circ \)). The representatives of the six clusters define a reduced version of the M32K25 SA (rSA).

Recurring rSA motifs in \( \beta \)-hairpins

Structures of \( \beta \)-hairpins were extracted from the ArchDB database (Apr/2013 release), where loops are grouped into classes according to the type of secondary structures flanking the loop, the length of the loop and its conformation as described by backbone \( \varphi \) and \( \psi \) angles. Classes are further
decomposed into subclasses by clustering the structures of the class members. The database has been recently revised adding to the original Density Search (DS) classification a new clustering method based on the Markov Clustering (MCL) algorithm. Both DS and MCL subclasses were included in the present work. The database includes 8154 β-hairpins with loops of length 4 (BN.4) extracted from a non-redundant subset of the PDB and classified in 517 subclasses. Among these, 21 have more than 100 members and were considered for further analysis (ArchDB.BN4.100). The structures were downloaded from the PDB and encoded with the M32K25 SA and its reduced version rSA. Only the loops with all the Cα atoms resolved were considered, for a total of 3314 structures.

The loops in the 5 most populated rSA motifs shown in Figure 5 were clustered with the gromos1 method and a cutoff of 0.6 Å on Cα atoms.

The ArchDB.BN4.100 dataset was further filtered for β-hairpin structures from immunoglobulin-like (Ig-like) domains (ArchDB.BN4.100.Ig) using Pfam5 annotation. The β-hairpins from the Ig domain clan in Pfam (CL0011) were selected, for a total of 30 structures and 5 ArchDB subclasses (Table S3).

Selection of KUN loops

We identified loops with similar shape from structurally related proteins by filtering the ArchDB.BN4.100 database for β-hairpin loops in Immunoglobulin-like domains (ArchDB.BN4.100.Ig in Table S3), which are ubiquitous and have a very well conserved fold6. The loops with one of the most populated rSA encodings (‘KUN’, Table S3) were then selected. A structural superimposition showed that KUN structures are clustered in two groups (KUNg1 and KUNg2), with KUNg2 structures featuring a more bent conformation around the residue in position p4 compared to KUNg1 and significantly different Ramachandran plots at the p4 and p5 position (Figure 6). Inspection of the sequences (Figure S5A) shows that a charged or aromatic residue in position p5 in KUNg1 loops is replaced by a Gly residue in the KUNg2 group, which allows for the larger bending of the loop in KUNg2. Two sequences representative of the two groups were
selected for the simulations (bold in Figure S5A), namely the D-E loop from the Novel immune-
type receptor 10 (KUNg1) and the C-D loop from the B and T lymphocyte attenuator (KUNg2).

System setup – folding simulations

The GROMACS 4.5.5 program was used to prepare the initial system coordinates. The Amber99SB*-ildn force field was used for all the simulations. The experimental structures of the fast folders GB1 β-hairpin (number of residues \(N = 16\)) and Trp-Cage (\(N = 20\)) and of the immunoglobulin β-hairpins KUNg1 (\(N = 14\)) and KUNg2 (\(N = 10\)) were extracted from the PDB (Table S10). The protein termini were capped with ACE (N-term) and NME (C-term) groups. The charge of the ionisable residues was set to that of their standard protonation state at pH 7. To generate starting unfolded conformations, preliminary simulations were run where the number of \(C^\alpha\) contacts was steered towards low values. Unfolding simulations were stopped when extended conformations with no significant secondary structure content were found. The unfolded structures were then solvated with an octahedral box of TIP3P water molecules, with an initial minimal distance between the protein and the box boundaries of 8 Å. The systems were then neutralized by adding the appropriate number of counterions, resulting in systems with a total number of atoms ranging from ~10,000 to ~30,000 (Table S10).

System setup – CASP8 target for model refinement

The protein used for model refinement (XF2673 from Xylella fastidiosa, UniProt ID Q87A02) is the TR464 target of the model refinement category in the 8th community wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP8). Simulations were performed on the initial model provided to the CASP8 participants by the organisers. This was the best model among those submitted in the normal structure prediction exercise. The target description provided to the participants indicated the regions around residues 24 and 42 (41 and 59 in the best model numbering) as problematic, i.e. as having a particularly large deviation from the experimental structure. A PSIPRED\textsuperscript{10-11} prediction of secondary structure (version 3.3, default parameters) indicated irregular or extended elements in the region around residue 24 (residues 19-
28) (Figure S6A) instead of the helical elements found in the model structure. Residue 42 is instead found in a β-hairpin loop (residues 40-43). Comparison with the experimental structure shows that in the best model this loop is in a non-native conformation that does not allow the formation of a E5-R42 salt bridge (Figure 7A). These data were used to define the regions L1 (19-30) and L2 (39-44) included in the CVrSA used for the refinement. The starting TR464 model (residues 18-86, N = 69) was solvated with an octahedral box of TIP3P water molecules, with an initial minimal distance between the protein and the box boundaries of 10 Å. The system was then neutralized by adding 2 Na+ ions, resulting in systems with a total number of 13,922 atoms.

System equilibration

MD simulations were performed with GROMACS 4.5.57. Periodic boundary conditions were imposed. The equations of motion were integrated using the leap-frog method with a 2-fs time step. All the protein covalent bonds were frozen with the LINCS12 method, while SETTLE13 was used for water molecules. The electrostatic interactions were calculated with the particle mesh Ewald14 method, with a 9-Å cutoff for the direct space sums, a 1.2-Å FFT grid spacing, and a 4-order interpolation polynomial for the reciprocal space sums. A 9-Å cutoff was used for van der Waals interactions. Long-range corrections to the dispersion energy were included. The neighbour list for non-covalent interactions was updated every 5 steps.

The systems were first minimized with 2000 steps of steepest descent, followed by 2000 steps of conjugate gradient minimization. Harmonic positional restraints with a force constant of 4.8 kcal/mol/Å² were imposed onto the protein heavy atoms and gradually reduced to 1.2 kcal/mol/Å² in 160 ps, while the temperature was increased from 200 to 300 K at constant volume. The systems were then simulated at constant temperature (300 K) and pressure (1 bar) for 200 ps. After removal of harmonic restraints, 1 ns of equilibration was run in NPT conditions, followed by a 1-ns NVT simulation with the volume fixed to the NPT equilibrated value. The Berendsen algorithm15 was employed for temperature and pressure regulation during the equilibration runs, with coupling constants of 0.2 and 1 ps, respectively. In the last NVT equilibration, the thermostat was switched
to the v-rescale\textsuperscript{16} method with a 0.1-ps coupling constant, which was then used for the subsequent production runs. In the refinement of target TR464, after removal of heavy atom restraints, weak positional restraints were kept on secondary structure elements not included in regions L1 and L2 (C\textalpha{} atoms only), with a force constant of 0.1 kcal/mol/Å\textsuperscript{2}.

Enhanced sampling – Metadynamics

Enhanced sampling MD simulations were performed with GROMACS 4.5.5\textsuperscript{7} coupled with PLUMED-1.3\textsuperscript{17}. Metadynamics\textsuperscript{18-19} simulations were run starting from the equilibrated systems. Gaussians were deposited every 5 ps with height 0.48 kcal/mol. The CV\textsubscript{SA} parameters (eq. 1) were set to \( n = 8, m = 10 \) and \( \rho_0 = 0.6 \, \text{Å} \). The Gaussian width was set to 0.1 on the basis of preliminary unbiased simulations. In the folding simulations, a CV describing a generic (i.e. not containing information on the native state) global property was used in addition to the CV\textsubscript{SA}. Different global CVs were tested and the best performance was obtained when using a CV measuring the total number of C\textalpha{} contacts calculated considering all the possible pairs of residues. The CMAP CV implemented in PLUMED-1.3\textsuperscript{17} was used, with \( n = 8, m = 10 \), a contact threshold of 6.5 Å and a Gaussian width of 2. To reduce the sampling of completely extended conformations in the folding simulations, an upper limit was set on the radius of gyration \( R_g \), using a quartic function as restraining potential (UWALL option of PLUMED-1.3) with an energy constant of 9.6 kcal/mol/Å\textsuperscript{4}.

For the fast folders and the KUN hairpins, the \( R_g \) limit was estimated on the basis of the empirical formula approximating the experimental \( R_g \) of the unfolded state \( R_g^{\text{unf}} = R_g^0 \times N^v \), where \( R_g^0 = 1.927 \, \text{Å} \), \( v = 0.598 \) and \( N \) is the number of residues\textsuperscript{20}. This results in a rounded value of 10 Å for GB1 β-hairpin, 11 Å for Trp-Cage, 9 Å for KUNg1 and 8 Å for KUNg2. An upper limit on \( R_g \) was also set in the simulations for the refinement of TR464, where \( R_g \) was allowed to increase by at most 1 Å from the starting value of 10.5 Å. Each metadynamics simulation was run for a total of 100 ns. Coordinates were saved every 1 ps. In the analysis of the trajectories, high-CV\textsubscript{SA} ensembles were selected as composed of structures with CV\textsubscript{SA} \( \geq CV_{SA}^{\text{max}} - 2 \). Due to the reduced percentage of
high-CV\textsubscript{SA} structures in the Metadynamics simulation of Trp-Cage using the rSA encoding (0.03%), in this case the definition was extended to the structures with CV\textsubscript{SA} \textgreater= CV\textsubscript{SA}\textsuperscript{max} – 3 (2%).

Enhanced sampling – Steered Molecular Dynamics (SMD)

Steered MD simulations were started from the equilibrated systems. The CV\textsubscript{SA} was steered using a harmonic restraint with the reference value moving at constant velocity from the initial CV\textsubscript{SA} value to CV\textsubscript{SA}\textsuperscript{max}, corresponding to the number of fragments used for its definition. The CV\textsubscript{SA} parameters (eq. 1) were set to \( n = 8, m = 10 \) and \( \rho_0 = 0.6 \) Å. Multiple simulations were run starting from different atomic velocities (Table S11). For the GB1 and KUN \( \beta \)-hairpins and the TR464 refinement, 9-ns replicas were run with a CV\textsubscript{SA} steering rate of 0.00275 \( \text{ps}^{-1} \). For Trp-Cage, test simulations showed that lower steering rates were required to have the same proportion of final folded states, so that 15-ns replicas were run with a CV\textsubscript{SA} steering rate of 0.001375 \( \text{ps}^{-1} \). The force constant of the moving harmonic restraint was set to 7 kcal/mol/Å\textsuperscript{2} in all cases. The same upper limits on R\textsubscript{g} were set as in the Metadynamics simulations.

For each system, the distribution of the work values from all the replicas was analysed. As expected considering the high velocity used for the steering (~ 1.4-2.8 CV\textsubscript{SA} units per ns, Table S11), the transformations were associated with large work values, broadly distributed and with multiple peaks (Figure S8). In all the cases considered in this study, the trajectories associated with the first peak (low-work transformations) were found to have a higher success rate in producing a correctly folded structure, with an increase in productive trajectories compared to the whole set of SMD runs from 36 to 38% (\( \beta \)-hairpin, \( p^{\text{NatFil}W} \) in Table S2), from 33 to 44% (TrpCage, Table S2), from 36 to 39% (KUNg1, Table S5), from 72 to 82% (KUNg2, Table S5) and from 18 to 33% (TR464, Table S7). SMD runs were filtered so that only the trajectories with work values within the first peak (low-work runs) were retained. Additional simulations were run on the final structures of these trajectories to allow for side chain and solvent relaxation. The CV\textsubscript{SA} was restrained to its final CV\textsubscript{SA}\textsuperscript{max} value, but with the RMSD cutoff threshold \( \rho_0 \) (eq. 1) increased from 0.6 to 0.8 Å, thus allowing also for small adjustments of the backbone conformation. These relaxation simulations
were run for 5 (fast folders and KUN β-hairpins) and 10 (TR464) ns (Table S11). They were usually very effective in recovering key side chain interactions (Figure 7C and S5B), together with small improvements in the average RMSD$_{\text{nat}}$ (Table S2, S5 and S7).

**Clustering of trajectories**

Representative structures of high-CV$_{\text{SA}}$ ensembles from Metadynamics trajectories were extracted with the gromos clustering method$^1$ applied to C$^\alpha$ coordinates. The cutoff values were estimated for each trajectory on the basis of the distribution of pairwise RMSD values$^3$, resulting in cutoff values of 2.0 (KUNg1), 1.2 (KUNg2) and 1.7 (TR464) Å. The representative of the most populated clusters are reported in Figure 6B/C and 7B.

**Rescoring with Rosetta**

Structures from the blind loop refinement of the TR464 CASP target were independently rescored with Rosetta$^{21-23}$ 3.3. Conformations with high CV$_{\text{rSA}}$ values ($\geq CV_{\text{rSA}}\text{max} - 2$) were extracted from the trajectories. Structures were converted to PDB format and atom names were remapped to Rosetta definition using an in-house python script (available for download from https://afornililab.wordpress.com/software or http://people.brunel.ac.uk/~csstaap2/software.html).

Rescoring was performed with the Rosetta full-atom scoring function$^{22-23}$.

**REFERENCES**

1. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W.; Mark, A., Peptide folding: When simulation meets experiment. *Angew. Chem.-Int. Edit.* 1999, 38, 236-240.

2. Rousseeuw, P., Silhouettes - A graphical aid to the interpretation and validation of cluster-analysis. *J. Comput. Appl. Math.* 1987, 20, 53-65.

3. Micheletti, C.; Seno, F.; Maritan, A., Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* 2000, 40, 662-674.

4. Bonet, J.; Planas-Iglesias, J.; Garcia-Garcia, J.; Marin-Lopez, M. A.; Fernandez-Fuentes, N.; Oliva, B., ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res.* 2014, 42, D315-9.

5. Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.; Finn, R. D., The Pfam protein families database. *Nucleic Acids Res.* 2012, 40, D290-301.

6. Bork, P.; Holm, L.; Sander, C., The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* 1994, 242, 309-320.

7. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput
8. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950-1958.

9. Maccallum, J. L.; Hua, L.; Schnieders, M. J.; Pande, V. S.; Jacobson, M. P.; Dill, K. A., Assessment of the protein-structure refinement category in CASP8. *Proteins* **2009**, *77* Suppl 9, 66-80.

10. Buchan, D. W. A.; Minneci, F.; Nugent, T. C. O.; Bryson, K.; Jones, D. T., Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **2013**, *41*, W349-57.

11. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195-202.

12. Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J., LINCS: a linear constraint solver for molecular simulations. *J. Comp. Chem.* **1997**, *18*, 1463-1472.

13. Miyamoto, S.; Kollman, P., SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comp. Chem.* **1992**, *13*, 952-962.

14. Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L., A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577.

15. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684-3690.

16. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

17. Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A., PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961-1972.

18. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562-12566.

19. Laio, A.; Gervasio, F. L., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71*, 126601.

20. Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W., Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 12491-12496.

21. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209-225.

22. Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D., Protein structure prediction using Rosetta. *Methods Enzymol.* **2004**, *383*, 66-93.

23. Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Cnorn, J. E.; Kim, D. E.; Lyskov, S.; Berzondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P., ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **2011**, *487*, 545-574.
Figure S1. Time evolution of the percentage of native C\(^\beta\) contacts (pNat(C\(^\beta\))) of the GB1 \(\beta\)-hairpin (upper panels) and the Trp-Cage mini-protein (lower panels) during Metadynamics simulations. A threshold of 9 Å on C\(^\beta\) distances was used for the definition of the contacts. The CV\(_{SA}\) was defined using the SA (left panels) and rSA (right panels) encodings of the experimental structures. High-CV\(_{SA}\) points are coloured in blue. For each panel, a representative high-CV\(_{SA}\) structure with a non-native arrangement of the fragments is reported (blue cartoon), together with the corresponding pNat(C\(^\beta\)) and CV\(_{SA}\) values. It is to be noted that the latter are comparable to the CV\(_{SA}\) values of the native-like structures represented in Figure 2.
Figure S2. Time evolution of the CV<sub>SA</sub> components during a productive SMD simulation of the GB1 β-hairpin (rSA target encoding). The value of the switching function for each fragment is reported as vertical lines coloured from white (0) to grey (1). The RMSD of C<sup>α</sup> atoms from the experimental structure is reported in blue. Representative structures from the trajectory are also shown as cartoon.
**Figure S3.** Time evolution of the percentage of native $\beta$ contacts (A/B) and of the $\alpha$ RMSD from the experimental structure (C/D) of the GB1 $\beta$-hairpin (left) and the Trp-Cage mini-protein (right) averaged over productive (blue) and non-productive (orange) SMD simulations. A SMD run is defined productive if $\text{RMSD}_{\text{nat}} < 2$ Å at the end of the simulation.
**Figure S4.** Final structures from productive (blue) and non-productive (orange) SMD simulations of the GB1 β-hairpin. Residues Trp3 and Tyr5 (licorice) are correctly aligned on one side of the molecule in productive runs, while they get stuck between the two strands by non-native contacts in non-productive runs, hindering the formation of the β-sheet.
Figure S5. (A) Sequences of the KUNg1 and KUNg2 β-hairpins. (B) Superimposition of structures from productive SMD runs. Final structures are represented for the initial SMD runs (left) and the relaxation runs (right). Only the structures from low-work runs (see Supplementary Methods) were selected for relaxation. The His3 and Gly5 residues are represented as licorice. The hydrogen bond between the His3 imidazole group and the Gly5 NH group is represented with dots.
Figure S6. (A) PSIPRED prediction of secondary structures for the L1 and L2 regions (H=Helix, E=Strand, C=Coil). The confidence of the prediction (Conf) is also reported, together with the amino acidic sequence. (B) Upper panel: time evolution of the Cα RMSD from the native structure during the Metadynamics simulation (rSA Target encoding) for the L1 (light green) and L2 (dark green) regions. Lower panel: time evolution of the CVSA (blue) and of the E5-R42 minimum distance calculated over all pairs of non-hydrogen atoms (red) during the Metadynamics simulation (rSA Target encoding). The CVSA reaches its maximum value only at the end of the simulation, when both L1 and L2 are in their target conformation. The E5-R42 distance drops to values <= 3.5 Å in the same time interval, indicating that the formation of the salt bridge is correlated with the rearrangement of L2.
Figure S7. Comparison of productive and non-productive SMD runs for the TR464 target of CASP8. A SMD run is defined productive if the final structure has Cα RMSD from the experimental native structure ≤ 2.5 Å. (A and B) Time evolution of the CV_{SA} components during productive (left) and non-productive (right) SMD simulations of TR464 (rSA target encoding). The value of the switching function for each fragment is reported as vertical lines coloured from white (0) to grey (1). The percentage of native Cβ contacts (pNat(Cβ)) averaged over the SMD runs is also reported in red. (C and D) Superimposition of the final L1 and L2 structures from productive (blue) and non-productive (orange) SMD runs of TR464. The experimental structure is also reported as reference (white cartoon).
**Figure S8.** Distribution of work values for SMD simulations of the GB1 β-hairpin (A) and the Trp-Cage mini-protein (B).
Table S1. Clustering of the M32K25 SA into the reduced SA (rSA).

|     | $\phi_1$ (°) | $\phi_2$ (°) | $\theta$ (°) | cluster description |
|-----|--------------|--------------|--------------|---------------------|
| A   | 122.4        | 119.4        | 164.2        | β-strand            |
| B   | 129.8        | 135.6        | -176.6       |                     |
| C   | 117.1        | 111.0        | -142.2       |                     |
| D   | 118.4        | 126.9        | -146.1       |                     |
| E   | 116.7        | 138.6        | 168.7        |                     |
| G   | 135.3        | 118.6        | -148.5       |                     |
| I   | 133.6        | 117.1        | -120.8       |                     |
| L   | 110.0        | 90.8         | -158.8       |                     |
| M   | 110.0        | 100.8        | 177.0        |                     |
| F   | 115.6        | 112.9        | -117.9       | loop/β-strand       |
| H   | 120.1        | 114.3        | -90.7        |                     |
| J   | 115.9        | 91.4         | -134.6       |                     |
| K   | 119.7        | 90.4         | -105.9       |                     |
| O   | 92.4         | 91.2         | -127.4       |                     |
| P   | 91.8         | 96.7         | -104.8       |                     |
| T   | 93.0         | 92.8         | 83.1         | helix               |
| U   | 91.4         | 90.7         | 49.8         |                     |
| V   | 93.3         | 89.1         | 68.3         |                     |
| W   | 93.8         | 105.2        | 32.3         |                     |
| X   | 111.4        | 94.6         | 21.8         |                     |
| Q   | 95.9         | 117.7        | 136.0        | loop/helix          |
| R   | 94.5         | 112.6        | 115.0        |                     |
| S   | 96.3         | 94.7         | 112.0        |                     |
| N   | 90.1         | 138.2        | 19.6         | turn                |
| Y   | 89.0         | 95.1         | -54.4        | turn                |

- The representative letter of the cluster is highlighted in bold.
- $\phi_1$ Pseudo-angle defined by the fragment C$^\alpha$ atoms 1, 2 and 3.
- $\phi_2$ Pseudo-angle defined by fragment C$^\alpha$ atoms 2, 3 and 4.
- $\theta$ Pseudo-torsion around fragment C$^\alpha$ atoms 2 and 3.
Table S2. Fraction of native-like final structures in SMD simulations of the GB1 β-hairpin and the Trp-Cage mini-protein.

|         | \( p^{\text{Nat}} \) (%) | \( \text{RMSD}_{\text{nat}} \) (Å) | \( p^{\text{Nat}}_{\text{FiltW}} \) (%) | \( \text{RMSD}_{\text{nat}} \) (Å) | \( p^{\text{Nat}}_{\text{RelaxW}} \) (%) | \( \text{RMSD}_{\text{nat}} \) (Å) |
|---------|---------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| GB1 β-Hairpin (rSA) | 36.0 (9/25) | 1.00 (0.75) | 37.5 (9/24) | 1.00 (0.75) | 33.5 (8/24) | 0.80 (0.70) |
| Trp-Cage (rSA) | 33.3 (4/12) | 1.79 (1.60) | 44.4 (4/9) | 1.79 (1.60) | 77.8 (7/9) | 1.53 (0.82) |

\( a \) Percentage of productive SMD trajectories (final structure with \( \text{RMSD}_{\text{nat}} \leq 2 \) Å). The number of productive over total SMD runs is reported in parentheses.

\( b \) \( C^\alpha \) RMSD from the native structure. Average values are reported, calculated over the final structures of productive SMD trajectories. The minimum final \( \text{RMSD}_{\text{nat}} \) value is reported in parentheses.

\( c \) Percentage of productive SMD trajectories after filtering for low-work trajectories.

\( d \) \( \text{RMSD}_{\text{nat}} \) values calculated as in \( b \) after filtering for low-work trajectories.

\( e \) Percentage of productive SMD trajectories after filtering for low-work trajectories and after relaxation.

\( f \) \( \text{RMSD}_{\text{nat}} \) values calculated as in \( b \) after filtering for low-work trajectories and after relaxation.
Table S3. rSA motifs in the ArchDB.BN4.100 and ArchDB.BN4.100.Ig datasets

|        | ArchDB.BN4.100 |        | ArchDB.BN4.100.Ig |
|--------|----------------|--------|-------------------|
|        | npDB<sup>a</sup> | nSAenc<sup>b</sup> | nDSclass<sup>c</sup> | nMCLclass<sup>d</sup> | npDB<sup>a</sup> | nSAenc<sup>b</sup> | nDSclass<sup>c</sup> | nMCLclass<sup>d</sup> |
| UUK    | 1638            | 37     | 3                 | 6                   | 1138             | 6                 | 3                 | 6                   |
| UKR    | 609             | 10     | 1                 | 2                   | 509              | 10                | 1                 | 2                   |
| KUN    | 577             | 16     | 3                 | 1                   | 517              | 16                | 3                 | 1                   |
| KUK    | 264             | 18     | 3                 | 4                   | 234              | 18                | 3                 | 4                   |
| NKU    | 133             | 4      | 1                 | 0                   | 113             | 4                 | 1                 | 0                   |
| KUR    | 43              | 5      | 2                 | 1                   | 43              | 5                 | 2                 | 1                   |
| UYR    | 16              | 1      | 2                 | 2                   | 16              | 1                 | 2                 | 2                   |
| KYR    | 5               | 2      | 1                 | 1                   | 5               | 2                 | 1                 | 1                   |
| KNN    | 3               | 2      | 2                 | 0                   | 3               | 2                 | 2                 | 0                   |
| KKU    | 3               | 1      | 1                 | 0                   | 3               | 1                 | 1                 | 0                   |
| UKU    | 3               | 2      | 1                 | 0                   | 3               | 2                 | 1                 | 0                   |
| UKA    | 2               | 1      | 1                 | 2                   | 2               | 1                 | 1                 | 2                   |
| UUA    | 2               | 2      | 1                 | 2                   | 2               | 2                 | 1                 | 2                   |
| YUR    | 2               | 1      | 1                 | 0                   | 2               | 1                 | 1                 | 0                   |
| KUU    | 2               | 2      | 2                 | 1                   | 2               | 2                 | 2                 | 1                   |
| YYR    | 2               | 1      | 2                 | 0                   | 2               | 1                 | 2                 | 0                   |
| RYA    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |
| AUK    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |
| NYN    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |
| RYK    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |
| NAU    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |
| KYN    | 1               | 1      | 1                 | 1                   | 1               | 1                 | 1                 | 1                   |
| UYN    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |
| KRK    | 1               | 1      | 1                 | 2                   | 1               | 1                 | 1                 | 2                   |
| URK    | 1               | 1      | 1                 | 2                   | 1               | 1                 | 1                 | 2                   |
| YUN    | 1               | 1      | 1                 | 0                   | 1               | 1                 | 1                 | 0                   |

<sup>a</sup> Number of PDB entries in the dataset for each rSA motif.
<sup>b</sup> Number of different SA motifs in the dataset mapping into each rSA motif.
<sup>c</sup> Number of different ArchDB DS classes in the dataset for each rSA motif.
<sup>d</sup> Number of different ArchDB MCL classes in the dataset for each rSA motif.
Table S4. Pairwise Cα RMSD of the top 5 rSA motifs in the ArchDB.BN4.100 dataset.

|     | UUK | UKR | KUN | KUK | NKU |
|-----|-----|-----|-----|-----|-----|
| UUK | 0.53<sup>a</sup> | -   | -   | -   | -   |
| UKR | 2.17<sup>b</sup> | 0.60| -   | -   | -   |
| KUN | 1.86 | 2.21| 0.78| -   | -   |
| KUK | 1.80 | 2.95| 1.43| 1.03| -   |
| NKU | 2.70 | 2.96| 2.15| 1.65| 0.46|

<sup>a</sup> Diagonal elements: average pairwise Cα RMSD between β-hairpin structures with the same rSA motifs.

<sup>b</sup> Off-diagonal elements: Cα RMSD between the representatives of the most populated cluster of each rSA motif.
Table S5. Fraction of native-like structures in Metadynamics and SMD simulations of the KUN β-hairpins.

| Metadyn |  | SMD |  |  |  |  |  |
|---------|---|-----|---|---|---|---|---|
| p<sub>Nat</sub> (%)<sup>a</sup> | p<sub>Nat</sub> (%)<sup>b</sup> | RMSD<sub>Nat</sub> (Å)<sup>d</sup> | p<sub>FiltCVSA</sub> (%)<sup>c</sup> | RMSD<sub>Nat</sub>(Å)<sup>f</sup> | p<sub>FiltW</sub> (%)<sup>e</sup> | RMSD<sub>Nat</sub>(Å)<sup>g</sup> | RMSD<sub>Nat</sub>(Å)<sup>h</sup> |
| KUNg1   | 4.5 | 26.9 | 36.0 | 1.25 (1.00) | 38.9 | 1.20 (1.00) | 38.9 | 1.31 (0.80) |
|         | (9/25) |     | (7/18) |       | (18/22) |       | (18/22) |       |
| KUNg2   | 9.7 | 35.5 | 72.0 | 1.34 (1.09) | 81.8 | 1.34 (1.09) | 86.4 | 1.28 (0.97) |
|         | (18/25) |     |       |       |       |       |       |       |

<sup>a</sup>Percentage of structures in the whole trajectory with RMSD<sub>Nat</sub> <= 2 Å.
<sup>b</sup>Percentage of structures in the high-CV<sub>SA</sub> ensemble (CV<sub>SA</sub> >= CV<sub>SA</sub> max – 2) with RMSD<sub>Nat</sub> <= 2 Å.
<sup>c</sup>Percentage of productive SMD trajectories (final structure with RMSD<sub>Nat</sub> <= 2 Å). The number of productive over total SMD runs is reported in parentheses.
<sup>d</sup> Cα RMSD from the native structure. Average values are reported, calculated over the final structures of productive SMD trajectories. The minimum final RMSD<sub>Nat</sub> value is reported in parentheses.
<sup>e</sup>Percentage of productive SMD trajectories after filtering for low-work trajectories. The number of productive over low-work SMD runs is reported in parentheses.
<sup>f</sup>RMSD<sub>Nat</sub> values calculated as in d after filtering for low-work trajectories.
<sup>g</sup>Percentage of productive SMD trajectories after filtering for low-work trajectories and relaxation. The number of productive over low-work SMD runs is reported in parentheses.
<sup>h</sup>RMSD<sub>Nat</sub> values calculated as in d after filtering for low-work trajectories and after relaxation.
Table S6. Matrix of Cα RMSD values (Å) between simulated and experimental structures of KUN β-hairpins.

|             | KUNg1-Xray | KUNg2-Xray |
|-------------|------------|------------|
| KUNg1-Metadyn\(^a\) | 1.15       | 1.40       |
| KUNg2-Metadyn\(^a\) | 1.40       | 0.67       |
| KUNg1-SMD\(^b\)    | 1.04       | 1.89       |
| KUNg2-SMD\(^b\)    | 1.20       | 0.94       |

\(^a\) Representative of the high-CV\(_{SA}\) Metadynamics trajectory extracted with a cluster analysis.

\(^b\) Average value calculated over the final structures of productive SMD runs.
Table S7. Fraction of refined final structures in SMD simulations of the CASP8 Target TR464.

| SMD (Target) | $p_{Nat}$ (%) | $p_{Nat}(L1)$ (%) | $p_{Nat}(L2)$ (%) | RMSD$_{nat}$ (Å) | $p_{Nat\ Filt\ W}$ (%) | $p_{Nat\ Filt\ W}(L1)$ (%) | $p_{Nat\ Filt\ W}(L2)$ (%) | RMSD$_{nat}$ (Å) | $p_{Nat\ Filt\ WR}$ (%) | $p_{Nat\ Filt\ WR}(L1)$ (%) | $p_{Nat\ Filt\ WR}(L2)$ (%) | RMSD$_{nat}$ (Å) |
|-------------|----------------|-------------------|-------------------|------------------|------------------------|-------------------------------|------------------------|------------------|------------------------|------------------------|------------------------|------------------|
|             | 18.0 (9/50)    | 42.0              | 40.0              | 2.27             | 33.3                   | 42.9                          | 71.4                   | 2.24             | 33.3                   | 71.4                   | 85.7                   | 2.09             |

* SMD simulations run using the Target rSA string (Table 1).

b Percentage of productive SMD trajectories (final structure with RMSD$_{nat} <= 2.5$ Å). The number of productive over total SMD runs is reported in parenthesis.

c Percentage of SMD trajectories with RMSD$_{nat}(L1/L2) <= 2.5$ Å. Values in parentheses are calculated using local fit RMSD$_{nat}(L1/L2)$ values and a threshold of 1 Å.

d $C_{\alpha}$ RMSD from the native structure. Average values are reported, calculated over the final structures of productive SMD trajectories. The minimum final RMSD$_{nat}$ value is reported in parenthesis.

e Percentage of productive SMD trajectories after filtering for low-work trajectories. The number of productive over low-work SMD runs is reported in parentheses.

f Percentage of SMD trajectories calculated as in c after filtering for low-work trajectories.

g RMSD$_{nat}$ calculated as in d after filtering for low-work trajectories.

h Percentage of productive SMD trajectories after filtering for low-work trajectories and after relaxation. The number of productive over low-work SMD runs is reported in parenthesis.

i Percentage of SMD trajectories calculated as in c after filtering for low-work trajectories and after relaxation.

j RMSD$_{nat}$ calculated as in d after filtering for low-work trajectories and after relaxation.
Table S8. Top 20 structures from Metadynamics simulations of TR464 (BlindL2 encoding) ranked using the Rosetta scoring function\(^a\).  

| Rosetta score\(^b\) | CV\(_{SA}\) | RMSD\(_{nat}\) (Å)\(^c\) | RMSD\(_{nat}\) (L2-LF) (Å)\(^d\) | RMSD\(_{nat}\) (L2-GF) (Å)\(^e\) | Label    |
|---------------------|-----------|----------------|----------------|----------------|---------|
| -12.23              | 12.0      | 2.67           | 0.66           | 3.01           | BlindL2a|
| -11.83              | 10.6      | 2.33           | 1.88           | 3.81           | BlindL2a|
| -11.79              | 10.9      | 2.44           | 0.92           | 2.66           | BlindL2a|
| -11.23              | 10.1      | 2.22           | 1.35           | 2.61           | BlindL2a|
| -10.63              | 10.2      | 2.70           | 1.12           | 3.31           | BlindL2a|
| -10.10              | 11.5      | 2.53           | 0.72           | 2.64           | BlindL2a|
| -10.00              | 10.2      | 2.13           | 2.37           | 3.45           | BlindL2a|
| -9.90               | 11.1      | 2.79           | 1.61           | 5.02           | BlindL2c|
| -9.87               | 10.6      | 2.61           | 1.19           | 2.86           | BlindL2a|
| -9.67               | 11.2      | 2.46           | 0.96           | 2.86           | BlindL2a|
| -8.49               | 10.0      | 2.49           | 1.52           | 3.54           | BlindL2a|
| -8.22               | 10.2      | 2.33           | 1.64           | 3.27           | BlindL2c|
| -8.15               | 12.0      | 2.61           | 0.67           | 2.67           | BlindL2a|
| -8.01               | 11.1      | 3.22           | 2.20           | 3.77           | BlindL2b|
| -7.73               | 10.1      | 2.29           | 1.35           | 2.51           | BlindL2a|
| -7.49               | 11.0      | 2.84           | 2.21           | 3.19           | BlindL2b|
| -7.48               | 10.8      | 2.50           | 0.91           | 2.55           | BlindL2a|
| -7.42               | 10.8      | 2.62           | 1.01           | 3.23           | BlindL2a|
| -7.19               | 10.5      | 2.85           | 2.29           | 4.27           | BlindL2b|
| -6.72               | 11.4      | 3.05           | 2.27           | 3.09           | BlindL2b|

\(^a\) Structures from the BlindL2a simulation are highlighted in grey. RMSD\(_{nat}\)(L2-GF) and RMSD\(_{nat}\)(L2-LF) values that are at least 1.0 Å lower than the starting values (4.35 Å for GF and 2.02 Å for LF) are reported in bold.

\(^b\) Rosetta Energy Units.

\(^c\) \(\alpha\) RMSD from the native structure.

\(^d\) \(\alpha\) RMSD from the native structure for the L2 region calculated in local fit mode.

\(^e\) \(\alpha\) RMSD from the native structure for the L2 region calculated in global fit mode.
Table S9. Clustering of the M32K25 SA.

| cutoff (Å) \(^a\) | nclusters \(^b\) | <silhouette> \(^c\) |
|-------------------|-----------------|-----------------|
| 0.35              | 17              | 0.10            |
| 0.40              | 13              | 0.13            |
| 0.45              | 11              | 0.14            |
| 0.50              | 12              | 0.00            |
| 0.55              | 9               | 0.15            |
| 0.60              | 8               | 0.19            |
| **0.65**          | **6**           | **0.35**        |
| 0.70              | 7               | 0.20            |
| 0.75              | 6               | 0.24            |
| 0.80              | 6               | 0.11            |
| 0.85              | 6               | 0.08            |
| 0.90              | 6               | 0.04            |
| 0.95              | 5               | 0.15            |
| 1.00              | 4               | 0.23            |
| 1.05              | 4               | 0.23            |
| 1.10              | 4               | 0.23            |

\(^a\) Cutoff value used in the cluster analysis.
\(^b\) Number of clusters for each cutoff value.
\(^c\) Average silhouette width calculated over all clusters for each cutoff value.
| Molecule          | PDB ID | chain | residues<sup>a</sup> | n<sub>wat</sub><sup>b</sup> | n<sub>ions</sub><sup>c</sup> | n<sub>atoms</sub><sup>d</sup> |
|-------------------|--------|-------|----------------------|----------------|----------------|----------------|
| GB1 β-hairpin     | 1PGB   | A     | 41-56                | 5805           | 3 (Na<sup>+</sup>) | 17674          |
| Trp-Cage          | 1L2Y   | A     | 1-20                 | 8275           | 1 (Cl<sup>-</sup>)  | 25139          |
| KUNg1 β-hairpin   | 2QHL   | C     | 67-80                | 9741           | 1 (Na<sup>+</sup>) | 29463          |
| KUNg2 β-hairpin   | 1XAU   | A     | 49-58                | 3350           | 1 (Cl<sup>-</sup>)  | 10221          |

<sup>a</sup> Residue indices (PDB numbering).
<sup>b</sup> Total number of water molecules.
<sup>c</sup> Total number of counterions (the ion type is indicated in parentheses).
<sup>d</sup> Total number of atoms in the system including the solvent.
Table S11. Number and length of SMD runs for each simulated system.

|       | number | length      | rate$^a$ |
|-------|--------|-------------|----------|
| B-hairpin | 25      | 9+5 ns     | 0.00275  |
| Trp-Cage  | 12      | 15+5 ns    | 0.001375 |
| KUNg1     | 25      | 9+5 ns     | 0.00275  |
| KUNg2     | 25      | 9+5 ns     | 0.00275  |
| TR464     | 50      | 9+10 ns    | 0.00275  |

$^a$ CV$_{SA}$ steering rate in ps$^{-1}$.
Appendix A1 – CV<sub>SA</sub> section of the Plumed input files for the simulations of the GB1 β-hairpin. Atom indices refer to all the C<sub>α</sub> atoms in the molecule.

ALIGN_ATOMS LIST <alist>
alist->
9 16 31 55 69 90 102 114 124 138 160 174 194 208 224 238
alist<-

SARMSD LIST <fragments> R_0 0.060 NN 8 MM 12 ANGSTROM_SCALE 0.1 NOPBC ENCODE D B A A L V W O Q A B A B

fragments->
9 16 31 55
16 31 55 69
31 55 69 90
55 69 90 102
69 90 102 114
90 102 114 124
102 114 124 138
114 124 138 160
124 138 160 174
138 160 174 194
160 174 194 208
174 194 208 224
194 208 224 238
fragments<->
Appendix A2 – CV<sub>SA</sub> section of the Plumed input files for the simulations of TR464 (Target rSA encoding). Atom indices refer to the C<sup>α</sup> atoms of loops L1 (18-31) and L2 (38-45).

ALIGN_ATOMS LIST <alist>
alist->
  5  24  35  54  73  88  99  121  128  147  162  172  191  205  227
  237  256  266  277  284  298  314  331  355  365  377  384  395  414  431
  445  462  473  492  509  524  534  553  572  586  610  622  646  668  685
  701  721  745  764  779  791  802  829  835  851  870  897  903  918  933
  943  962  972  986  1005  1022  1041  1051  1065
alist<-

SARMSD LIST <fragments> R_0 0.060 NN 8 MM 12 ANGSTROM_SCALE 0.1 NOPBC ENCODE R R A A K U K R A A A K U K R N

fragments->
  266  277  284  298
  277  284  298  314
  284  298  314  331
  298  314  331  355
  314  331  355  365
  331  355  365  377
  355  365  377  384
  365  377  384  395
  377  384  395  414
  384  395  414  431
  395  414  431  445
  553  572  586  610
  572  586  610  622
  586  610  622  646
  610  622  646  668
  622  646  668  685
fragments<-