Low-resolution description of the conformational space for intrinsically disordered proteins

Daniel Förster, Leo Liberti, Antonio Mucherino, Jung-Hsin Lin, Thérèse E Malliavin, Jérôme Idier

Low-resolution description of the conformational space for intrinsically disordered proteins

Daniel Förster (1), Jérôme Idier (2), Leo Liberti (3), Antonio Mucherino (4), Jung-Hsin Lin (5) and Thérèse E. Malliavin (6,7,8)

(1) UMR7374 Interfaces, Confinement, Matériaux et Nanostructures, Université d'Orléans, France

(2) UMR6004 Laboratoire des Sciences du Numérique de Nantes, France

(3) LIX UMR 7161 CNRS École Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France

(4) IRISA, University of Rennes 1, France

(5) Biomedical Translation Research Center, Academia Sinica, Taiwan

(6) Institut Pasteur, Université Paris Cité, CNRS UMR3528, Unité de Bioinformatique Structurale, F-75015 Paris, France

(7) Laboratoire de Physique et Chimie Théoriques (LPCT), University of Lorraine, Vandoeuvre-lès-Nancy, France.

(8) Laboratoire International Associé, CNRS and University of Illinois at Urbana-Champaign, Vandoeuvre-lès-Nancy, France.

Corresponding authors:

Thérèse E. Malliavin, therese.malliavin@univ-lorraine.fr

Jérôme Idier, jerome.idier@ls2n.fr

Short title

Conformational space of IDPs Sic1 and pSic1

August 29, 2022

Abstract

Intrinsically disordered proteins (IDP) are at the center of numerous biological processes, and attract consequently extreme interest in structural biology. A systematic enumeration of protein conformations, carried out using the TAiBP approach based on the distance geometry, was performed on two proteins, Sic1 and pSic1, corresponding to unphosphorylated and phosphorylated states of an IDP. The populated conformations were then obtained by fitting SAXS curves as well as Ramachandran probability maps, the original finite mixture approach RamaMix being developed for this second task. The similarity between profiles of local gyration radii provides to a certain extend a converged view of the Sic1 and pSic1 conformational space. Profiles and populations are thus proposed for describing IDP conformations. Different variations of the resulting gyration radius between phosphorylated and unphosphorylated states are observed, depending on the set of enumerated conformations as well as on the methods used for obtaining the populations.

Intrinsically disordered proteins (IDP) are at the center of the attention in the structural biology of proteins. Indeed, disordered residues are expected to constitute 35 to 50% of

the human proteome and, depending on the organism type, the overall percentage of amino acids predicted to be disordered ranges from about 12% up to 50%.[1] In addition, the conformational plasticity of the disordered regions of proteins allows them to interact with numerous partners in the cell, as for example for the three intrinsically disordered domains of the tumor protein P53.[2] This moonlighting[3] behavior explains the strong impact of IDPs in cellular signaling, regulation, and control, and the differences observed in their interactomes with respect to globular proteins.[4]

Intrinsically disordered proteins represent a challenge for structural biology for several reasons. In solution, the nuclear Overhauser effects measuring distance between hydrogens is usually not available. On the other hand, crystallization processes are hampered by the conformational disorder, or the variability of conformations in the crystal or in the electron cryogenic maps makes impossible the observation of electronic density for disordered regions. Numerous approaches have been proposed[5–8] for the calculation of protein conformations, based on molecular dynamics or Monte Carlo simulations for generating molecular conformations.

We propose here to explore a new approach for the exploration of the conformational space of IDPs, based on a systematic enumeration of conformations in the frame of the distance geometry problem. We amend here our previous work introducing TAiBP as a new tool to investigate structural ensembles of IDPs in a systematic way, by predicting populations and consequently selecting pools of representative conformations. This approach, initiated as the interval Branch-and-Prune (iBP) algorithm by Mucherino and coworkers,[9] was adapted to the protein molecular modeling as threading-augmented interval Branch-and-Prune (TAiBP).[10,11] Based on distance geometry, TAiBP, explores the entire conformational

space compatible with NMR chemical shifts retaining conformations that are most different from one another yielding thus a diverse set of conformations to be analyzed further. This is in contrast to Monte Carlo methods which are informed by force fields and explore the part of the configurational space that is thermodynamically relevant in more detail. TAiBP was shown recently[12] to allow the analysis of the conformational space of a tandem domain of protein whirlin, in which a disordered linker induces a large orientation variability of two PDZ domains.[13] The application of TAiBP to the tandem domain was made possible by the analysis of unprocessed output of the neural network TALOS-N,[14] the Ramachandran likelihood maps. Indeed, drawing boxes on the most probable regions of these maps, allowed the determination of intervals on backbone angles, which serve as inputs for the TAiBP algorithm. It should be noticed that the approach MERA has been developed[15] for the prediction of the ϕ, ψ distributions for IDPs.

In the present work, we apply TAiBP to a well-know example of IDP.[16,17] The obtained IDP conformations will be filtered and their relative populations determined by BioEn[18] using SAXS data. In parallel, we propose an original method, RamaMix, to select the main conformations, as well as their populations, according the Ramachandran likelihood maps predicted by TALOS-N.[14] The principle of RamaMix is to fit a bivariate, periodic, finite mixture model to the output of TALOS-N. The N terminal fragment of the intrinsically disordered protein Sic1, as well as it phosphorylated form pSic1, each one spanning 90 residues, will be studied.

Sic1 prevents premature S-phase entry in the budding yeast *Saccharomyces cerevisiae* by inhibiting the complex Cdk1-Clb. At the START point in the yeast cell cycle, Sic1 is phosphorylated on three Threonines (residues 7, 35, and 47) and three Serines (residues 71,

78, and 82) in order to be degraded by the proteosome. Sic1 as well as pSic1 were shown[16,19] to contain significant amount of transient secondary structures.

The comparison of repeated runs of TAiBP on Sic1 and pSic1 reveals a good reproducibility of global conformational shape. Qualitatively similar but quantitatively different populations are obtained either by fitting distinct SAXS curves or Ramachandran maps. The sets of individual conformations selected from the fitting of various data are partially distinct, but better convergence is observed for the profiles of local gyration radius. These profiles could be proposed as a low resolution description of the IDP conformational space. Depending on the way the TAiBP conformations are generated, and on the processing method to obtain the populations, different patterns of variations are observed for the resulting gyration radius of Sic1 and pSic1.

Results

Enumeration of protein conformations

The TALOS-N[14] prediction was obtained using the chemical shifts measured for the nuclei Hα, HN, ^{15}N, ^{13}Cα, ^{13}Cβ of Sic1 and pSic1 residues, and was used to determine boxes of ϕ and ψ values, giving the limits in which the conformations will be enumerated. Indeed, from the NMR chemical shifts and the protein sequence information, the TALOS-N neural network predicts the likelihood that a given residue n has backbone torsion angles that fall in any of the 324 voxels, of 20° × 20° each, that make up the Ramachandran map.[14] Following the approach proposed in Ref. 12, we define boxes (Figures S1-S4) using Ramachandran regions

displaying largest likelihood for the TALOS-N prediction, and corresponding supposedly to protein conformations populated in solutions.

rev1mera In order to probe the reliability of the (ϕ,ψ) boxes obtained from the TALOS-N likelihood maps, these boxes were compared to the predictions performed using the approach MERA,[15] which predicts the residue-by-residue Ramachandran map distributions for disordered proteins using short-range NOEs, chemical shifts, J couplings and spectral density derived from the N^{15} relaxation measurement. As only chemical shifts were available for Sic1 and pSic1, the MERA prediction was performed putting to zeros all other possible inputs. The MERA Ramachandran map distributions are plotted for all successful predicted residues, along with the input boxes derived from the TALOS-N prediction (Figures S10 and S11), showing a reasonable agreement between the two methods.

Two replicates of boxes were generated for Sic1 and pSic1, using threshold values of 0.01 and 0.011 on the Ramachandran probability maps as described in section "Extraction of boxes from Ramachandran likelihood" in the Supplementary Material. Using these sets of input boxes, five TAiBP runs were performed, named $Sic1^1$, $Sic1^2$, $pSic1^1$, $pSic1^2$ and $pSic1^3$. $pSic1^3$ The run $pSic1^3$ differs from the others by the procedure for selecting more extended representative conformations after the SOM clustering, as described in the section "Clustering of generated conformations" in the Supplementary Information.

The two replicates of TAiBP calculations introduced in the previous subsection were based on similar numbers of fragments: 14 and 13 for $Sic1^1$ and $Sic1^2$, 17 for for $pSic1^1$ and $pSic1^2$ and 18 for $pSic1^3$ (Table S1). The larger number of fragments used for pSic1 arises from the regions of residues 5-9, 33-37, 45-49, 69-73, 76-84 for which TALOS-N was unable to give a prediction due to the phosphorylated residues and for which generic boxes

(Table S2) were used. These boxes being formed of three components, they increase the combinatorics of the enumeration and shorter fragments have to be used, requiring a larger number of fragments to span the protein sequence.

The boxes used as inputs for the TAiBP runs (Figures S1-S4) are quite similar. The loop region (positive ϕ) is slightly more populated for runs pSic1[1] and pSic1[2]. For the iBP and assembly steps forming the TAiBP approach, the duplicate runs, marked in colors red and green in Figure 1, produces parameter values similar in most of the protein sequence.

For the iBP steps, three parameters were compared (Figure 1, first and second lines) along the residue number located at the middle of each fragment: the number of individual iBP runs (N_{iBPrun}), the number of saved conformations ($N_{iBPconf}$) and the number of obtained conformations after clustering ($N_{clustiBP}$). The three analyzed parameters are located in similar ranges for all calculations. Nevertheless, N_{iBPrun} displays the largest observed values (3888) around the positions of phosphorylated Threonines in agreement with the larger generic boxes used in these protein regions (Table S2). Such increase is not observed for phosphorylated Serines due to shorter fragments used in the region 50-90 (Table S1). For every calculation, $N_{iBPconf}$ is smaller than 10^9, which is the input given for the maximum number of solutions: all individual iBP trees have been thus completely parsed. The $N_{iBPconf}$ profiles display smaller values, mostly in the range 10^6-10^7, for all calculations in the region of residues 60-90. At the contrary of $N_{iBPconf}$, the numbers of clustered conformations ($N_{clustiBP}$) display relatively flat profiles for Sic1, but a decrease in the number of conformations of pSic1 in the region of residues 60-90. This larger reduction of conformations due to the clustering is the sign that the conformations generated by iBP in the region 60-90 are more diverse in Sic1 than in pSic1. In all calculations, the C terminal fragments

which are smaller than the others (Table S1), display smaller N_{iBP}, $N_{iBPconf}$ and $N_{clustiBP}$. The results obtained for the run pSic1[3] (blue crosses) are quite similar to those of the run pSic[2], which is not suprising as the fragment definition are the same, except around residues 40-60 (Table S1).

Three parameters are plotted (Figure 1, third and fourth lines) along the assembled fragments: the number of conformations rejected due to Cα atoms closer than 1Å ($N_{clashes}$), the number of saved conformations (N_{saved}) and the number of clustered conformations (N_{clust}).L ooking at the relative ranges of values of $N_{clashes}$ and N_{saved}, between 10% and 15% of the assembled fragments are rejected due to the steric clashes. The profiles of N_{clust} are different for Sic1 and pSic1, as the number of clustered conformations increases up to the last fragment, whereas this number already starts to decrease in the region of residues 60-90 in pSic1. This effect can be put in parallel with the decrease of $N_{clustiBP}$ in the same region during the iBP step. The last fragments of proteins have strong decreasing effects on N_{clust}, due to their smaller size (Table S1) which induces probably less variability in the generated conformations. Smaller numbers of clashes are mostly obtained for run pSic1[3] (blue crosses), which is probably due to the larger extension of conformations. The number of saved conformations is often larger than in other runs, which may be a consequence of the smaller numbers of clashes. Unsurprisingly, the number of clustered conformations increases along the number of saved conformations.

After the distance geometry calculations, a refinement by molecular dynamics (MD), described in section "Molecular dynamics refinement in implicit solvent" of Supplementary Material, was applied to the generated conformations. The protein conformations do not vary much during MD trajectories. Indeed, the cumulative sums of differences between initial

and final values of backbone angles produce values in the range 4.2-4.9° for ϕ and in the range 0.04-2.4° for ψ. Similarly, the average coordinate RMSD between the initial and final frames of the refinement trajectories are 0.6 Å for the four runs Sic1^1, Sic1^2, pSic1^1 and pSic1^2. The drift is larger for pSic1^3, with backbone angle values in the ranges -24 to 6° for ϕ and -40 to -1° for ψ, and an average coordinate RMSD of 0.7 Å. The conformations displaying potential energy smaller than -50 kcal/mol for the runs Sic1^1 and Sic1^2 and smaller than -600 kcal/mol for the runs pSic1^1 and pSic1^2, were selected for further analyses. This selection produces sets of 98 (Sic1^1), 133 (Sic1^2), 161 (pSic1^1), 121 (pSic1^2) and 148 (pSic1^3) conformations.

Comparison of the conformations between duplicate TAiBP runs

The distributions of gyration radii R_g and maximal diameters D_{max} (Figure 2 top) are quite similar for the duplicate runs on Sic1 and pSic1. The global envelope of generated conformation is thus reproducible between the replicated TAiBP runs. The distribution of gyration radii R_g and maximal diameters D_{max} have been plotted in magenta for the run pSic^3 to display the larger extension of the obtained conformations.

The individual conformations generated for the duplicated runs of Sic1 and pSic1 were then compared by calculating the two-by-two coordinate root-mean-square deviation (RMSD, Å). The distributions of the minimum RMSD values (Figure 2 bottom left panels) observed for each conformation of one run to the conformations of the other run are quite reproducible whatever is the performed comparison. They display sets of values in the ranges of 8-16 Å for both proteins, with a maximum around 11 Å for Sic1 and around 10 Å for pSic1. This drift of pSic1 maximum towards smaller values agrees with a larger compaction of pSic1

conformations. Nevertheless, the range 8-16 Å of RMSD values means that the individual conformations of a given run are not reproducible in the replicated run. This excludes a high resolution determination of representative conformations which is not surprising due to the enormous size of the conformational space to explore and the heavy clustering procedure used along the TAiBP approach.

By analogy to the cross-sectional gyration radius, we propose here the profiles of local gyration radii to describe the local variation in the shape of conformations. These profiles P_q of local gyration radii are calculated along residue number n for each conformation q in the following way:

$$P_q(n) = \sqrt{\frac{1}{N_n} \sum_{i=n-N_{win}}^{n+N_{win}} (\mathbf{X}_i - \mathbf{X}_n^{ave})^2} \quad (1)$$

where $\mathbf{X}_i$ represents the vector of atomic coordinates for the backbone atoms of residue i in the range $n - N_{win}$, $n + N_{win}$, and N_{win}=5 is the residue window around n on which a local gyration radii is calculated, N_n being the number of backbone atoms located in this window. $\mathbf{X}_n^{ave}$ is the coordinate vector of the centroid of the atomic coordinates of the backbone atoms of residues in the range $n - N_{win}$, $n + N_{win}$.

The profiles P_q of local gyration radii were compared two-by-two between conformations using Euclidean distance. The distributions of minimal distance between P_q (Figure 2 bottom right panels) are similar to those observed for minimal RMSD values (Figure 2 bottom left panels), but are drifted toward ranges of 4-11 Å. The comparison between local gyration profiles shows that one half of the obtained conformations displays a distance between profiles located between 1/6 and 1/3 of the average gyration radius. The profile distance smaller than the average gyration radius is the sign of a reduced variation of the profiles P_q with respect to

the coordinate RMSD. The P_q profiles, inspired by the cross-sectional gyration radius, seems thus to capture a better convergence between the duplicate runs than the coordinate RMSD. In the following, the conformations selected by the fitting of SAXS curves and Ramachandran maps will be compared through their P_q profiles.

Quite similar global shape of conformations are populated in the duplicated TAiBP runs. The profiles P_q of local gyration radii display also some similarity. But, the comparison of atomic coordinates reveals a large variability of the individual conformations selected by the TAiBP approach, which is not surprising due to the enormous considered conformational space.

Validation of the finite mixture model on synthetic data

Once a set of conformations have been selected using TAiBP, one needs to detect the conformations significantly populated and to evaluate their relative populations. Indeed, the systematic enumeration along all possible combination of the ϕ/ψ boxes induces the generation of conformations spanning a space possibly larger than the conformations effectively populated. The populations were determined, from one side, using BioEn[18] on SAXS data, and on the other side, using on the Ramachandran maps, a finite mixture model, RamaMix, specially developed for this purpose. We first present in this section a validation of RamaMix on synthetic data.

A pseudo Ramachandran map has been generated by randomly choosing up 15 couples of ϕ, ψ values located in most populated regions of the Ramachandran map (Figure S5). Several sets of more or less scattered values, represented by different colors, have been generated,

to investigate the effect of conformational superimposition on the population determination. Corresponding populations were also chosen randomly (see caption of Figure S5). Noise levels of 0.2, 1, 2, 3, 5 and 10 were added to the histogram obtained from the pseudo Ramachandran map, the maximum value of the histogram being around 15. The starting points for each RamaMix run was the ϕ_0, ψ_0 values from the synthetic Ramachandran plot, and random population values. During each RamaMix run, several upper limits were imposed to the drift of the backbone angles during the optimization, with values of: 1°, 10°, 20°, 30°, 40° and 50°. For each Ramachandran synthetic map, each noise level and each drifting limit value, one hundred runs are performed producing sets of backbone angles (ϕ_0 and ψ_0) (Eq. 7), von Mises parameters (Eq. 8) ($\kappa 1$, $\kappa 2$ and ρ) and populations γ_q (Eq. 2). Over the 12600 individual RamaMix runs, only 275 runs were terminated without convergence of the optimization. Averages and standard deviations were calculated from the sets of obtained parameters. The differences between the averaged and the input values, as well as the standard deviations (Figure 3) are used to evaluate RamaMix.

The differences between average and initial populations (Figure 3E) as well as the standard deviations of populations are mostly smaller than 30%. Thus, the determination of populations is not much influenced by the level of noise, but the population values are rather qualitative. Interestingly, the standard deviation is of the order of value of the difference.

The efficiency of the determination of backbone angles (Figure 3A-D) for noise levels of 0.2, 1, 2, 3 and 5, is not much influenced by the scattering of synthetic Ramachandran maps, but rather by the drifting limit imposed on the ϕ, ψ values. Increasing the allowed drift induces larger differences and standard deviations: this would support not allowing large drift for the calculations. Interestingly, for the large scattered Ramachandran map

(bullets in Figure 3), the effect of a large drift is more pronounced than for other synthetic Ramachandran maps. For most of the cases, the standard deviations display larger values than the difference: allowing a drift induces more error on the precision of the calculation than on the average value of angles.

The parameters describing the von Mises distribution (Figure 3G-L) display contrasted results: the differences are larger for ρ than for $\kappa 1$ and $\kappa 2$. For $\kappa 1$ and $\kappa 2$, the standard deviations are much larger than the differences whereas they are similar for ρ. The differences between ρ and $\kappa 1$ and $\kappa 2$, arise from the definition of these parameters (Eq. 8) in which ρ occupies a different place than $\kappa 1$ and $\kappa 2$.

Determination of populations

The TAiBP conformations were fitted to the SAXS curves and Ramachandran probability maps using BioEn[18] and RamaMix.

The following sets of conformations were processed: the conformations obtained from runs Sic1^1, Sic1^2, pSic1^1, pSic1^2 and pSic1^3, as well as two mixed sets of conformations obtained by pooling the conformations from pSic1^1 and pSic1^3 and the conformations from pSic1^2 and pSic1^3. These mixed sets of conformations will be denoted pSic1^{13} and pSic1^{23} and encompass respectively 309 and 269 conformations.

BioEn calculations were performed using each of the three SAXS curves available (Tables 1, 2 and S5). The populations larger than 1% found for a given TAiBP run and the fitting of a given SAXS curve, reveal that the same conformations are repeatedly selected: the conformation numbers selected more than once have been written in bold in the Tables.

Most of the conformations selected only once, display populations smaller than 15%. But the populations vary significantly from one analysis to another as for example for the conformation 109 from the run Sic1^2 (Table 1B) which display populations of 26.6, 40.9 and 43.8% for the three SAXS curve processing. Normalized χ^2 values smaller than one are found for each calculation along with null final S_{KL} values, in agreement with the definition of S_{KL} as the Kullback-Leibler divergence.[18,20]

Tables 3 and S6 present the populations obtained by RamaMix from the fitting of the Ramachandran probability maps on the same sets of conformations. The variations of backbone angles ϕ and ψ during the RamaMix optimization are smaller than 0.25° for ϕ and 0.1° for ψ during all considered calculations. These variations are smaller for pSic1^3 with 0.12° and 0.03° for ϕ et ψ, and even smaller for the mixed pools of conformations with 0.06 and 0.02○. Among six of the seven sets of TAiBP conformations, conformations (marked in bold) already repeatedly selected by BioEn, were also selected by RamaMix (Tables 1, 2 and S5).

Similarly to the populations obtained by BioEn between the different SAXS data, the populations found using RamaMix are quite different than the ones determined by BioEn. Another difference between BioEn and RamaMix processing is the smaller number of conformations selected by RamaMix, it can arise from the essential difference between the data, as the SAXS curves describe a global picture of the conformations whereas the Ramachandran maps give a local information. A smaller number of conformations are selected from the sets where more extended conformations were included: this may be due to the important conformational drift induced by the systematic choice of extended conformations during the clustering step (Supplementary information section "Clustering of generated conformations").

In order to compare the conformations selected by BioEn on the three SAXS curves, several curves superimpositions have been realized. The superimposition of SAXS curves reconstructed from the conformations selected from Tables 1 and 2 to the corresponding fitted SAXS curves (Figure S6) displays a reasonable agreement with χ^2 in the range 0.6-2.06. These values are larger than the ones given in the Tables 1 and 2, due to the fact that conformations displaying populations smaller than 1% have been removed. Besides, a comparison of all sets of BioEn conformations with all SAXS curves (Table S3) reveals that the conformations and populations determined from the fit of one SAXS curve display χ^2 values with another SAXS curve going up to 4.42. The variability between the three SAXS curves induces thus a drift between conformations and populations selected from the fit of each curve.

A similar comparison has been performed between the SAXS curves and the conformations and populations determined with RamaMix (Figure S7). In this comparison, the χ^2 values are in the range 0.98-4.24 which is similar to what is observed for BioEn selected conformations in Table S3. The variability between the fits to Ramachandran maps and SAXS curves is thus similar to the variability of fit between different SAXS curves.

In order to investigate the possible convergence between the different conformations and populations detected using BioEn and RamaMix, systematic comparison of Euclidean distances between profiles of local gyration P_q (Eq. 1) was performed (Figures 4, S8 and S9) The Euclidean distances within each set of conformations selected by BioEn reveal (Figure 4, three left columns) that, for several cases, distances smaller than 8 Å are observed between different conformations. In many cases, such small distances are observed between conformations (labeled with asterisk Figure 4) for which populations smaller than 10% are observed.

The comparison of profiles P_q (Eq. 1) between conformations selected by RamaMix (Figure 4, right column) reveals two features. When few conformations have been selected (as for Sic1^2 and pSic1^2), the distances between their profiles P_q are larger than 8 Å. When more conformations are selected (as for Sic1^1 and pSic1^1), profile distances smaller than 8 Å are observed. The small P_q distances reveal a certain convergence of the profiles P_q.

The comparison of conformations selected by BioEn and RamaMix as well as the comparison between conformations selected from the fit of the various SAXS curves is displayed in Figures S8 and S9. A close inspection of these distance matrices for BioEn conformations (Figure S8) shows that, if one excludes the conformations populated less than 10%, there are only three conformations displaying profile distances larger than 8 Å and selected in two distinct BioEn runs: (i) for Sic1^2, the conformation 106 selected on the SAXS curve BioEn1 compared to the conformations selected on the two other SAXS curves; (ii) for pSic1^2, the conformation 74, selected in the runs BioEn1 and BioEn2, and compared to the conformations selected from the run BioEn3; (iii) for pSic1^2, the conformation 139 selected from the run BioEn3, and compared to the conformations from the run BioEn1. Overall, most of the conformations populated more than 10% from the fitting of different SAXS curves display profile distances smaller than 8 Å, supporting a convergence of the profiles in the different fits.

On the other hand, the comparison between BioEn and RamaMix fitting (Figure S9) displays contrasted behaviors between the duplicated TAiBP runs. For Sic1^2 and pSic1^2, all RamaMix conformations display profiles closer than 8 Å to the profiles of BioEn conformations. For pSic1^1, this is also the case for three RamaMix conformations (16, 98, 101) over five. For Sic1^1, only the conformation 79 displays profile distances smaller than 8 Å for the

three comparisons.

Examples of profiles P_q superimposition have been chosen accordingly to the values of their distances (Figure 5) and give an estimation of the connection between the information related to atomic coordinates and the distance between the profiles. These examples represent distances in the 4.05-7.88 Å range. The examination of Figure 5 reveals that the profile peaks are mostly located at similar places in the protein sequence. This gives a qualitative description of the conformations separated in extended regions (profile maxima) and in aggregated regions (profile minima).

The description of IDP conformations by P_q profiles permits to detect some convergence between the various Bioen fits and also between RamaMix and BioEn fit. This is extremely encouraging due to the enormous conformational size and to the heterogeneity of the measurements (SAXS, NMR) used for fitting the populations. Nevertheless, this comparison remains extremely qualitative, and far from any high resolution description. It could represent a starting point for deeper investigation of IDP conformations.

Comparison with PED conformations and link with biological activity

The sets of Sic1 and pSic1 conformations selected from the fitting of SAXS curves and of Ramachandran probability maps, were compared to the sets of protein conformations deposited in the Protein Ensemble Database `proteinensemble.org`.[21]

rgyr The values of the resulting gyration radii were calculated (Table 4) from the populations determined by BioEn and RamaMix, and using the individual gyration radii of selected

conformations. Globally, the resulting gyration radii display orders of values agreeing with the measurements reported in Figure 2E of Ref. 17. For the conformations extracted from the data-sets Sic1^1 and Sic1^2, the resulting gyration radii agree with the measurement of 3.0 $\pm$ 4.1 Å given in Figure 2E of Ref. 17. But, for pSic1^1 and pSic1^2, the resulting gyration radii are smaller than the measurements of Ref. 17: this is particularly true for the BioEn processing whereas the RamaMix processing displays values closer to those of Gomes et al.[17] On the more extended conformations (pSic1^3), all resulting gyration radii are significantly closer to the Gomes et al[17] measurements, for BioEn and RamaMix processing. Pooling pSic1^3 with the conformations of pSic1^1 or pSic1^2 (sets pSic1^{13} and pSic1^{23}) produces different effects for BioEn and RamaMix processing. For these mixed data-sets, the resulting gyration radii obtained from the BioEn processing (range 27.2-28.1 Å) decrease to reach a level just slightly larger than the one obtained for pSic1^1 and pSic1^2 (range 26.1-27.9 Å). At the contrary, the gyration radii obtained by RamaMix processing are the same than the ones obtained for pSic1^3. Overall, the BioEn processing is more sensitive than RamaMix to the presence of conformation with lower gyration radii. The discrepancy of the results obtained here on pSic1 with those shown in Ref. 17 arises in part from the tendency to obtain smaller gyration radii by processing of the whole SAXS curve with respect to the larger gyration radii obtained by using the Guinier approximation within the low-q region of the SAXS data.

The selected TAiBP conformations were also compared to the PED conformations by realizing a principal component analysis (PCA) of the atomic coordinates. The coordinates projected on the first and second or on the second and third component (Figure 6) reveal that most of the TAiBP conformations are located in similar space regions than the PED conformations.

rev3hbond The presence of phosphorylated residues decreases obviously the global charge of pSic1 with respect to Sic1. It was pointed out that the induced variation in long-range electrostatic interactions plays a role in the electrostatic interaction of pSic1 with its target Cdc14.[22] But, the variations of charges gives also various opportunities for the formation of hydrogen bonds, which were analyzed for the whole set of conformations from the TAiBP runs as well as for the PED sets of conformations. All PED conformations were submitted to the same refinement than the one used on TAiBP conformations described in the section "Molecular dynamics refinement in implicit solvent" of the Supplementary Material, using positional restraints on protein backbone atoms with a constant force of 50 kcal/mol. The cumulative variations of ψ and ϕ angles during the refinement were in the range 0.8-1.2° for ϕ and in the range 0.3-0.8° for ψ, and the coordinate RMSD around 0.1 Å. All hydrogen bonds were detected on the refined PED conformations as well as on the TAiBP conformations. Cumulative contact maps (Figure S12) display these hydrogen bonds, according to the involved residues, the hydrogen bonds involving phosphorylated residues being colored in magenta. The inspection of these contact maps reveals that the PED and TAiBP conformations display distinct tendencies. Long range hydrogen bonds involving phosphorylated residues are more present in the less extended set of TAiBP conformations $pSic^1$ and $pSic^2$. On the other hand, the conformation set PED161 of pSic1 displays the largest number of long-range hydrogen bonds involving the sidechains of phosphorylated residues. Thus, the presence of phosphorylated residues can induce the appearance of long-range hydrogen bonds whatever is the variation of resulting gyration radius.

Discussion

The TAiBP approach enumerating the protein conformations in the frame of the distance geometry problem has been used for describing the conformational space of two IDPs, Sic1 and pSic1, corresponding to the unphosphorylated and phosphorylated states of a disordered region involved in the control of S phase in the cellular cycle. The present study represents a test for a new approach able to systematically enumerate protein conformations in the frame of a distance geometry approach. Indeed, up to now, most of the approaches for calculating IDP conformations are based on Monte Carlo approaches[8,23,24] which do not guarantee an exhaustive exploration of the conformational space.

One should notice that TAiBP overcome the exponential complexity of the branch-and-prune algorithm, due to the parallel calculations on fragments, to the rejection of too close solutions, and to the systematic use of clustering. A major advantage of this approach is the availability of a systematic procedure. Nevertheless, the obtained conformations are only representative conformations, and the represented conformational space has still to be defined.

The use of TAiBP approach permits to avoid the question of the convergence of solutions for protein conformations. The introduction of the profiles of local gyration radii P_q along with their relative populations allows the reintroduction of a convergence criterion into the problem, and this is essential for validation purposes. In the present work, the validity of this convergence criterion has been assessed by the comparison of the profiles P_q obtained from independent fits. In that frame, the profile of local gyration radii could be proposed for describing the IDP conformational space: the knowledge, even qualitative, of the profiles

should provide geometrical restraints allowing a more precise exploration of the conformational space. label1reviewer1 The profiles are closer between the conformations selected by the various fits of SAXS curves, than between the conformations selected by BioEn and RamaMix. This is expected as the various fits of SAXS curves use an homogeneous information. More surprisingly, similar profiles are observed between conformations selected by RamaMix and BioEn, for the runs Sic1^2 and pSic1^2, and for many conformations of the runs Sic1^1 and pSic1^{1}s.

One specific advantage of the mixture method RamaMix for determining populations of conformations from the likelihood Ramachandran maps is that it has a larger domain of applicability that the BioEn method based on the SAXS curves. Indeed, polydispersity in protein solutions can make difficult to extract conformational information from the SAXS curve. In addition, the chemical shifts from which the likelihood Ramachandran maps are extracted, can be measured in solution as well as in in-cell NMR or for an IDP sequence inserted in a larger protein.[25]

rgyr The comparison of the resulting gyration radii obtained from the BioEn and RamaMix processing with the values measured in Ref. 17 showed (Table 4) that various ranges of gyration radii are obtained, depending on the clustering procedure in TAiBP, as well as on the method for SAXS processing. In particular, the processing of the whole SAXS curve with BioEn displays a tendency to underestimate the gyration value with respect to the processing of the Guinier curve. The determination of populations from the Ramachandran probability maps, using RamaMix, seems to be less prone to the underestimation of the gyration radius.

song The discrepancy between resulting gyration radii obtained by processing the whole

SAXS curve (BioEn) or restricting the analysis to the low-q region (Guinier approximation[17]) agrees with independent calculations performed using coarse-grained protein model,[26] in which various distribution of gyration values produce very similar SAXS spectra (Figure 10a of[26]) or different disordered ensemble produce similar Kratky plots (Figure 13 of[26]).

Methods

Origins of the data

Three sets of conformations for Sic1 and pSic1 were available from the Protein Ensemble Database (PED) `proteinensemble.org`:[21] PED159 and PED160 for pSic1 and PED161 for Sic1.[17] The residue numbering used here is the one proposed in the PED. The NMR chemical shifts were downloaded from the Biological Magnetic Resonance Data Bank (BMRB)[27] as entries: 16657 for Sic1[17] and 16659 for pSic1.[19] The SAXS data-sets recorded as triplicate sets in the conditions described in Ref.[17] were provided by Tanja Mittag.

Enumeration of conformations using TAiBP

The protein conformations have been enumerated using the recently proposed TAiBP approach,[10–12] which generalizes the interval branch-and-prune (iBP) algorithm[9,28–31] so as to overcome the combinatorial barrier arising from the enormous space of IDP conformations.[32] TAiBP is composed of two steps: (i) the enumeration of conformations for peptide fragments (Table S1) spanning the studied protein using individual iBP calculations; (ii) the enumeration of Sic1 and pSic1 conformations by systematic assembly of fragment conformations in

a way similar to what is used in the field of protein prediction.[33]

The boxes of backbone angles ϕ and ψ used as inputs for the iBP step were determined from the Ramachandran likelihood maps predicted by TALOS-N[14] (see section "Extraction of boxes from Ramachandran likelihood maps" and Figures S1-S4 in Supplementary Material). The ϕ/ψ boxes were systematically combined by permutation to prepare individual iBP calculations as in Ref 11. The enumeration of conformations is realized by the building of a tree, each node of the tree corresponding to an atomic position. The tree building allows the enumeration of the various possibilities for atom positions (branching step) whereas additional geometric information is used to accept or reject a newly built branch (pruning step). As the angles ϕ and ψ are straightforwardly related to distances between atoms C and N of residues successive in the sequence,[11,12] the discretization of intervals of these angles is used in the branching step. In the iBP step, the pruning was applied by preventing atoms to be closer than the sum of their van der Waals radii and by checking that the improper angle values are correct. In addition, each solution displaying a coordinate root-mean-square deviation (RMSD) smaller than 2 Å with the previously stored solution, is rejected. The details of the iBP step calculation are described in the section "Enumeration of conformations" of the Supplementary Material.

The assembly step is also performed with a branch-and-prune approach using as elementary blocks, not the atoms, but the fragment conformations previously determined during the iBP step. Two peptide fragments are assembled by superimposing the three last and initial residues of the fragments successive in the protein sequence. The fragments are then merged in the following way: the atom at which the smallest distance was observed between corresponding atoms in the two peptides was used to decide where to stop with the first

peptide and to continue with the second one. The assembled conformations in which Cα atoms closer than 1 Å are observed, were pruned from the calculation. The fragment assembly was implemented using python scripting based on the MDAnalysis[34,35] and numpy 1.7.1[36] packages.

To scale down the combinatorial explosion of the calculation, a clustering approach based on **S**elf-**O**rganizing **M**aps (SOM)[37–40] was systematically applied to the generated sets of conformations larger than 100 during the iBP and assembly steps. The details of this approach are described in the section "Clustering of generated conformations" of the Supplementary Material.

After the assembly step, the sidechains have been added to the conformation backbones, and the conformations were refined by molecular dynamics simulations as described in the section "Molecular dynamics refinement in implicit solvent" in the Supplementary Material.

Determining the population from Ramachandran maps

The approach RamaMix, based on a finite mixture model, was designed to determine the populations of conformations by fitting on the Ramachandran probability maps. The setting-up of this approach is based on the hypothesis that the likelihood maps describing the likelihood of the TALOS-N prediction[14] can be transformed by normalization into the probability density of the presence of ϕ and ψ values in the set of conformations populated in solution.

Consequently, for each residue n, the Ramachandran probability map is denoted as a 2D probability density $p^n(\phi, \psi)$, modeled as a mixture of probability densities $p^n_q(\phi, \psi)$ deter-

mined on each conformation q:

$$p^n(\phi,\psi) = \sum_{q=1}^{Q} \gamma_q p_q^n(\phi,\psi) \tag{2}$$

where $\gamma_q \geq 0$ is the population of conformation q in solution.

RamaMix intends to decompose the probability map $p^n(\phi,\psi)$ according to Eq. 2 along the following lines: (i) the total number Q of conformations is taken from output of TAiBP; (ii) for each conformation q and each residue n, $p_q^n(\phi,\psi)$ is a periodized Gaussian density characterized by averaged values of backbone angles (ϕ_q^n, ψ_q^n) and by a 2×2 covariance matrix C_q^n; (iii) the populations γ_q have to be adjusted in order to maximize the fit between the Ramachandran probability maps and the mixture model (Eq. 2).

The Ramachandran probability maps $p^n(\phi,\psi)$ are jointly fitted to the finite mixture model (Eq. 2) using a discrepancy measure between both probability maps given by the Kullback-Leibler divergence:

$$D_{KL}(p_1||p_2) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx. \tag{3}$$

Calculations detailed in the Supplementary material (sections "Determination of the populations from the Ramachandran maps" and "Maximum likelihood estimation for bivariate sine mixtures") show that using the Kullback-Leibler divergence is equivalent to the maximization of the log-likelihood of the data:[41]

$$\mathcal{L}(y;\theta) = \sum_{n=1}^{N} \sum_{m=1}^{M} \ln p^n(\phi_m,\psi_m). \tag{4}$$

For the sake of clarity, let us first introduce a standard, non-periodized Gaussian density $p_q^n(\phi,\psi)$ for the residue n in conformation q:

$$p_q^n(\phi,\psi) = \frac{1}{2\pi} \det(C_q^n)^{-1/2} \exp(-V_q^n(\phi,\psi)) \tag{5}$$

where $V_q^n(\phi,\psi)$ represents the free energy surface for the basin around the conformation q. The free energy surface is described in the frame of an elastic network model on the backbone dihedral angles:[42–45]

$$V_q^n(\phi,\psi) = \frac{1}{2}\theta_q^t [C_q^n]^{-1}\theta_q \tag{6}$$

where: $\theta_q = (\phi - \phi_q^n, \psi - \psi_q^n)^t$, ϕ_q^n and ψ_q^n are the values of dihedral angles of the residue n in the conformation q and C_q^n is the corresponding covariance. The software IMOD[42] was used for determining the full Hessian (N,N) (N is the total number of residues in the protein) matrix H_q along the backbone dihedral angles. The Hessian matrix is then inversed to produce: $C_q = H_q^{-1}$. The covariance matrix C_q^n of the angles ϕ and ψ of the considered residue n is the (2,2) sub-matrix of C_q, centered on the two ϕ_q^n and ψ_q^n angles. The inverse of this matrix $[C_q^n]^{-1}$ is used in Eq. 6.

As the protein conformations are described by couples of angles, we must consider that the support of the probability densities $p_q^n(\phi,\psi)$ is a torus, i.e., that they are doubly circular. Following,[46–48] we replaced Eq. 6 by a bivariate extension of the von Mises distribution, as being more easily tractable than a Gaussian density wrapped on the torus. More precisely, we adopt a bivariate periodic sine model:[46]

$$p(\phi,\psi) = \frac{1}{T}\exp(W(\phi-\phi_0,\psi-\psi_0)) \tag{7}$$

with

$$W(\phi,\psi) = \kappa_1\cos\phi + \kappa_2\cos\psi + \lambda\sin\phi\sin\psi, \tag{8}$$

$\kappa_1,\kappa_2 \geq 0$ and $\lambda^2 < \kappa_1\kappa_2$. According to Ref. 46, the integration constant T is expressed as

an infinite series, depending on parameters $(\kappa_1, \kappa_2, \lambda)$:

$$T = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2} \right)^m I_m(\kappa_1) I_m(\kappa_2) \tag{9}$$

where I_m denotes the modified Bessel functions of the first kind of order m.[49]

In Ref.,[46] expressions of $(\kappa_1, \kappa_2, \lambda)$ are given as functions of the parameters $(\sigma_1^2, \sigma_2^2, \rho)$ of a bivariate Gaussian where $\rho \in (-1, 1)$ denotes the normalized correlation coefficient between the two components of the bivariate Gaussian:

$$\sigma_1^2 = \frac{\kappa_2}{\kappa_1\kappa_2 - \lambda^2}, \quad \sigma_2^2 = \frac{\kappa_1}{\kappa_1\kappa_2 - \lambda^2}, \quad \rho = \frac{\lambda}{\sqrt{\kappa_1\kappa_2}}. \tag{10}$$

These expressions are valid only in the case where σ_1^2 and σ_2^2 are small. They are easily inverted as

$$\kappa_1 = \frac{1}{\sigma_1^2} \frac{1}{1-\rho^2}, \quad \kappa_2 = \frac{1}{\sigma_2^2} \frac{1}{1-\rho^2}, \quad \lambda = \frac{1}{\sigma_1\sigma_2} \frac{\rho}{1-\rho^2}. \tag{11}$$

Using (11), we can replace a Gaussian mode p_q^n by a periodized version, with approximately the same location and the same spread. In the following, we will describe basin shapes around conformations using the triplets of parameters $(\kappa_1, \kappa_2, \rho)$ rather than $(\kappa_1, \kappa_2, \lambda)$, since $\rho^2 < 1$ is a simpler constraint than its counterpart on λ.

A well-known local optimization scheme to identify finite mixture models by maximum likelihood is the Expectation-Maximization (EM) algorithm.[50,51] Unfortunately, the M step of the EM has no analytical expression in the case of mixtures of bivariate Von-Mises densities. Therefore, we have performed local optimization based on L-BFGS-B[52] instead, given that both the likelihood and its gradient can be evaluated efficiently, and that some parameters are subject to box constraints. The implementation details and equations are given in the sections "Determination of the populations from the Ramachandran maps" and "Maximum likelihood estimation for bivariate sine mixtures" of the Supplementary Material.

By optimization of the log-likelihood, the RamaMix approach will thus produce the Q normalized populations γ_q, the $Q \times N$ couples of backbone angles ϕ_q^n and ϕ_q^n, as well as the $Q \times N$ triplets $(\kappa_1^n, \kappa_2, \rho_q^n)$ describing the von Mises distributions. The calculations were performed starting from the ϕ and ψ values observed in the set of TAiBP conformations, complemented by von Mises parameters allowing us to approximate the Gaussian distributions determined by IMOD. Moreover, the variation of ϕ and ψ values was limited by a threshold of 15° during the optimization in order to avoid inappropriate drift.

The RamaMix approach was implemented in Fortran90, and the software is available at `github.com/tmalliavin/RamaMix`.

Determining the populations from SAXS data

The software BioEn 0.1.1[18] was used in order to determine the populations from SAXS data. On each considered conformation, theoretical SAXS curves were calculated using CRYSOL[53] available in the package ATSAS 3.0.3[54] with 847 points, a maximum scattering vector of 0.503 nm^{-1} and a maximum order of harmonics of 18. A 1D cubic interpolation[55] was used to obtained the theoretical SAXS values at the same sets of scattering vectors q than the ones at which the experimental SAXS curve was recorded.

The processing with BioEn was performed in the following way. For each TAiBP run and each SAXS curve, the optimization was run for 1000 steps using the GSL library.[56] Ten runs were performed independently on all considered conformations, and the subset of conformations for which the sum of observed populations is larger than 0.01, was selected. Ten additional BioEn runs were performed on the subset of conformations, and from the

results of these ten repetitions, average values and standard deviations were computed for the populations.

Acknowledgments

The project ANR-19-CE45-0019 (multiBioStruct) is acknowledged for funding, as well as Institut Pasteur, CNRS, Ecole Polytechnique and University of Rennes. Tanja Mittag is acknowledged for providing SAXS data recorded as triplicate sets in the conditions described in Ref. 17. Cyprien Bertran is acknowledged for fruitful discussions.

Availability of Data and Materials

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References

[1] Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**, 553–584 (2014).

[2] Kumar, A., Kumar, P., Kumari, S., Uversky, V. N. & Giri, R. Folding and structural polymorphism of p53 C-terminal domain: One peptide with many conformations. *Arch Biochem Biophys* **684**, 108342 (2020).

[3] Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem Rev* **116**, 6424–6462 (2016).

[4] Teilum, K., Olsen, J. G. & Kragelund, B. B. On the specificity of protein–protein interactions in the context of disorder. *Biochemical Journal* **478**, 2035–2050 (2021).

[5] Bernadó, P. *et al.* A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* **102**, 17002–17007 (2005).

[6] Allison, J. R., Varnai, P., Dobson, C. M. & Vendruscolo, M. Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements. *J Am Chem Soc* **131**, 18314–18326 (2009).

[7] Fisher, C. K., Huang, A. & Stultz, C. M. Modeling intrinsically disordered proteins with bayesian statistics. *J Am Chem Soc* **132**, 14919–14927 (2010).

[8] Krzeminski, M., Marsh, J. A., Neale, C., Choy, W. Y. & Forman-Kay, J. D. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* **29**, 398–399 (2013).

[9] Lavor, C., Liberti, L. & Mucherino, A. The interval Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances. *J Glob Optim* **56**, 855–871 (2013).

[10] Worley, B. *et al.* Tuning interval Branch-and-Prune for protein structure determination. *Journal of Global Optimization* **72**, 109–127 (2018).

[11] Malliavin, T. E., Mucherino, A., Lavor, C. & Liberti, L. Systematic Exploration of Protein Conformational Space Using a Distance Geometry Approach. *J Chem Inf Model* **59**, 4486–4503 (2019).

[12] Malliavin, T. E. Tandem domain structure determination based on a systematic enumeration of conformations. *Sci Rep* **11**, 16925 (2021).

[13] Delhommel, F. *et al.* Structural Characterization of Whirlin Reveals an Unexpected and Dynamic Supramodule Conformation of Its PDZ Tandem. *Structure* **25**, 1645–1656 (2017).

[14] Shen, Y. & Bax, A. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol* **1260**, 17–32 (2015).

[15] Mantsyzov, A. B. *et al.* A maximum entropy approach to the study of residue-specific backbone angle distributions in α-synuclein, an intrinsically disordered protein. *Protein Sci* **23**, 1275–1290 (2014).

[16] Mittag, T. *et al.* Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**, 494–506 (2010).

[17] Gomes, G. W. *et al.* Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc* **142**, 15697–15710 (2020).

[18] Köfinger, J. *et al.* Efficient Ensemble Refinement by Reweighting. *J Chem Theory Comput* **15**, 3390–3401 (2019).

[19] Mittag, T. *et al.* Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc Natl Acad Sci U S A* **105**, 17772–17777 (2008).

[20] Różycki, B., Kim, Y. C. & Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **19**, 109–116 (2011).

[21] Lazar, T. *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res* **49**, D404–D411 (2021).

[22] Borg, M. *et al.* Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc Natl Acad Sci U S A* **104**, 9650–9655 (2007).

[23] Bernadó, P. *et al.* A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* **102**, 17002–17007 (2005).

[24] Ozenne, V. *et al.* Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **28**, 1463–1470 (2012).

[25] Bondarenko, V. *et al.* Structures of highly flexible intracellular domain of human α7 nicotinic acetylcholine receptor. *Nat Commun* **13**, 793 (2022).

[26] Song, J., Li, J. & Chan, H. S. Small-Angle X-ray Scattering Signatures of Conformational Heterogeneity and Homogeneity of Disordered Protein Ensembles. *J Phys Chem B* **125**, 6451–6478 (2021).

[27] Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res* **36**, D402–408 (2008).

[28] Lavor, C., Liberti, L., Maculan, N. & Mucherino, A. The Discretizable Molecular Distance Geometry Problem. *Computational Optimization and Applications* **52**, 115–146 (2012).

[29] Liberti, L., Lavor, C. & Mucherino, A. The discretizable molecular distance geometry problem seems easier on proteins. *Distance Geometry: Theory, Methods and Applications. Mucherino, Lavor, Liberti, Maculan (eds.)* 47–60 (2014).

[30] Liberti, L., Lavor, C., Maculan, N. & Mucherino, A. Euclidean Distance Geometry and Applications. *SIAM Rev* **56**, 3–69 (2014).

[31] Lavor, C., Alves, R., Figueiredo, W., Petraglia, A. & Maculan, N. Clifford Algebra and the Discretizable Molecular Distance Geometry Problem. *Adv. Appl. Clifford Algebras* **25**, 925–942 (2015).

[32] Levinthal, C. Are there pathways for protein folding? *J Chem Phys* **65**, 44–45 (1968).

[33] Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. & Baker, D. Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One* **6**, e23294 (2011).

[34] Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

[35] Richard J. Gowers *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Conference*, 98–105 (2016).

[36] Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020). URL `https://doi.org/10.1038/s41586-020-2649-2`.

[37] Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol Cybern* **43**, 59–69 (1982).

[38] Kohonen, T. Self-organizing maps. *Springer Series in Information Sciences, Heidelberg, Germany.* (2001).

[39] Miri, L. *et al.* Stabilization of the integrase-DNA complex by Mg^{2+} ions and prediction of key residues for binding HIV-1 integrase inhibitors. *Proteins* **82**, 466–478 (2014).

[40] Bouvier, G. *et al.* Functional motions modulating VanA ligand binding unraveled by self-organizing maps. *J Chem Inf Model* **54**, 289–301 (2014).

[41] Lehmann, E. L. & Casella, G. *Theory of point estimation.* Springer Texts in Statistics (Springer-Verlag, New York, NY, 1998), 2nd edn.

[42] Lopéz-Blanco, J. R., Garzón, J. I. & Chacón, P. iMod: multipurpose normal mode analysis in internal coordinates. *Bioinformatics* **27**, 2843–2850 (2011).

[43] Wako, H. & Endo, S. Normal mode analysis based on an elastic network model for biomolecules in the Protein Data Bank, which uses dihedral angles as independent variables. *Comput Biol Chem* **44**, 22–30 (2013).

[44] Na, H. & Song, G. Bridging between normal mode analysis and elastic network models. *Proteins* **82**, 2157–2168 (2014).

[45] Tirion, M. M. & ben Avraham, D. Atomic torsional modal analysis for high-resolution proteins. *Phys Rev E Stat Nonlin Soft Matter Phys* **91**, 032712 (2015).

[46] Singh, H., Hnizdo, V. & Demchuk, E. Probabilistic model for two dependent circular variables. *Biometrika* **89**, 719–723 (2002).

[47] Mardia, K. V., Hughes, G., Taylor, C. C. & Singh, H. A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* **36**, 99–109 (2008).

[48] Boomsma, W. *et al.* A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci U S A* **105**, 8932–8937 (2008).

[49] Clenshaw, C. Chebyshev series for mathematical functions. *NPL Mathematical Tables* **5** (1962).

[50] McLachlan, G. J. & Krishnan, T. *The EM Algorithm and Extensions.* Wiley series in probability and statistics (John Wiley and Sons, Inc., 1997).

[51] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[52] Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208 (1995).

[53] Svergun, D. I., Barberato, C. & Koch, M. CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Cryst.* **28**, 768–773 (1995).

[54] Manalastas-Cantos, K. *et al.* ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J Appl Crystallogr* **54**, 343–355 (2021).

[55] Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

[56] Galassi, M. *GNU Scientific Library Reference Manual (3rd Ed.)* (Network Theory Ltd., 2009).

Figures

Figure 1

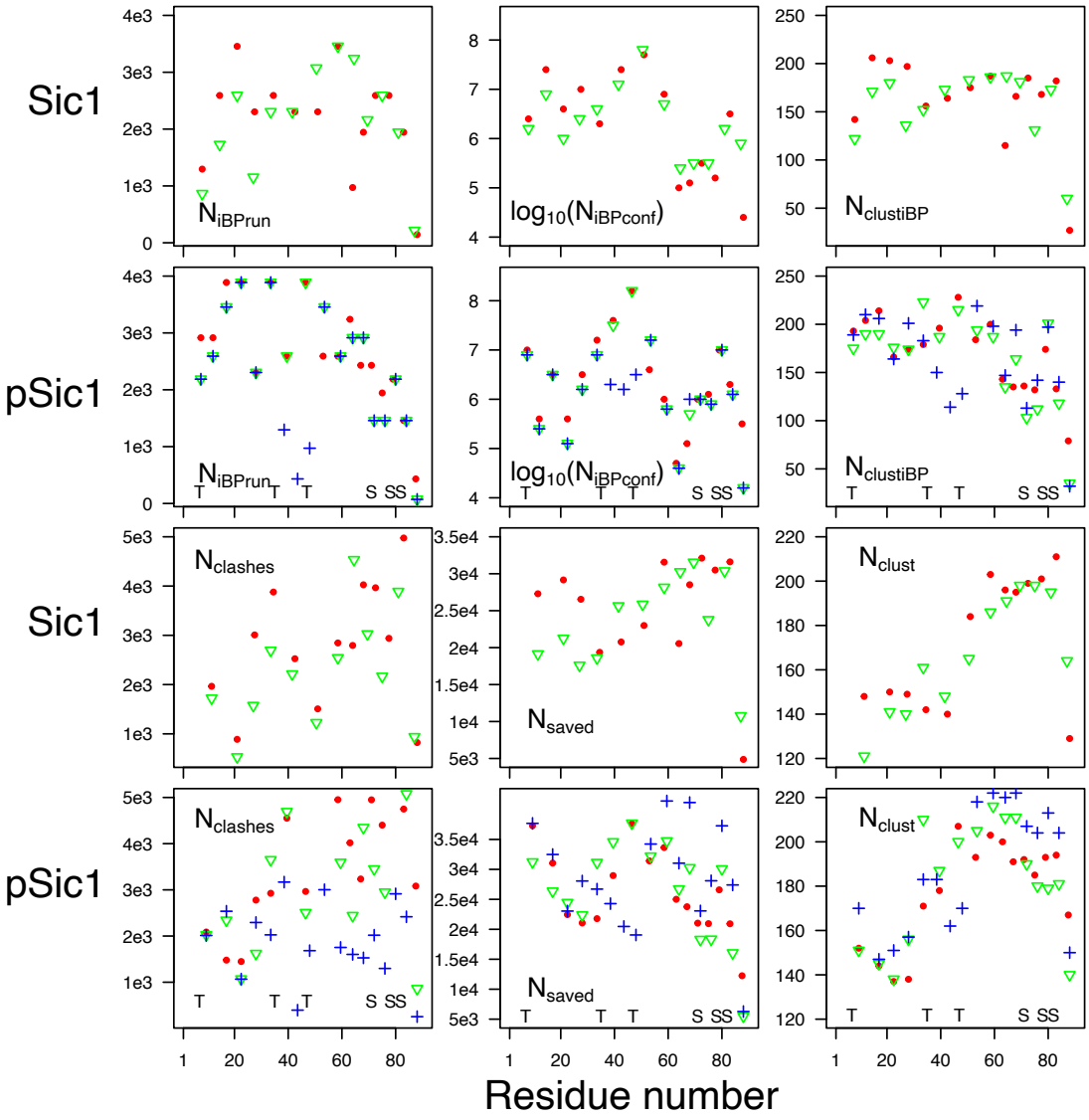


Figure 1: Parameters of the iBP and assembly steps of the TAiBP procedure. The signs red and green correspond respectively to the duplicated runs in which thresholds of 0.01 and 0.011 have been applied on the probability Ramachandran map. The blue crosses correspond to the run pSic1[3] producing more extended conformations. The positions of phosphorylated Threonines and Serines are marked with T and S for the runs on pSic1. The parameters are plotted along the number of the residue located at the middle of the fragment (iBP step) or at the middle of the last attached fragment (assembly step).

741 Figure 2

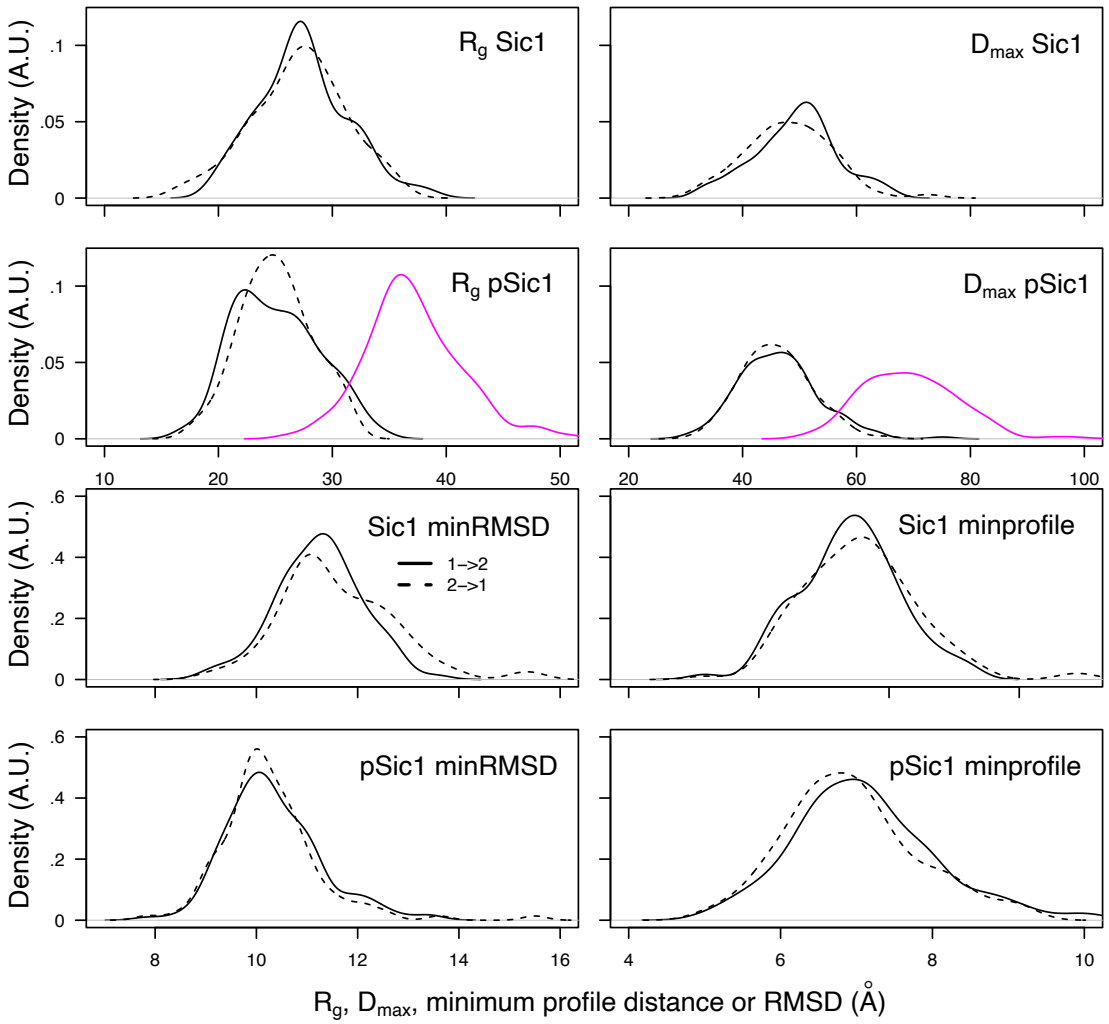


Figure 2: Four panels on top: Distribution of the gyration radii R_g and of maximal diameters D_{max} values in the two sets of TAiBP obtained during the first runs Sic1^1 and pSic1^1 (solid line) and the second runs Sic1^2 and pSic1^2 (dashed line) runs. The R_g and D_{max} distribution obtained for the run pSic1^3 are plotted in magenta. Four panels on bottom: Distribution of the minimum RMSD values (Å) and of the minimum distances (Å) between profiles for the duplicate runs performed for Sic1^1 and Sic^2 and for pSic1^1 and pSic1^2. full line: first run with respect to the second one, dashed line: second run with respect to the first one.

742 Figure 3

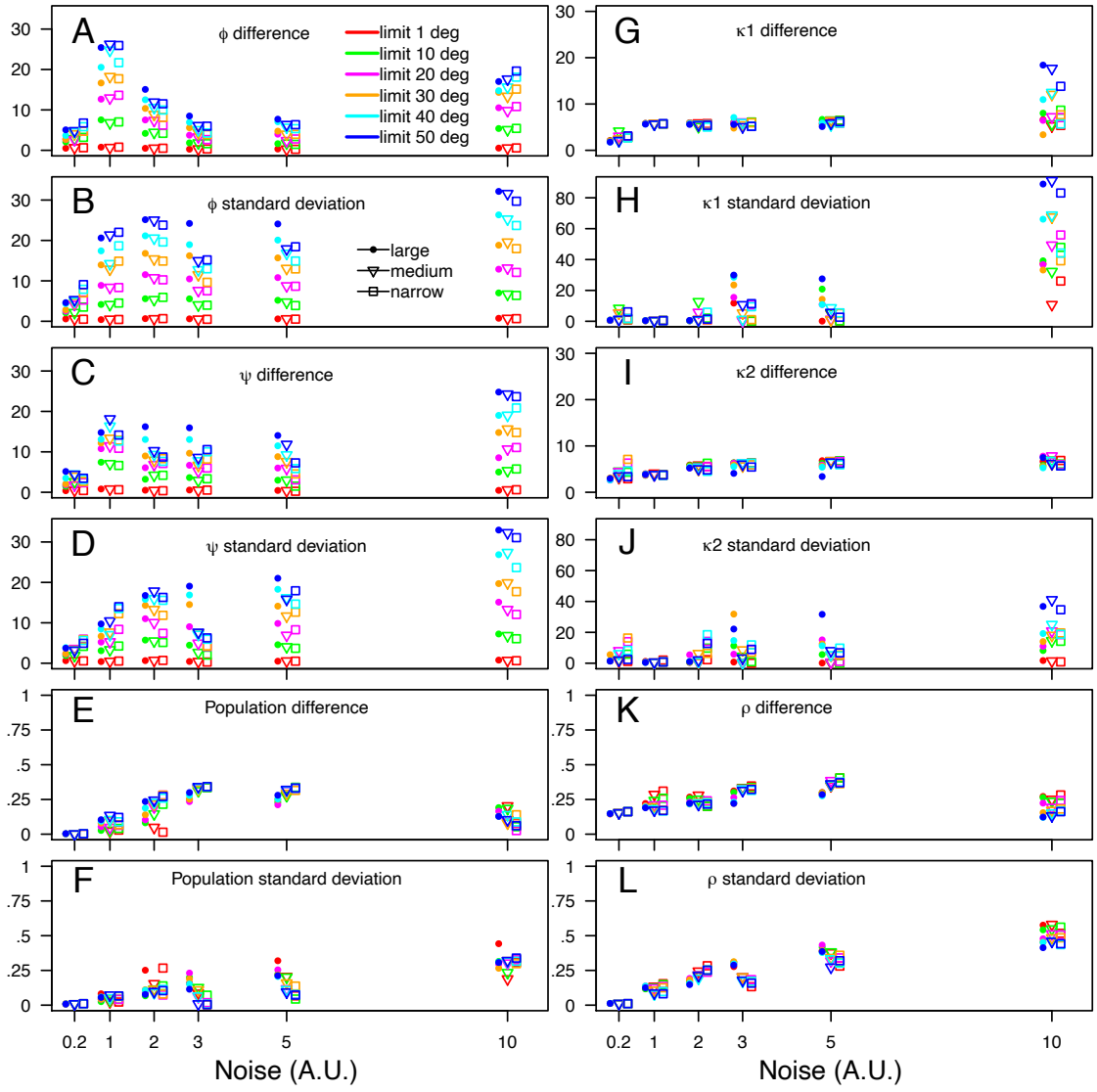


Figure 3: Efficiency of RamaMix for determining the ϕ_0, ψ_0 positions (A-D: Eq. 7), the von Mises shape parameters $\kappa 1$, $\kappa 2$ and ρ (G-L: Eq. 8), and the populations γ_q (E-F: Eq. 2) using synthetic data and various noise levels described in Figure S5. The results obtained for large, medium and narrow scattered synthetic Ramachandran maps are drawn as bullets, triangles and squares.

743 Figure 4

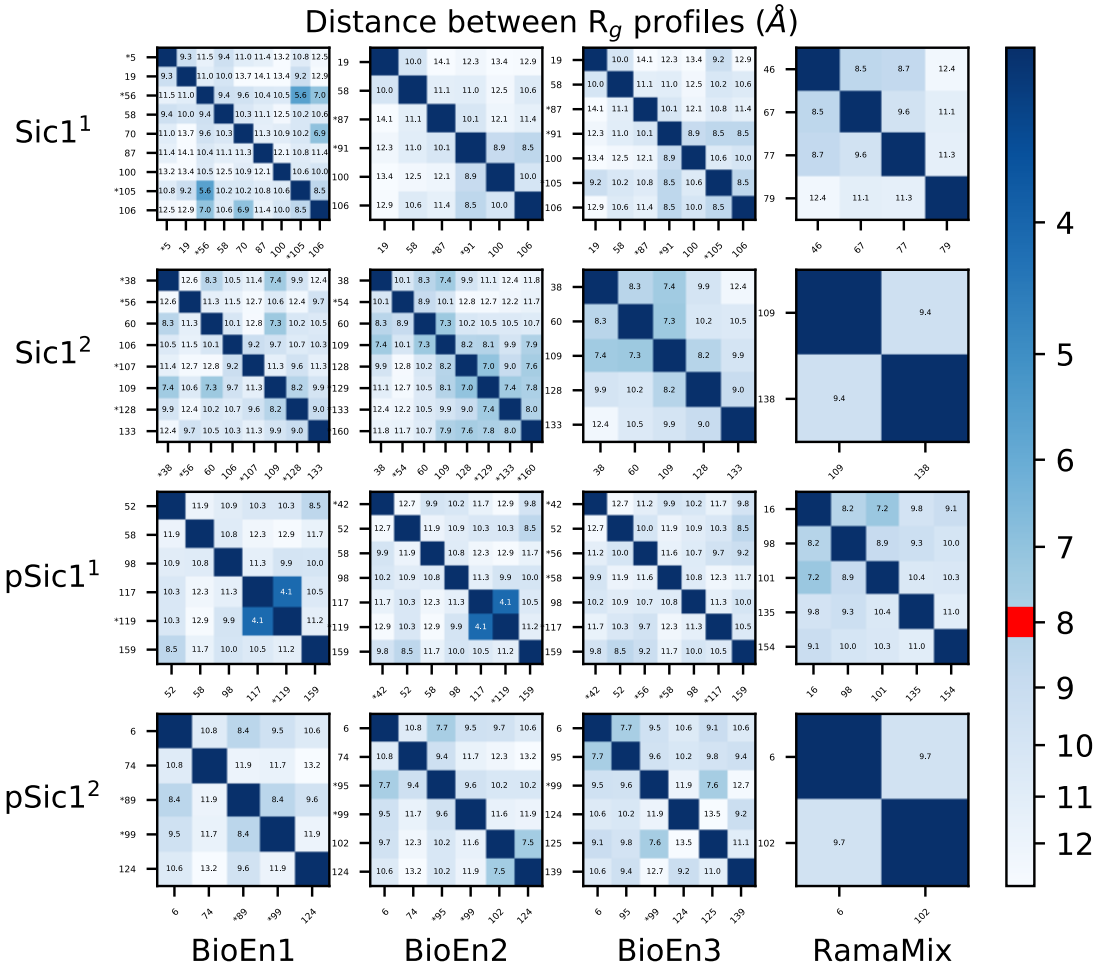


Figure 4: Distances between the profiles P_q (Eq. 1) of local gyration radii between the conformations selected from the fit of SAXS curves (BioEn1,BioEn2, BioEn3) or Ramachandran maps (RamaMix). The conformations for which populations smaller than 10% were calculated, are labeled with an asterisk. The diagonals correspond to the comparison of the same conformations and are thus not annotated with distance value. The limit of 8 Å used to display superimposed plots of profiles P_q (Figure 5) is drawn in red on the scale of distance.

744 Figure 5

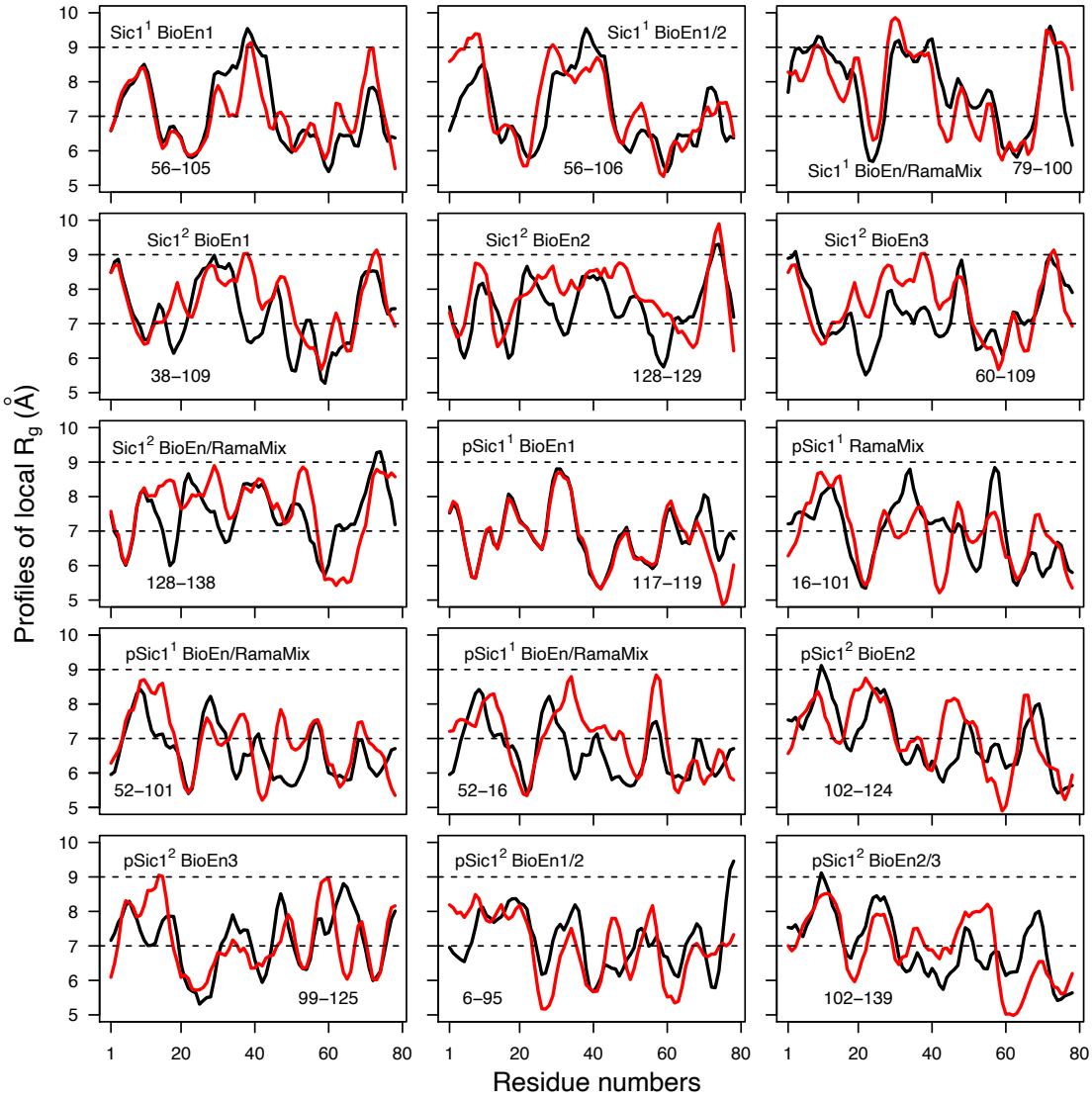


Figure 5: Superposition of profiles P_q (Eq. 1) displaying distances smaller than 8 Å extracted from Figures 4 and S8, S9. The name of the run ($Sic1^1$, $Sic1^2$, $pSic1^1$, $pSic1^2$) is given, along with the name of the considered fits (BioEn1, BioEn2, BioEn3, RamaMix, RamaMix/BioEn) and the conformations numbers. The labels RamaMix/BioEn correspond to the comparison of conformations selected by BioEn on one side and RamaMix on the other side. The labels BioEn1/2 and BioEn2/3 correspond to the comparison of conformations selected by BioEn from two different SAXS curves.

745 Figure 6

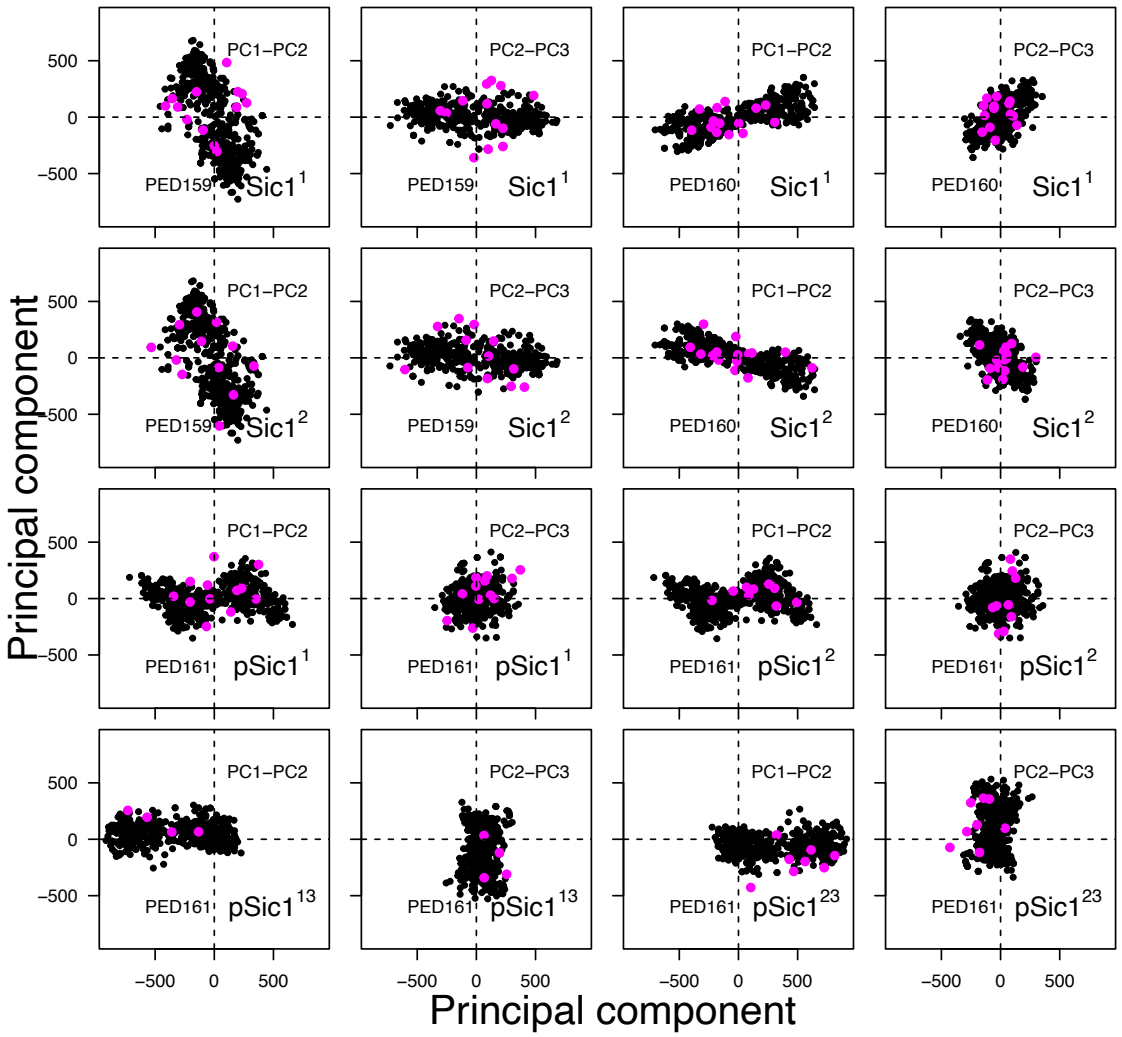


Figure 6: Projections of the Sic1 and pSic1 conformations along the three largest components of their principal component analysis (PCA). On these projections, the TAiBP conformations selected by BioEn or RamaMix are colored in magenta and the conformations stored in PED[21] are colored in black.

A. Sic1[1]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	100	15.5 ± 2.2	**100**	19.8 ± 0.2	**100**	21.7 ± 0.1
	105	1.4 ± 2.7	**106**	15.0 ± 1.6	**105**	5.0 ± 0.1
	106	13.0 ± 1.7	**19**	25.6 ± 0.2	**106**	18.2 ± 0.6
	19	18.3 ± 2.2	**58**	24.4 ± 0.4	**19**	21.8 ± 0.1
	56	6.7 ± 3.4	**87**	9.5 ± 0.3	**58**	23.9 ± 0.3
	58	4.4 ± 2.2	**91**	5.5 ± 1.8	**87**	7.0 ± 0.0
	5	6.4 ± 2.6			**91**	2.3 ± 0.8
	70	14.6 ± 5.6				
	87	19.1 ± 2.8				
Average final $\chi 2$		0.4		0.4		0.3
Average final S_{KL}		-1.7e-9		-7.9e-10		-5.0e-10
B. Sic1[2]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	106	17.2 ± 1.0	**38**	13.7 ± 4.6	**109**	43.6 ± 1.3
	107	7.1 ± 0.3	**109**	41.7 ± 7.6	**128**	15.2 ± 1.5
	109	27.1 ± 0.9	**128**	12.5 ± 1.4	**133**	10.4 ± 3.5
	128	6.6 ± 0.5	129	9.1 ± 3.0	**38**	19.9 ± 0.7
	133	17.3 ± 1.2	**133**	1.2 ± 3.5	**60**	10.4 ± 0.2
	38	6.9 ± 0.3	160	1.3 ± 3.3		
	56	5.7 ± 0.3	54	3.0 ± 4.9		
	60	11.7 ± 0.3	**60**	16.1 ± 1.1		
Average final $\chi 2$		0.4		0.4		0.3
Average final S_{KL}		-1.0e-8		-1.7e-8		-3.1e-9

Table 1: Conformations and populations selected using BioEn 0.1.1[18] on the three sets of SAXS curves. The conformations were generated by the runs Sic1[1] and Sic1[2]. For each SAXS curve and set of protein conformations, after ten runs starting from random values of populations and performed on the whole set of conformations, all conformations for which the sum of populations over the ten runs was larger than 0.01 were gathered, and a second run of ten additional BioEn calculations was performed on this reduced set of conformations. The average and standard deviation values of populations obtained for each selected conformation from the second set of BioEn runs, are given in the Table, along with the final average values of reduced χ^2 and of entropy S_{KL}. The labels of conformations selected in at least two runs are written in bold. The conformations displaying average populations smaller than 1% were removed from the final set.

A. pSic1[1]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	117	16.1 ± 0.6	**117**	21.2 ± 0.6	**117**	9.1 ± 4.6
	119	2.7 ± 0.9	**119**	2.1 ± 1.3	**159**	35.9 ± 3.6
	159	32.1 ± 0.1	**159**	22.2 ± 0.6	**42**	7.3 ± 7.6
	52	18.9 ± 0.4	**42**	4.4 ± 2.5	**52**	11.0 ± 4.3
	58	16.9 ± 0.5	**52**	10.9 ± 0.9	56	2.2 ± 4.3
	98	13.2 ± 0.3	**58**	23.0 ± 0.9	**58**	8.1 ± 4.1
			98	16.2 ± 1.6	**98**	26.4 ± 3.4
Average final $\chi 2$		0.9		1.1		0.7
Average final S_{KL}		-1.9e-9		-2.5e-9	-2.3e-10	
B. pSic1[2]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	124	42.3 ± 0.6	102	13.9 ± 7.4	**124**	24.8 ± 1.6
	6	39.7 ± 0.4	**124**	37.1 ± 8.4	125	10.2 ± 3.4
	74	13.4 ± 0.2	**6**	30.3 ± 4.3	139	30.5 ± 1.3
	89	2.0 ± 0.3	**74**	13.8 ± 0.7	**6**	13.1 ± 0.7
	99	2.1 ± 0.7	**95**	1.3 ± 3.6	**95**	20.2 ± 1.2
			99	3.6 ± 2.5	**99**	1.2 ± 1.8
Average final $\chi 2$		0.8	0.9		0.7	
Average final S_{KL}		-3.8e-9		-1.8e-8		-3.8e-10

Table 2: Conformations and populations selected using BioEn 0.1.1[18] on the three sets of SAXS curves. The conformations were generated by the runs pSic1[1] and pSic1[2]. The Table caption is the same than for Table 1

A. Sic1^1	conformation numbers	populations percentages
	79	44.7 ± 0.5
	77	23.4 ± 0.6
	67	21.7 ± 0.4
	46	10.2 ± 0.4
B. Sic1^2	conformation numbers	populations percentages
	109	67.8 ± 2.9
	138	32.1 ± 0.8
C. pSic1^1	conformation numbers	populations percentages
	98	23.2 ± 1.4
	154	22.7 ± 2.0
	101	21.2 ± 0.7
	135	19.2 ± 3.3
	16	13.7 ± 1.0
D. pSic1^2	conformation numbers	populations percentages
	6	59.2 ± 3.7
	102	40.7 ± 3.0

Table 3: Conformations and populations selected by fitting of the Ramachandran maps using RamaMix. For each set of protein conformations, 100 runs were performed starting from random values for the populations. The few converged optimizations which did not converge, were discarded: 6 for Sic1^1, 2 for Sic1^2, 3 for pSic1^1 and 3 for pSic1^2. The backbone angles ϕ and ψ were allowed to move up to 15°. The populations of conformations for the converged runs were averaged and these mean values are given as percentages in the Table along with the corresponding standard deviation values. The labels of conformations also selected by BioEn are written in bold.

Data-set	BioEn1	BioEn2	BioEn3	RamaMix
$Sic1^1$	27.8	28.7	28.5	31.3
$Sic1^2$	27.7	28.4	28.4	27.1
$pSic1^1$	26.7	26.1	27.2	28.0
$pSic1^2$	27.4	27.1	27.9	30.0
$pSic1^3$	30.4	30.6	30.5	32.4
$pSic1^{13}$	27.4	27.2	28.0	32.5
$pSic1^{23}$	27.4	27.4	28.1	32.5

Table 4: Resulting gyration radii (Å) calculated from the individual gyration radii of the conformations selected by the BioEn and RamaMix analyses. The data-sets $Sic1^1$, $Sic1^2$ and $pSic1^1$, $pSic1^2$, $pSic1^3$ were obtained using the approach TAiBP on the proteins Sic1 and pSic1. The data-sets $pSic1^{13}$ and $pSic1^{23}$ were obtained by pooling together the conformations of $pSic1^3$ and $pSic1^1$ or the conformations of $pSic1^3$ and $pSic1^2$.

Supplementary information: Low-resolution description of the conformational space for intrinsically disordered proteins

Daniel Förster (1), Jérôme Idier (2), Leo Liberti (3), Antonio Mucherino (4), Jung-Hsin Lin (5) and Thérèse E. Malliavin (6,7,8)

(1) UMR7374 Interfaces, Confinement, Matériaux et Nanostructures, Université d'Orléans, France

(2) UMR6004 Laboratoire des Sciences du Numérique de Nantes, France

(3) LIX UMR 7161 CNRS École Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France

(4) IRISA, University of Rennes 1, France

(5) Biomedical Translation Research Center, Academia Sinica, Taiwan

(6) Institut Pasteur, Université Paris Cité, CNRS UMR3528, Unité de Bioinformatique Structurale, F-75015 Paris, France

(7) Laboratoire de Physique et Chimie Théoriques (LPCT), University of Lorraine, Vandoeuvre-lès-Nancy, France

(8) Laboratoire International Associé, CNRS and University of Illinois at Urbana-Champaign, Vandoeuvre-lès-Nancy, France

Corresponding authors:

Thérèse E. Malliavin, therese.malliavin@univ-lorraine.fr

Jérôme Idier, jerome.idier@ls2n.fr

Short title

Conformational space of IDPs Sic1 and pSic1

August 29, 2022

Extraction of boxes from Ramachandran likelihood

The likelihood Ramachandran maps, calculated by TALOS-N [1], were first normalized in order to get the sum of values equal to 1 and to produce probability maps. Each ϕ, ψ box was determined from these maps in the following way. In the current state of the Ramachandran map, the pixels belonging to a box are removed from the map. From the remaining pixels, the pixel of maximum probability value and larger than the threshold, is selected and a box is iteratively drawn around this position by testing systematically all pixels neighboring the current box limits. All neighbouring pixels containing values larger than a given threshold are included in the box. If values are smaller than the threshold, the calculation stops, the current box definition is kept for further analyses and the pixels selected from the box are removed from the Ramachandran map. This approach is iteratively applied to the map up to the situation where all remaining map pixels display values smaller than the threshold. In order to probe the reproducibility of TAiBP results, two sets of boxes have been determined with threshold values of 0.01 (Figures S1 and S3) and 0.011 (Figures S2 and S4).

In pSic1, the presence of phosphorylated Threonines 7, 35 and 47 and of phosphorylated Serines 71, 78 and 82 makes impossible the TALOS-N predictions for residues 5-9, 33-37,

45-49, 69-73, 76-84, due to the lack of phosphorylated proteins in the learning set of the neural network. Thus, for these residues, generic boxes (Table S2) have been used as input of TAiBP, in order to cover the Ramachandran regions corresponding to α-helix, extended, β strand and loop structures.

Enumeration of conformations

The enumeration of protein conformations was performed using boxes of backbone angles ϕ and ψ. These boxes (Figures S1-S4) have been extracted from the likelihood Ramachandran maps obtained by TALOS-N [1] as described in the previous section. During the tree building, each atomic position is determined by trilateration from the previously determined atomic positions, following a specific ordering (Table S4) [2]. More precisely, two out of three of the distances involved in trilateration must be known exactly, and one may be subject to uncertainty and represented by an interval [2, 3]. The iBP algorithm was the one described by Worley et al [4, 3].

The backbone dihedral angles ϕ and ψ can be straightforwardly related to bond lengths and bond angles and respectively to distances between atoms C of residues $i-1$ and i and between atoms N of residues i and $i+1$. This equivalence between the backbone dihedral angles and inter-atomic distances permits to use the angles ϕ and ψ for the so-called branching step. This branching step is performed by discretization of the intervals in order to generate new branches in the tree.

The bond lengths, bond angles, improper angles and van der Waals radii were taken from the force field protein-allhdg5-4 PARALLHDG (version 5.3) [5, 6]. The van der Waals radii

were scaled by a factor of 0.7.

For each fragment, two dummy residues were added at the N and C terminal extremities, and the ϕ and ψ dihedral angles of the inner peptide residues were sampled according to the box limits (Table S1). In order to avoid pruning due to slight discrepancy between distances, a tolerance of 0.05 Å has been added to the bounds of distance intervals. The maximum number of branches by discretized interval was set to 4. The minimum discretization factor, which is the minimum ratio between each distance interval to the number of tree branches generated within the interval, was set to 0.1 Å, in order to avoid that the branching over-samples small intervals. A maximum number of 10^9 saved conformations was permitted for each iBP run. The solutions were stored in a multiframe dcd format [7].

Clustering of generated conformations

The approach **S**elf-**O**rganizing **M**ap (SOM) [8, 9, 10, 11], used to cluster conformations, is an artificial neural network (ANN) trained using unsupervised learning. SOM displays the advantage with respect to the k-means clustering approach that it does not require the predetermined knowledge of the number of clusters. The SOM approach was used after each iBP calculation or assembly step as soon as the number of saved conformations was larger than 100. The conformations are encoded from the distances d_{ij} calculated between the n C_α atoms by diagonalizing the covariance matrix C:

$$C_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{n}(d_{i,k} - \bar{d}_i)(d_{l,j} - \bar{d}_j) \quad (1)$$

where $\bar{d}_s = \frac{1}{n}\sum_{p=1}^{n} d_{s,p}$. The information contained in the matrix C is equivalent to its four largest eigenvalues along with the corresponding eigenvectors, and is formatted as an

input vector of length $4(n+1)$. These vectors are used to train a periodic Euclidean 2D self-organizing map (SOM), which corresponds to a three-dimensional matrix. The first two matrix dimensions were chosen to be 100×100 and define the map size, the third dimension being equal to $4(n+1)$. Each vector along the third dimension defines a neuron of the map. The neurons of the self-organizing map are initialized with a random uniform distribution covering the range of values of the input vectors. At each step, an input vector is presented to the map, and the neurons closest to this input are updated. The training parameters were those previously described [12, 11].

Once the SOM has been determined, representative conformations are extracted from the conventional **U**nified **d**istance **m**atrix (U-matrix) calculated from the final SOM neurons. For each neuron ν, the corresponding U-matrix element is calculated as the average Euclidean distance between the neuron ν and its eight immediate neighbors:

$$\text{U-matrix}(\nu) = \frac{1}{8} \sum_{\nu \in N(\nu)} d(\nu, \mu) \tag{2}$$

where $N(\nu)$ is the set of neighbors, and $d(\nu, \mu)$ is the Euclidean distance between the neurons μ and ν. pSic1[3] The neurons corresponding to local minima of the U-matrix, and thus to local maxima of conformational homogeneity, are extracted and for all performed runs except pSic1[3], the protein conformation displaying the closest distance to this neuron is saved. In the case of pSic1[3], among the conformations saved in the neurons, the one displaying the longest distance between the Carbons α of the first and the last residues is saved, in order to obtain more extended conformations in agreement with the values of gyration radii measured in Ref. [13]. The conformations generated during the iBP or assembly steps are finally replaced by the sets of representative conformations extracted from local minima of U-matrix.

Molecular dynamics refinement in implicit solvent

Molecular dynamics (MD) trajectories were used to relax the Sic1 and pSic1 conformations obtained from the TAiBP approach. The MD trajectories were recorded using NAMD 2.13 [14]. Topology parameters were taken from the force fields c36 [15] and c36m [16]. The simulations were performed at a temperature of 300 K. A Generalized Born implicit solvent (GBIS) [17] model was used with an ion concentration of 0.3M, and a cutoff of 12 Å for calculating Born radius. A cutoff of 14 Å and a switching distance of 13 Å were defined for non-bonded interactions. The RATTLE algorithm [18, 19] was used to keep all covalent bonds involving hydrogens rigid, enabling a time step of 2 fs. Temperature was regulated according to a Langevin thermostat [20]. At the beginning of each trajectory, the system was first minimized for 1,000 steps, then heated up gradually from 0 K to 300 K in 30,000 integration steps. Finally, the system was equilibrated for 5,000 steps. During all steps, from minimization to production, positional restraints were applied on protein backbone atoms with a constant force of 1 kcal/mol. A production run of 100ps was then performed and the conformation of the final frame was saved as the relaxed conformation.

Determination of the populations from the Ramachandran maps

Using the neural network TALOS-N [1], it is possible, starting from the NMR chemical shifts measured on protein atoms, to determine for each residue n a 2D probability density $p^n_{\text{TALOS-N}}(\phi, \psi)$. As NMR analyses a sample containing a mixture of conformations, we

propose to decompose the probability map produced by TALOS-N as a mixture of probability densities $p_q^n(\phi, \psi)$, corresponding to a certain number of free energy basins q present in the experimental sample:

$$p^n_{\text{TALOS-N}}(\phi, \psi) \approx \sum_{q=1}^{Q} \gamma_q p_q^n(\phi, \psi) \tag{3}$$

where $\gamma_q \geq 0$ is the proportion of local basins q in the NMR sample. Thus:

$$\sum_{q=1}^{Q} \gamma_q = 1. \tag{4}$$

In the following, the problem described by Eq. 3 will be named as a Q-class mixture problem, each class corresponding to a conformation of the studied protein. In addition, for each conformation q, the couple of angles corresponding to the bottom of the basin will be named its location parameters.

To fit the set of N available TALOS-N probability densities using the mixture model (3), we have to rely on a discrepancy measure between both probability maps. Kullback-Leibler divergence is a standard choice:

$$D_{\text{KL}}(p_1||p_2) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx. \tag{5}$$

Here, we consider the sum of discrepancy measures over the N residues between TALOS-N probability densities and the corresponding mixture model densities $p^n = \sum_{q=1}^{Q} \gamma_q p_q^n$:

$$\sum_{n=1}^{N} D_{\text{KL}}(p^n_{\text{TALOS-N}} \,||\, p^n) = \sum_{n=1}^{N} \int p^n_{\text{TALOS-N}}(\phi, \psi) \ln \frac{p^n_{\text{TALOS-N}}(\phi, \psi)}{p^n(\phi, \psi)} d\phi d\psi. \tag{6}$$

In practice, TALOS-N probability densities are available on a finite rectangular grid $\{\phi_1, \ldots, \phi_I\} \times \{\psi_1, \ldots, \psi_J\}$. Let us modify (6) using a zeroth-order approximation of the integrals:

$$\sum_{n=1}^{N} D_{\text{KL}}(p^n_{\text{TALOS-N}} \,||\, p_2) = \sum_{n=1}^{N} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{p}^n_{ij} \ln \frac{\hat{p}^n_{ij}}{p^n_{ij}}. \tag{7}$$

where $\hat{p}^n_{ij} = p^n_{\text{TALOS-N}}(\phi_i, \psi_j)$ and

$$p^n_{ij} = p^n(\phi_i, \psi_j) = \sum_{q=1}^{Q} \gamma_q p^n_q(\phi_i, \psi_j). \tag{8}$$

Finally, the minimization of Eq. 7 with respect to $\gamma = (\gamma_1, \ldots, \gamma_Q)$ amounts to the maximization of

$$f(\gamma) = \sum_{n=1}^{N} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{p}^n_{ij} \ln p^n_{ij}, \tag{9}$$

$p^n_{ij}(\phi, \psi)$ being given by (8), under the constraints $\gamma_q \geq 0$ and $\sum_{q=1}^{Q} \gamma_q = 1$. Let us remark that f is a concave function of γ, so that its local maximization with respect to γ cannot be trapped in a local maximum.

In case TALOS-N yielded observations in the form of angle couples $y^n_m = (\phi^n_m, \psi^n_m)$, $m = 1, \ldots, M$, instead of a probability map, we would naturally maximize the log-likelihood of the data [21],

$$\mathcal{L}(y; \theta) = \sum_{n=1}^{N} \sum_{m=1}^{M} L^n(\phi^n_m, \psi^n_m), \tag{10}$$

a well-known local optimization scheme to reach this goal being the EM algorithm [22, 23]. Let us remark the similarity between Eqs. (9) and (10). In fact, Eq. (9) identifies with the log-likelihood of virtual data, each couple (ϕ_i, ψ_j) being observed $D^n_{ij} = \hat{p}^n_{ij} \times C$ times for the nth residual (C being an arbitrary constant). This corresponds to a well-known correspondance between the log-likelihood and the Kullback-Leibler divergence between the empirical data distribution and the parametrized one (see for instance [23]).

In the case where data points correspond to couples of angles, we must consider that the support of densities p_q is a torus, i.e., that they are doubly circular. The most natural circular extension of the univariate Gaussian is the wrapped normal distribution. However, the von Mises distribution is usually considered as a better option, being more easily tractable

[24]. Moreover, multivariate extensions exist for the latter. In particular, in the Ref. [25] a bivariate version was introduced, motivated by problems of modelling torsional angles in molecules, and a pseudo-maximum likelihood method was proposed [24] to estimate its parameters. Moreover, a so-called *cosine* version was investigated [26] and an Expectation-Maximization (EM) algorithm was used [26, 27] to solve a problem that is almost identical to ours.

Here, we adopt the same bivariate periodic sine model as [25]:

$$p(\phi, \psi) = \frac{1}{T} \exp(W(\phi - \phi_0, \psi - \psi_0)) \tag{11}$$

with

$$W(\phi, \psi) = \kappa_1 \cos\phi + \kappa_2 \cos\psi + \lambda \sin\phi \sin\psi \tag{12}$$

and $\kappa_1, \kappa_2 \geq 0$ and $\lambda^2 < \kappa_1 \kappa_2$. A difficulty is that the integration constant is expressed as an infinite series, depending on parameters $(\kappa_1, \kappa_2, \lambda)$:

$$T = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1 \kappa_2} \right)^m I_m(\kappa_1) I_m(\kappa_2). \tag{13}$$

In Ref. [25], expressions of κ_1, κ_2, λ are given as functions of the parameters $(\sigma_1^2, \sigma_2^2, \rho)$ of a bivariate Gaussian:

$$\sigma_1^2 = \frac{\kappa_2}{\kappa_1 \kappa_2 - \lambda^2}, \quad \sigma_2^2 = \frac{\kappa_1}{\kappa_1 \kappa_2 - \lambda^2}, \quad \rho = \frac{\lambda}{\sqrt{\kappa_1 \kappa_2}}. \tag{14}$$

where $\rho \in (-1, 1)$ denotes the normalized correlation coefficient between the two components of the bivariate Gaussian. These expressions are valid only in the case where σ_1^2 and σ_2^2 are small. They are easily inverted as

$$\kappa_1 = \frac{1}{\sigma_1^2} \frac{1}{1 - \rho^2}, \quad \kappa_2 = \frac{1}{\sigma_2^2} \frac{1}{1 - \rho^2}, \quad \lambda = \frac{1}{\sigma_1 \sigma_2} \frac{\rho}{1 - \rho^2}. \tag{15}$$

Using (15), we can replace a Gaussian mode p_q^n by a periodized version, with approximately the same location and the same spread. This is not specific to the Gaussian case, so it also holds for the bivariate von Mises-type model.

In the following section "Maximum likelihood estimation for bivariate sine mixtures", we are deriving the equations describing an original approach for solving the problem (3) by a maximum likelihood approach.

Maximum likelihood estimation for bivariate sine mixtures

Let $Y = (y_1, \ldots, y_D)$ stand for D iid datapoints. We make the assumption that each y_d is sampled from a Q-class mixture model, and we use the notation C_d to refer to the random class attached to y_d, taking values in $(1, \ldots, Q)$. For each d, we have

$$p(y_d; \theta) = \sum_{q=1}^{Q} \Pr(C_d = c_q)\, p(y_d | C_d = c_q; \zeta) = \sum_{q=1}^{Q} \gamma_q\, p(y_d; \zeta_q^{\mathrm{L}}, \zeta_q^{\mathrm{S}}) \tag{16}$$

with unknown parameters $\theta = (\gamma, \zeta) = (\gamma, \zeta^{\mathrm{L}}, \zeta^{\mathrm{S}})$, including

- normalized weights $\gamma = (\gamma_q)$,
- location parameters $\zeta^{\mathrm{L}} = (\zeta_q^{\mathrm{L}})$ where $\zeta_q^{\mathrm{L}} = (\phi_q, \psi_q)$ is specific to class q,
- shape parameters $\zeta^{\mathrm{S}} = (\zeta_q^{\mathrm{S}})$ where $\zeta_q^{\mathrm{S}} = (\kappa_{1q}, \kappa_{2q}, \lambda_q)$ is specific to class q.

We would like to estimate θ according to the maximum likelihood principle:

$$\hat{\theta} = \arg\max_{\theta} p(Y; \theta).$$

where $p(Y;\theta) = \prod_{d=1}^{D} p(y_d;\theta)$. Equivalently, $\hat{\theta}$ maximizes the log-likelihood, which reads

$$L(Y;\theta) = -\sum_{d=1}^{D} \ln\left(\sum_{q=1}^{Q} \gamma_q \, p(y_d;\zeta_q^{\mathrm{L}},\zeta_q^{\mathrm{S}})\right).$$

In the following, we are first describing what should be an Expectation-Maximization (EM) algorithm adapted for solving the maximum likelihood problem, to finally remark that the Maximization step of the EM cannot be solved analytically. Thus, we turn to a solution based on a well-grounded gradient-based optimization scheme. We derive explicit expressions for the gradient terms, on the basis of the Expectation step of the EM. At the end of this section, we present how to include into the optimization scheme, several Ramachandran probability maps corresponding to several protein residues.

Expectation-Maximization (EM) algorithm

The EM algorithm is a reference solution to determine $\hat{\theta}$ by iterative local optimization. Each EM iteration consists in solving the following auxiliary problem:

$$\theta^{\mathrm{new}} = \arg\max_{\theta} Q(\theta;\theta^{\mathrm{old}}), \tag{17}$$

where Q is the expectation of the log-likelihood of the "complete" dataset:

$$Q(\theta,\theta^{\mathrm{old}}) = \mathrm{E}\left[\ln\left(\Pr(C;\gamma)\, p(Y|C;\zeta)\right)|Y;\theta^{\mathrm{old}}\right] \tag{18}$$

$$= Q_0(\gamma,\theta^{\mathrm{old}}) + Q_1(\zeta,\theta^{\mathrm{old}}) \tag{19}$$

with

$$Q_0(\gamma,\theta^{\mathrm{old}}) = \mathrm{E}\left[\ln\Pr(C;\gamma)|Y;\theta^{\mathrm{old}}\right], \tag{20}$$

$$Q_1(\zeta,\theta^{\mathrm{old}}) = \mathrm{E}\left[\ln p(Y|C;\zeta)|Y;\theta^{\mathrm{old}}\right]. \tag{21}$$

On the one hand, along classical derivations, we get

$$Q_0(\gamma, \theta^{\text{old}}) = \sum_{q=1}^{Q} \left(\sum_{d=1}^{D} P_{qd} \right) \ln \gamma_q \tag{22}$$

where $P_{qd} = P_q(y_d)$, with

$$P_q(y) = \Pr(C = q|y; \theta^{\text{old}}) = \frac{\gamma_q^{\text{old}} p_q(y; \zeta^{\text{old}})}{\sum_{q'} \gamma_{q'}^{\text{old}} p_{q'}(y; \zeta^{\text{old}})}. \tag{23}$$

On the other hand,

$$Q_1(\zeta, \theta^{\text{old}}) = \sum_{d=1}^{D} \sum_{q=1}^{Q} P_{qd} \ln p(y_d; \zeta_q^{\text{L}}, \zeta_q^{\text{S}}). \tag{24}$$

The optimization problem (17) splits in two parts at each iteration, according to

$$\gamma^{\text{new}} = \arg\max_{\gamma} Q_0(\gamma, \theta^{\text{old}}), \tag{25}$$

$$\zeta^{\text{new}} = \arg\max_{\zeta} Q_1(\zeta, \theta^{\text{old}}). \tag{26}$$

The first subproblem is constrained by $\sum_q \gamma_q = 1$. It has a simple, explicit solution. Unfortunately, the second subproblem cannot be solved analytically for the sine model, neither for the shape parameters ζ^{S}, nor for the location parameters ζ^{L}. As a consequence, exact closed-form EM formulas do not exist for the sine model. To our best knowledge, the same holds for other von Mises type models, such as the cosine version of [26]. Indeed, we guess that the EM algorithm used therein solves the maximization step in an approximate way. We rather propose a different solution, relying on a well-grounded gradient-based optimization scheme (namely, the L-BFGS-B algorithm [28]) applied to the log-likelihood itself.

Gradient-based log-likelihood maximization

Fisher's identity [29] relates the gradient of Q to the gradient of the log-likelihood L:

$$\left. \frac{\partial}{\partial \theta} Q(\theta; \theta^{\text{old}}) \right|_{\theta=\theta^{\text{old}}} = \left. \frac{\partial}{\partial \theta} L(Y; \theta) \right|_{\theta=\theta^{\text{old}}} \tag{27}$$

This property is very useful when the M step is not closed-form, since it allows one to replace non-explicit EM iterations by explicit gradient-based iterations, directly applicable to the log-likelihood.

Partial derivative w.r.t. the weights γ

Given Eqs (19), (22) and (27), we have

$$\frac{\partial}{\partial \gamma_q} L(Y;\theta)\bigg|_{\theta=\theta^{\text{old}}} = \frac{\partial}{\partial \gamma_q} Q_0(\gamma, \theta^{\text{old}})\bigg|_{\theta=\theta^{\text{old}}} = \left(\sum_{d=1}^{D} P_{qd}\right) \frac{1}{\gamma_q}. \tag{28}$$

Optimization w.r.t. the weights must be conducted under the constraints of nonnegativity and sum-to-one. The latter can be easily handled using the simple reparameterization $\gamma_q = \frac{\gamma'_q}{\sum_r \gamma'_r}$. It is easy to establish that

$$\frac{\partial}{\partial \gamma'_q} L(Y;\theta)\bigg|_{\theta=\theta^{\text{old}}} = \left(\sum_{d=1}^{D} P_{qd}\right) \frac{1}{\gamma'_q} - \frac{D}{\sum_{r=1}^{Q} \gamma'_r}.$$

Partial derivative w.r.t. the shape parameters ζ^{S}

Given Eqs (19), (24) and (27), we have

$$\frac{\partial}{\partial \zeta_q^{\text{S}}} L(Y;\theta)\bigg|_{\theta=\theta^{\text{old}}} = \frac{\partial}{\partial \zeta_q^{\text{S}}} Q_1(\theta;\theta^{\text{old}})\bigg|_{\theta=\theta^{\text{old}}} = \sum_{d=1}^{D} P_{qd} \frac{\partial}{\partial \zeta_q^{\text{S}}} \ln p(y_d; \zeta_q^{\text{L}}, \zeta_q^{\text{S}}) \tag{29}$$

where $p(y_d; \zeta_q^{\text{L}}, \zeta_q^{\text{S}})$ is a sine density defined by (11). Explicit expressions for the partial derivative depend on each shape parameter, according to

$$\frac{\partial}{\partial \kappa_1} \ln p(\phi, \psi) = \cos(\phi - \phi_0) - \frac{1}{T} \frac{\partial T}{\partial \kappa_1} \tag{30}$$

$$\frac{\partial}{\partial \kappa_2} \ln p(\phi, \psi) = \cos(\psi - \psi_0) - \frac{1}{T} \frac{\partial T}{\partial \kappa_2} \tag{31}$$

$$\frac{\partial}{\partial \lambda} \ln p(\phi, \psi) = \sin(\phi - \phi_0) \sin(\psi - \psi_0) - \frac{1}{T} \frac{\partial T}{\partial \lambda}, \tag{32}$$

where, given the expression of T (Eq. (13)) and $I'_m(u) = \frac{m}{u} I_m(u) + I_{m+1}(u)$ (see [30]),

$$\frac{\partial T}{\partial \kappa_1} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_{m+1}(\kappa_1) I_m(\kappa_2) \tag{33}$$

$$\frac{\partial T}{\partial \kappa_2} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1) I_{m+1}(\kappa_2) \tag{34}$$

$$\frac{\partial T}{\partial \lambda} = \frac{8\pi^2}{\lambda} \sum_{m=1}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m m I_m(\kappa_1) I_m(\kappa_2). \tag{35}$$

Optimization must be performed under the nonlinear inequality constraint $\lambda^2 < \kappa_1\kappa_2$. A simpler alternative consists in replacing λ by $\rho = \lambda/\sqrt{\kappa_1\kappa_2}$ in the parameterization, so the constraint becomes $\rho \in (-1, 1)$. We then need to replace Eq. (12) by

$$W(\phi, \psi) = \kappa_1 \cos\phi + \kappa_2 \cos\psi + \sqrt{\kappa_1\kappa_2}\rho \sin\phi \sin\psi \tag{36}$$

and (30)-(32) by

$$\frac{\partial}{\partial \kappa_1} \ln p(\phi, \psi) = \cos(\phi - \phi_0) + \frac{\lambda}{2\kappa_1} \sin(\phi - \phi_0) \sin(\psi - \psi_0) - \frac{1}{T} \frac{\partial T}{\partial \kappa_1} \tag{37}$$

$$\frac{\partial}{\partial \kappa_2} \ln p(\phi, \psi) = \cos(\psi - \psi_0) + \frac{\lambda}{2\kappa_2} \sin(\phi - \phi_0) \sin(\psi - \psi_0) - \frac{1}{T} \frac{\partial T}{\partial \kappa_2} \tag{38}$$

$$\frac{\partial}{\partial \rho} \ln p(\phi, \psi) = \frac{\lambda}{\rho} \sin(\phi - \phi_0) \sin(\psi - \psi_0) - \frac{1}{T} \frac{\partial T}{\partial \rho}, \tag{39}$$

with

$$\frac{\partial T}{\partial \kappa_1} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\rho}{2}\right)^{2m} I'_m(\kappa_1) I_m(\kappa_2) \tag{40}$$

$$\frac{\partial T}{\partial \kappa_2} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\rho}{2}\right)^{2m} I_m(\kappa_1) I'_m(\kappa_2) \tag{41}$$

$$\frac{\partial T}{\partial \rho} = \frac{8\pi^2}{\rho} \sum_{m=1}^{\infty} \binom{2m}{m} \left(\frac{\rho}{2}\right)^{2m} m I_m(\kappa_1) I_m(\kappa_2), \tag{42}$$

given

$$T = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\rho}{2}\right)^{2m} I_m(\kappa_1) I_m(\kappa_2). \tag{43}$$

Partial derivative w.r.t. the location parameters ζ^{L}

Given Eqs (19), (22) and (27), we have

$$\frac{\partial}{\partial \zeta_q^{\mathrm{L}}} L(Y;\theta)\bigg|_{\theta=\theta^{\mathrm{old}}} = \frac{\partial}{\partial \zeta_q^{\mathrm{L}}} Q_1(\theta;\theta^{\mathrm{old}})\bigg|_{\theta=\theta^{\mathrm{old}}} = \sum_{d=1}^{D} P_{qd} \frac{\partial}{\partial \zeta_q^{\mathrm{L}}} \ln p(y_d;\zeta_q^{\mathrm{L}},\zeta_q^{\mathrm{S}}). \tag{44}$$

Moreover,

$$\frac{\partial}{\partial \phi_0} \ln p(\phi,\psi) = \kappa_1 \sin(\phi-\phi_0) - \lambda \cos(\phi-\phi_0)\sin(\psi-\psi_0) \tag{45}$$

$$\frac{\partial}{\partial \psi_0} \ln p(\phi,\psi) = \kappa_2 \sin(\psi-\psi_0) - \lambda \sin(\phi-\phi_0)\cos(\psi-\psi_0) \tag{46}$$

Case of multiple datasets

In the case where N residues are available, each conformation is characterized by a unique weight vector, whereas its location and shape parameters are specific to each residue. The identification problem then consists in estimating:

- Q normalized weights $\gamma = (\gamma_q)$ for the protein conformations (classes),
- $5NQ = 3NQ + 2NQ$ shape and location parameters specific to each conformation and each residue, respectively $\zeta_{qn}^{\mathrm{S}} = (\kappa_{1qn}, \kappa_{2qn}, \lambda_{qn})$ and $\zeta_{qn}^{\mathrm{L}} = (\phi_{qn}, \psi_{qn})$.

The log-likelihood then reads

$$L(Y;\theta) = \sum_{n=1}^{N} \sum_{d=1}^{D_n} \ln \left(\sum_{q=1}^{Q} \gamma_q \, p(y_{dn};\zeta_{qn}^{\mathrm{S}},\zeta_{qn}^{\mathrm{L}}) \right)$$

where the nth residue corresponds to D_n observed pairs of angles y_{dn}.

The gradient component relative to each shape or location parameter can still be calculated using the equations (37)-(42) and (44)-(46), respectively, while a summation of Eq.

(28) over all residues must be performed to obtain the gradient components relative to the weight parameters.

The scheme developed here has been used to calculate the relative weigths of the conformations by fitting the probability Ramachandran maps obtained using TALOS-N [1].

Figures

Figure S1

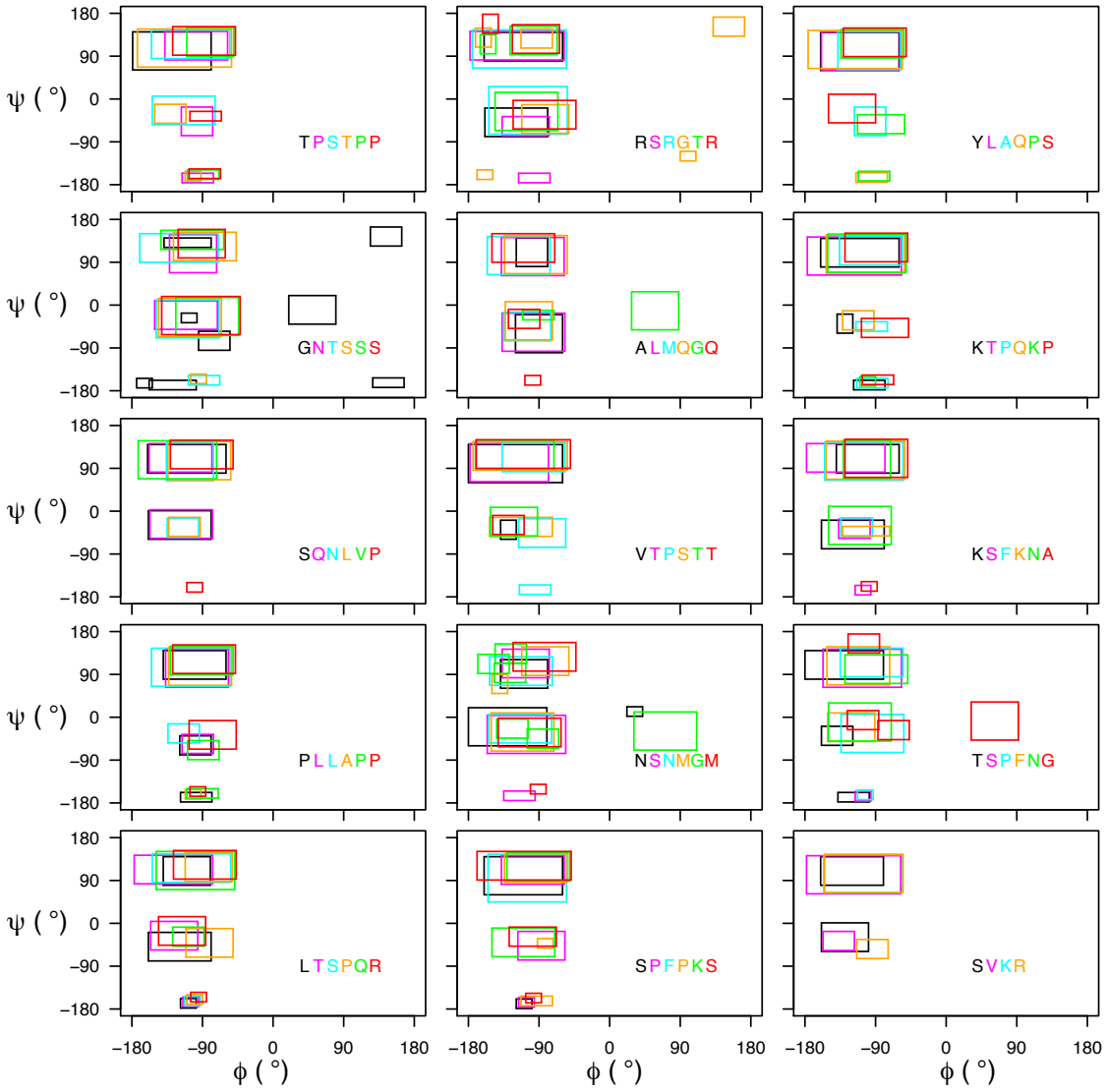


Figure S1: Boxes for backbone angles used as inputs for the run Sic1[1] and obtained from the Ramachandran maps using a threshold of 0.01. The boxes and the corresponding sequence are colored according to the considered residue.

229 Figure S2

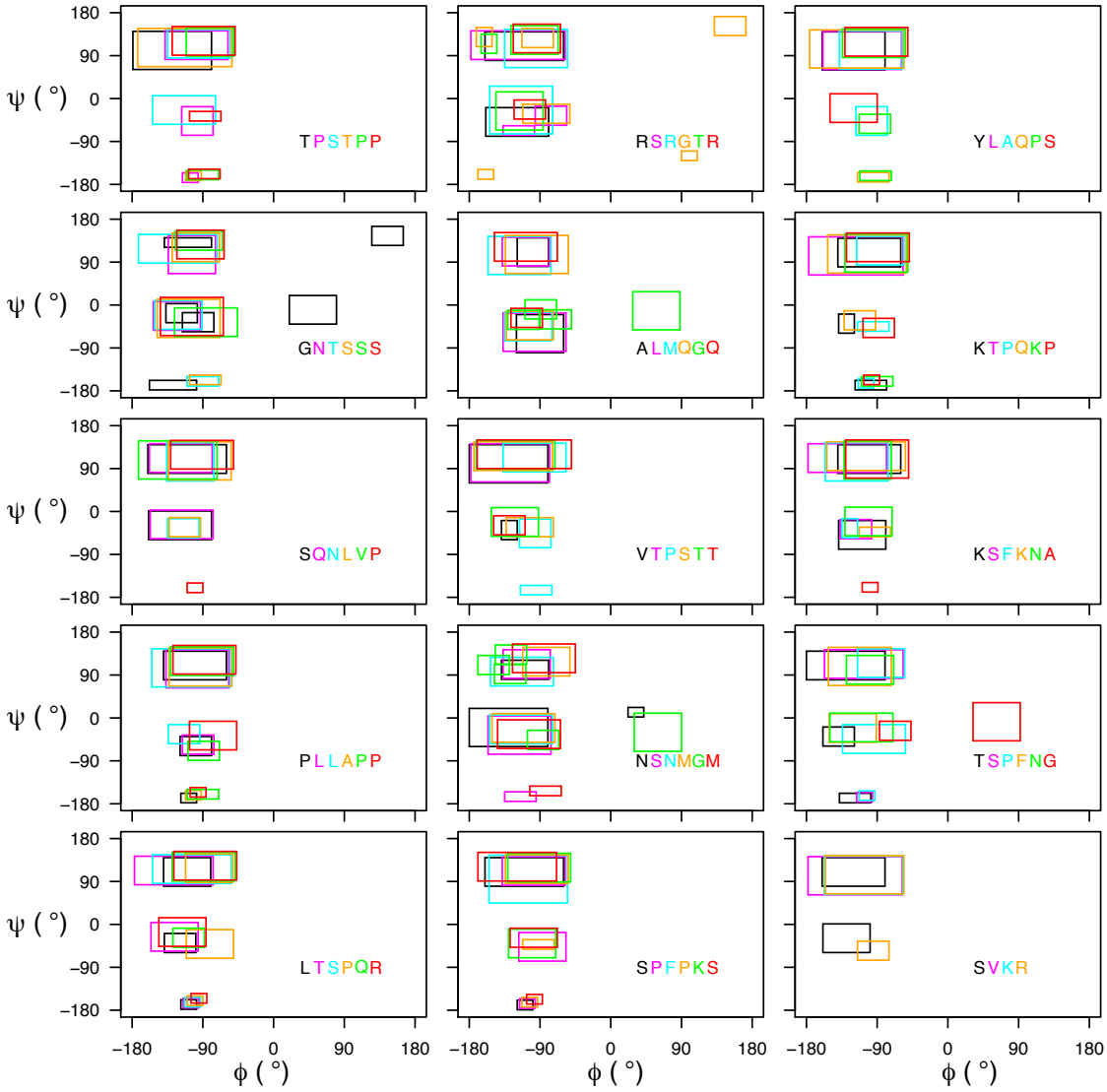


Figure S2: Boxes for backbone angles used as inputs for the run Sic1[2] and obtained from the Ramachandran maps using a threshold of 0.011. The boxes and the corresponding sequence are colored according to the considered residue.

Figure S3

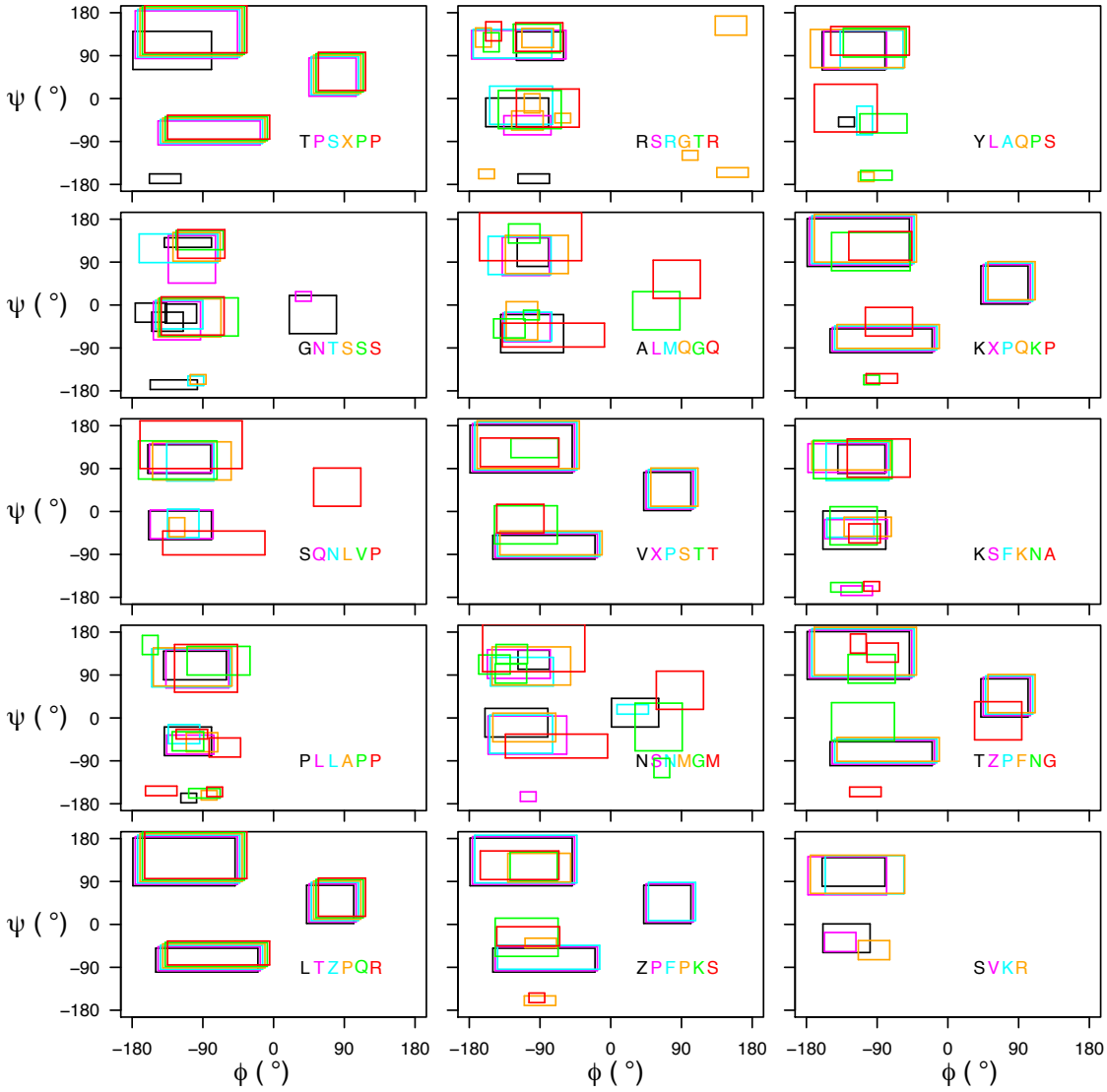


Figure S3: Boxes for backbone angles used as inputs for the run pSic1[1] and obtained from the Ramachandran maps using a threshold of 0.01. The boxes and the corresponding sequence are colored according to the considered residue. The pT and pS residues are marked as X and Z in the sequences.

Figure S4

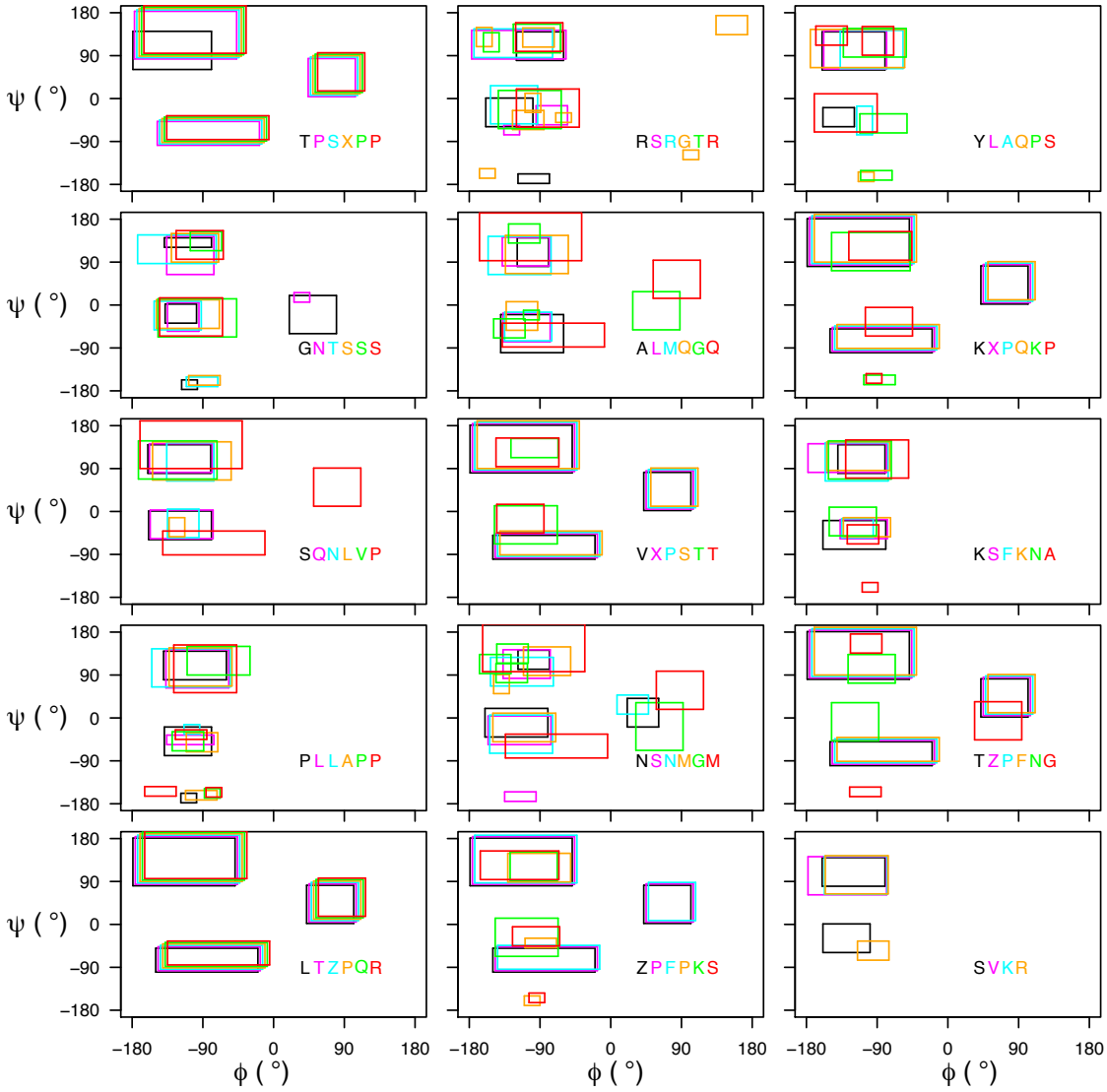


Figure S4: Boxes for backbone angle used as inputs for the run pSic1[2] using a threshold of 0.011. The boxes and the corresponding sequence are colored according to the considered residue. The pT and pS residues are marked as X and Z in the sequences.

232 Figure S5

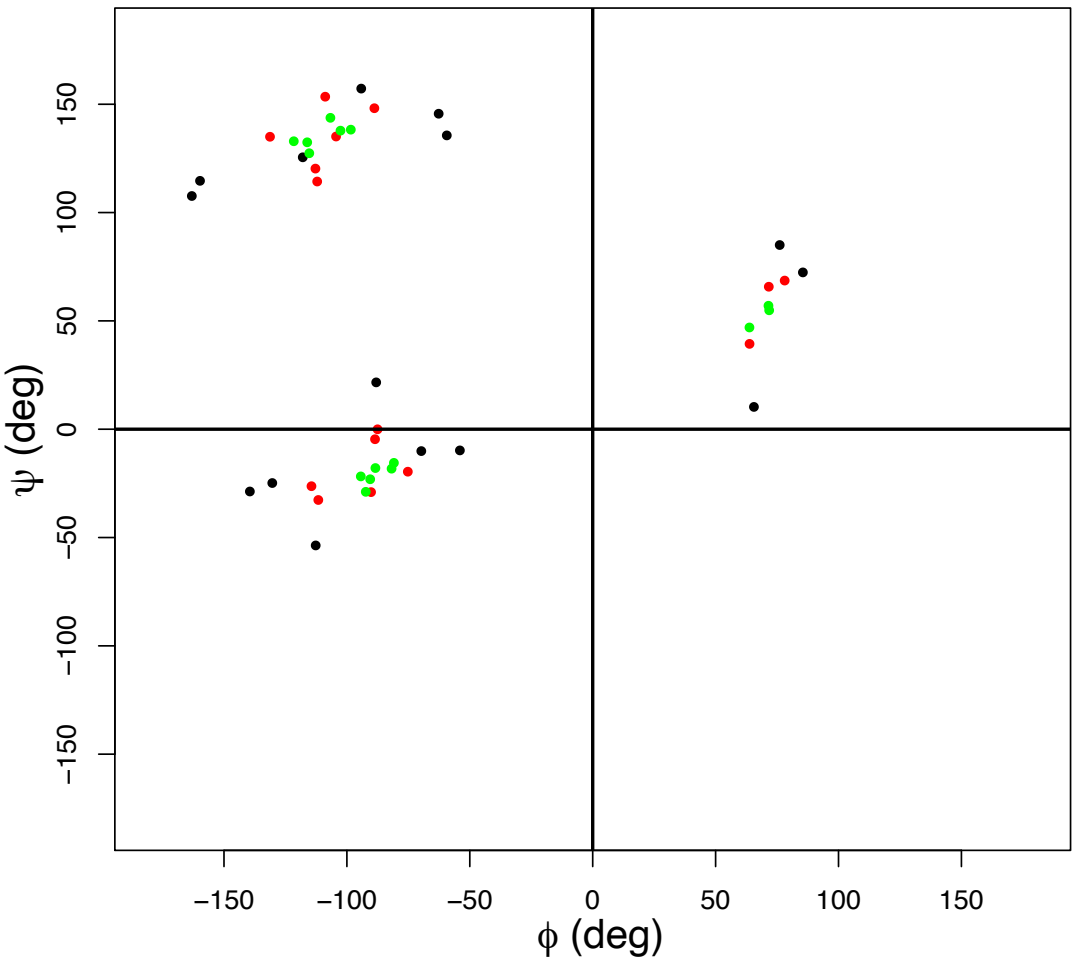


Figure S5: Synthetic Ramachandran plots used for validation of RamaMix. The colors black, red and green correspond to most (large), averaged (medium) and least (narrow) scattered 15 ϕ and ψ values. These synthetic data correspond to five hypothetical residues located in three conformations, the relative weights of conformations being 56.8%, 11.6 % and 31.6 %.

233 Figure S6

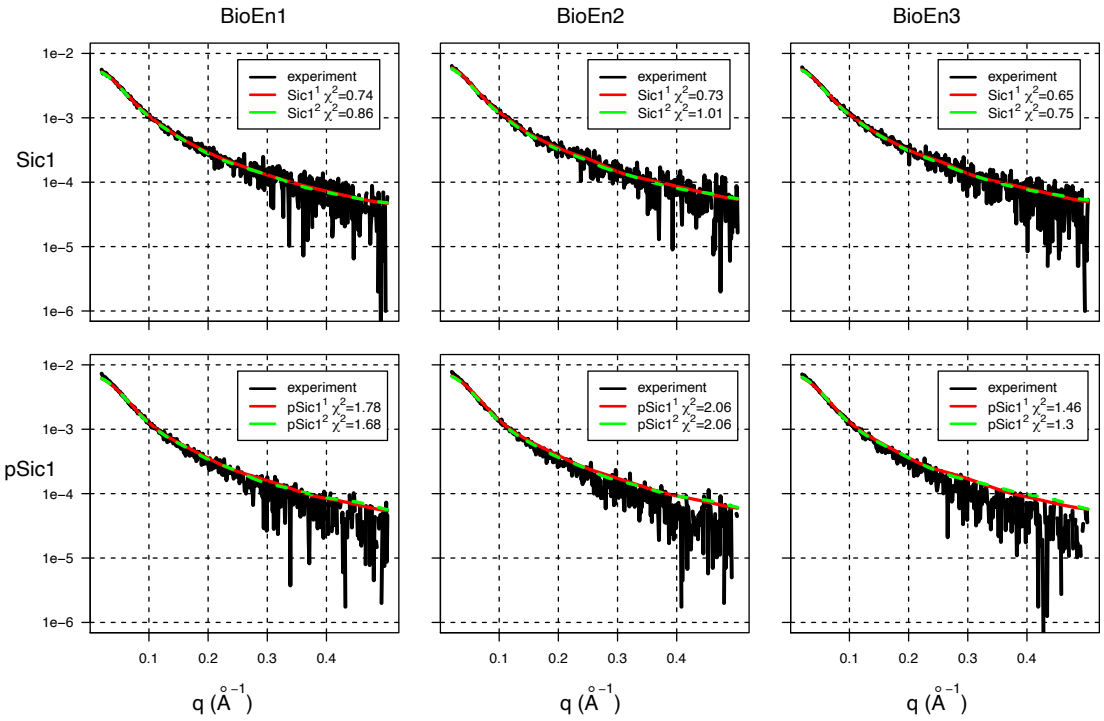


Figure S6: Superimposition of experimental SAXS curves with the reconstructed SAXS curves from the conformations of Sic1 and pSic1 selected by BioEn. The reconstructions of the SAXS curves from the selected conformations are plotted using red and green solid lines, depending on the TAiBP first (Sic1^1, pSic1^1) or second (Sic1^2, pSic1^2) run.

Figure S7

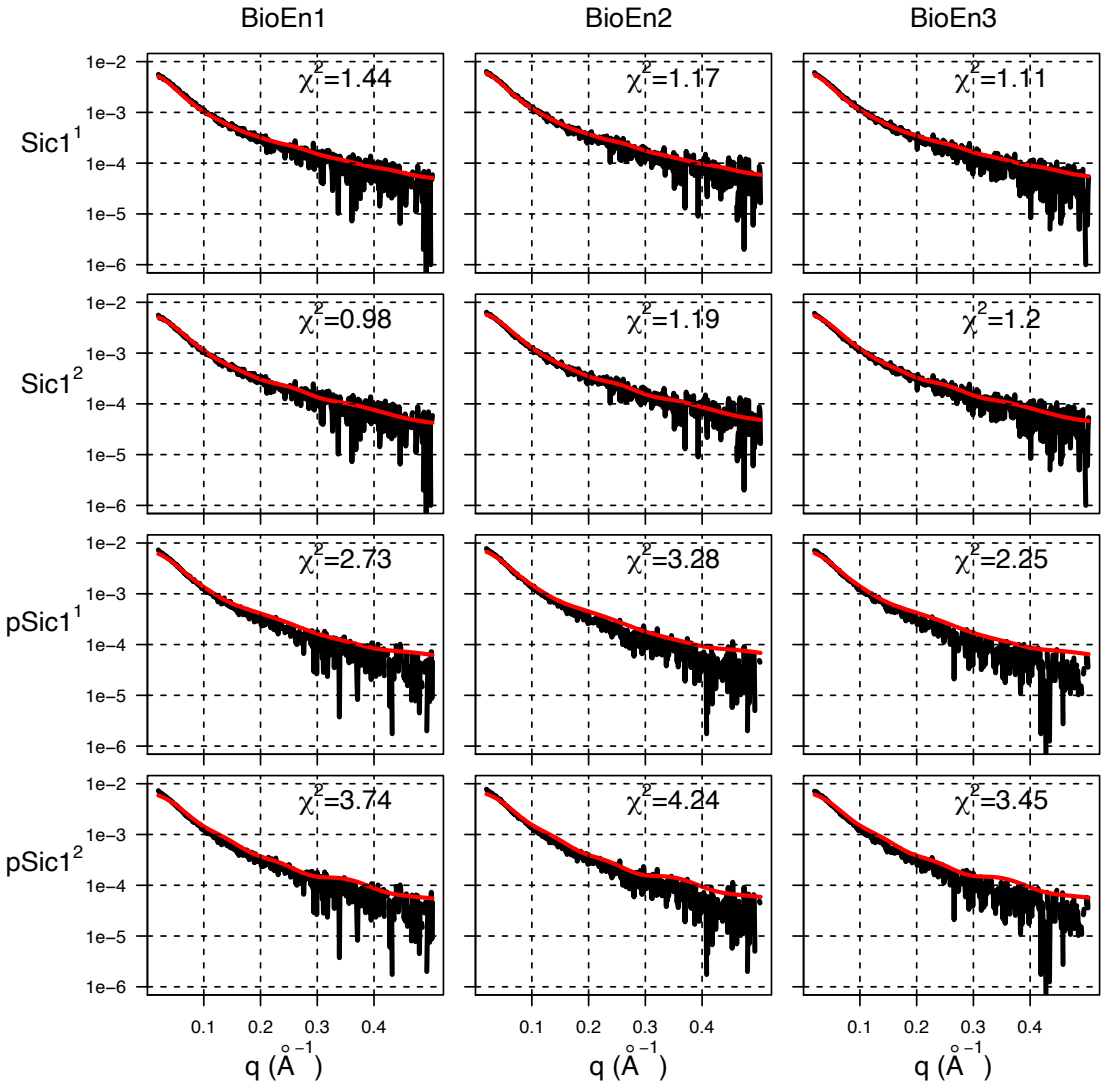


Figure S7: Superimposition of experimental SAXS curves (black) with the reconstructed SAXS curves (red) from the conformations of Sic1 and pSic1 selected by RamaMix.

Figure S8

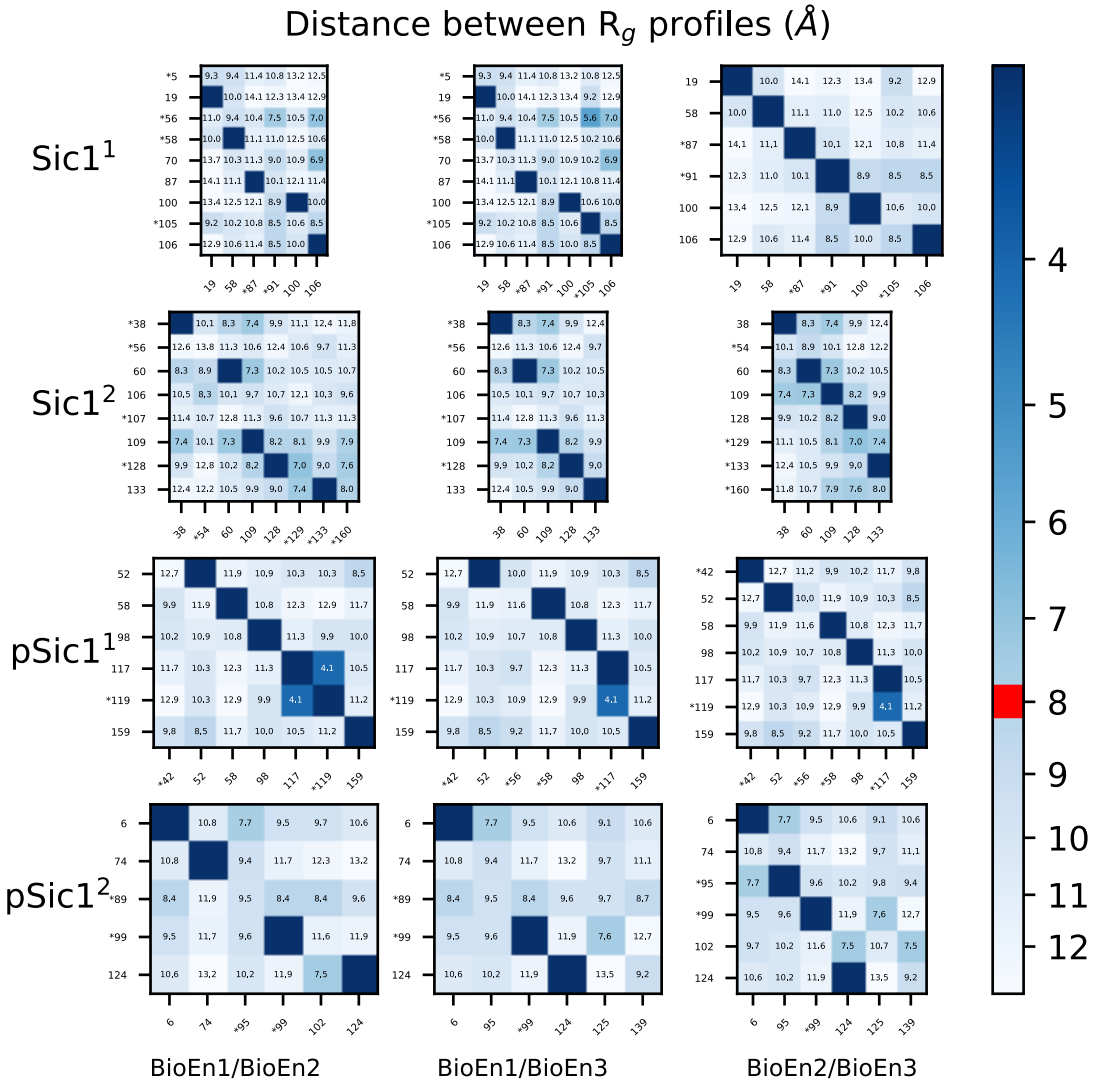


Figure S8: Distances between the profiles P_q (Eq. 11 of the main text) for local gyration radii between the conformations selected from different fittings of SAXS curves (BioEn1,BioEn2, BioEn3). The limit of 8 Å used for the superposed plots of profiles (Figure 5 of the main text) is drawn in red on the scale of distance.

Figure S9

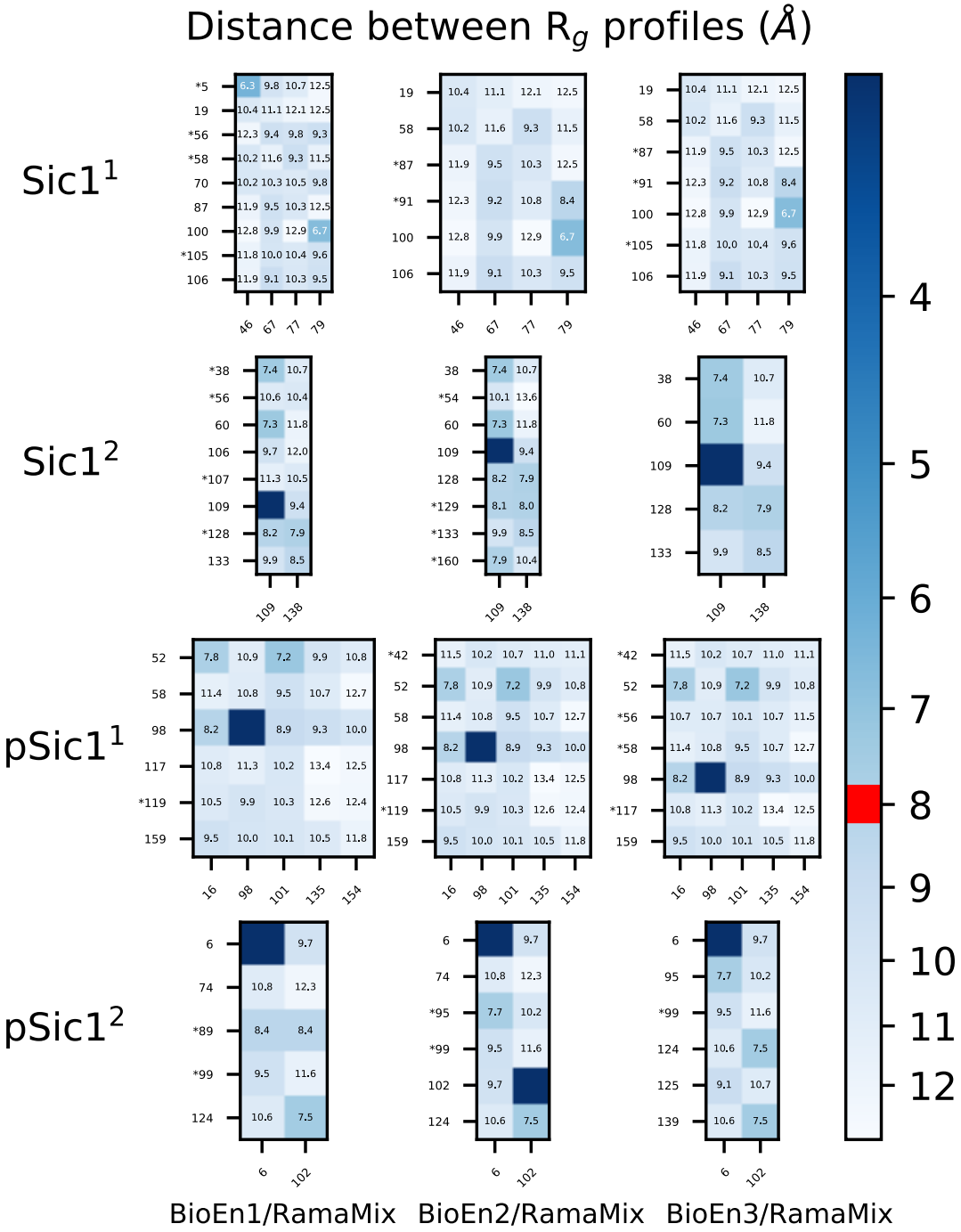
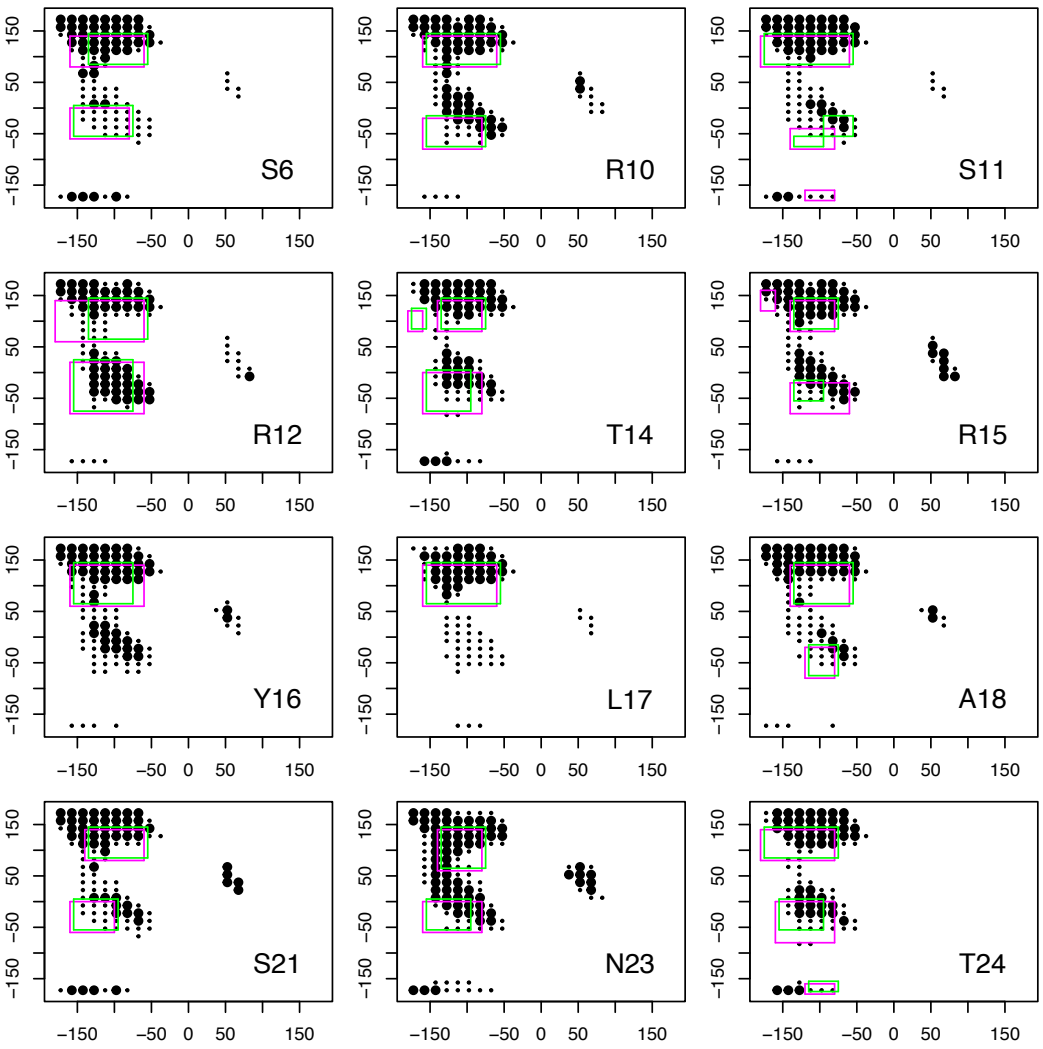
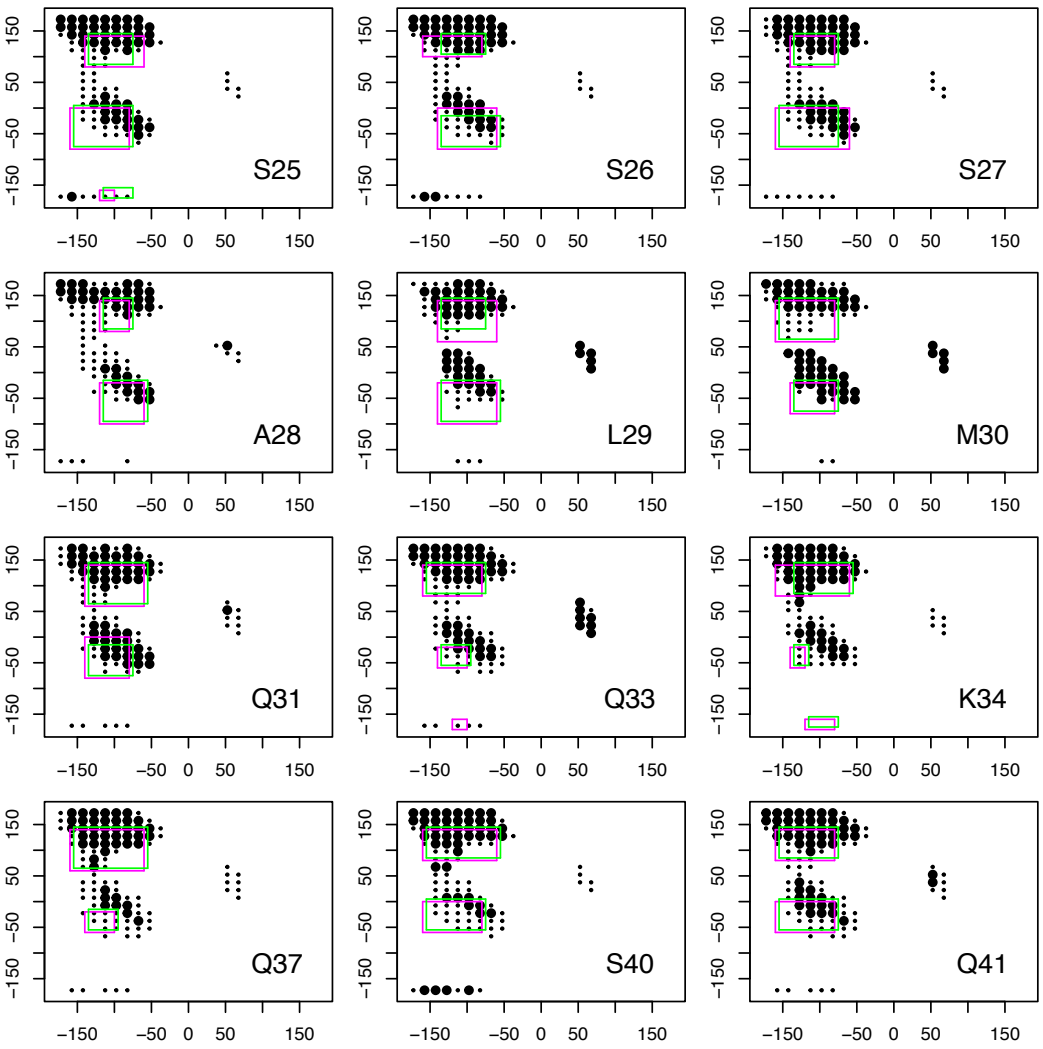


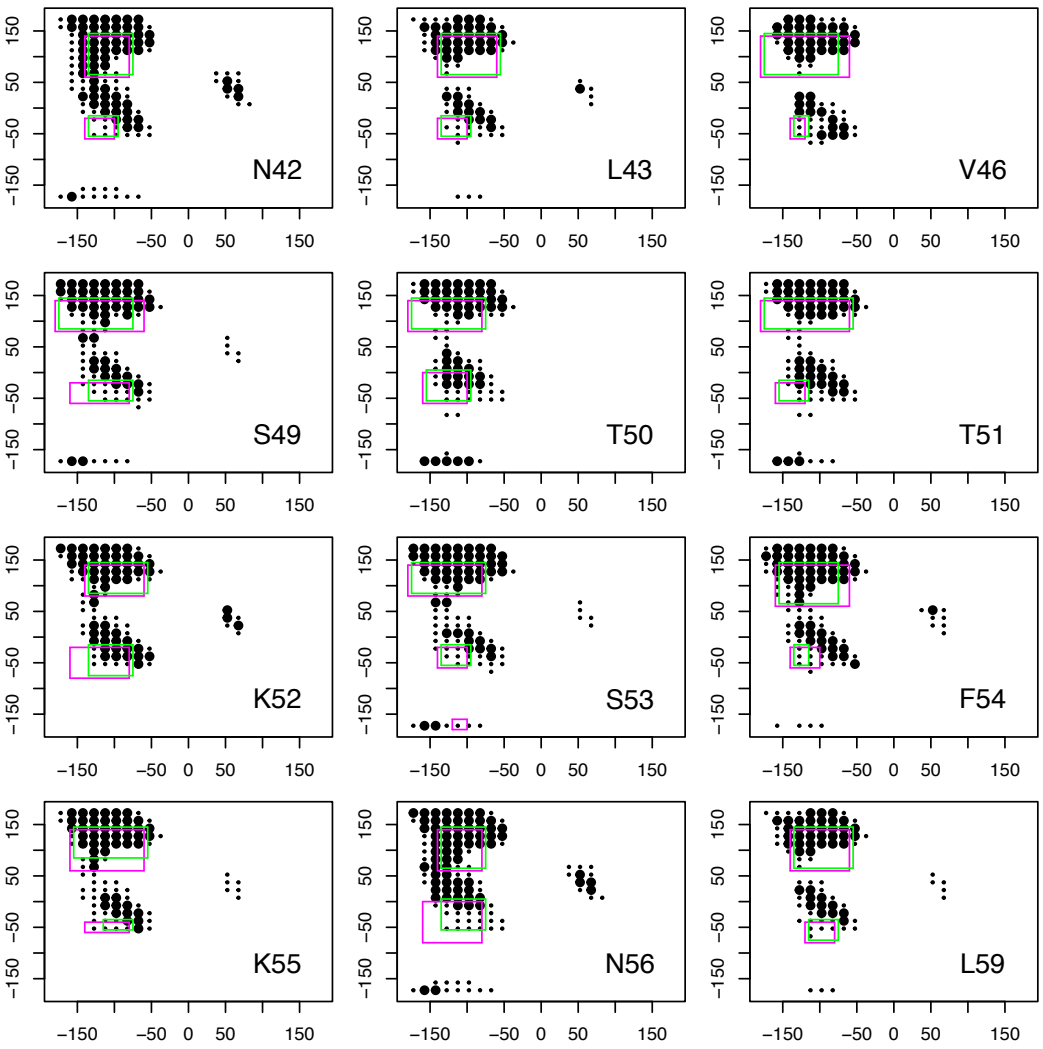
Figure S9: Distances between the profiles P_q (Eq. 11 of the main text) for local gyration radii between the conformations selected from fitting of SAXS curves (BioEn1,BioEn2, BioEn3) and of Ramachandran maps (RamaMix). The limit of 8 Å used for the superposed plots of profiles (Figure 5 of the main text) is drawn in red on the scale of distance.

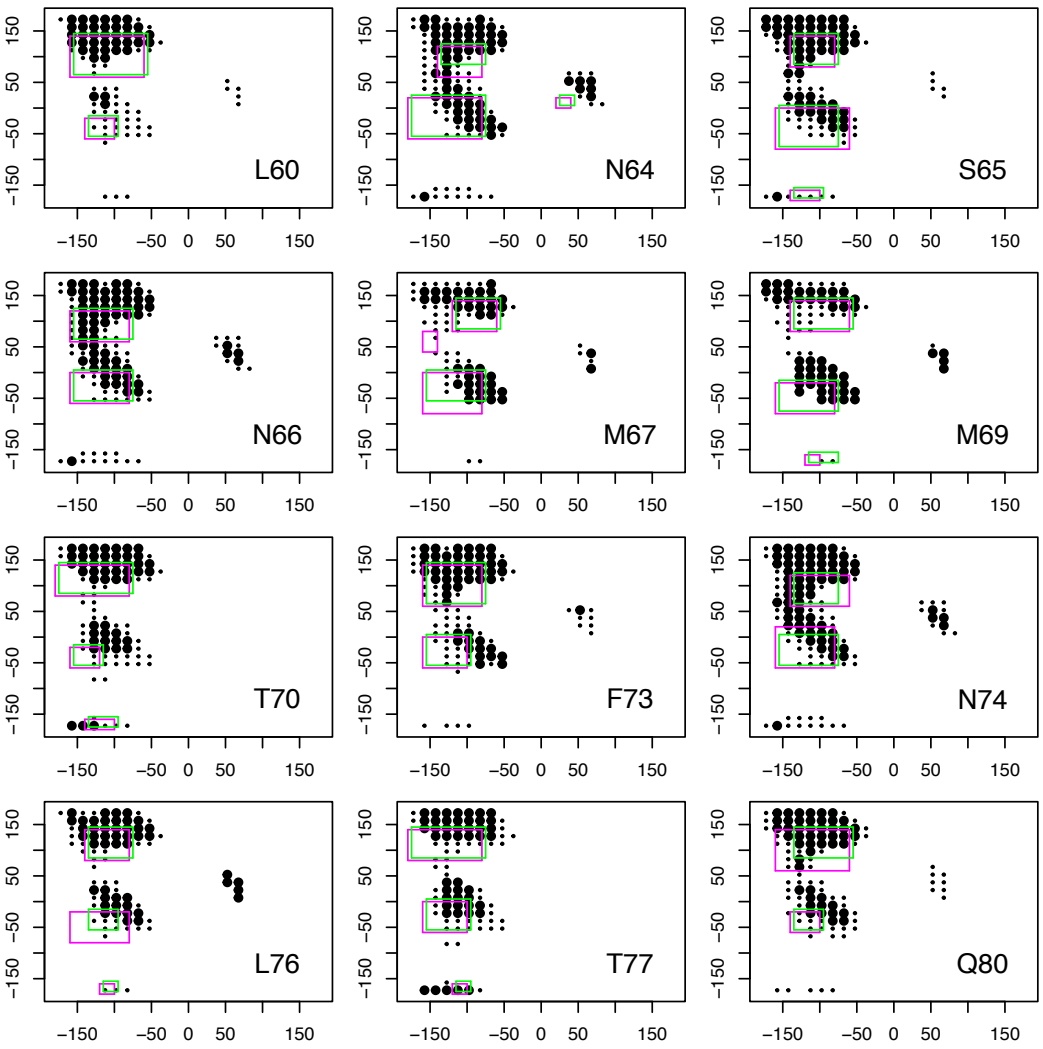
Figure S10 (5 next pages)

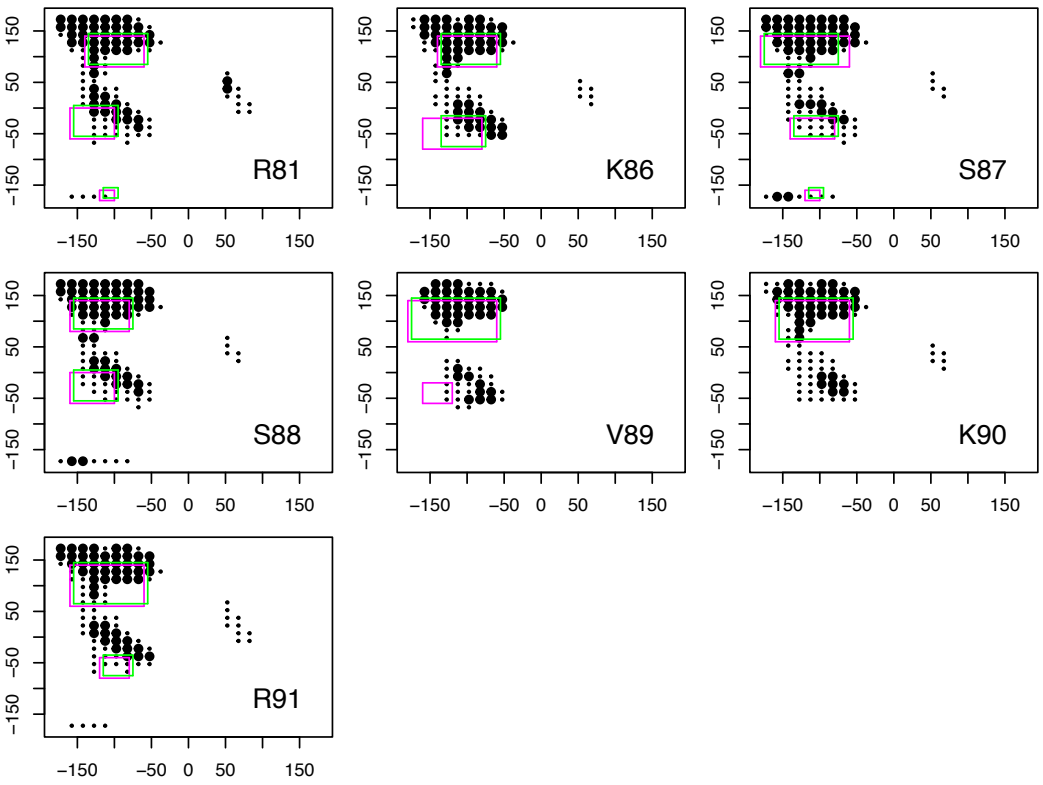
Figure S10: Superimposition of the MERA ϕ, ψ distributions obtained on residues of Sic1 with the (ϕ, ψ) input boxes for TAiBP. The size of the points on MERA distribution is large for predicted probability values larger than 0.005 and small for the other probability values. The TAiBP input boxes are colored in magenta and green for the duplicated TAiBP runs: Sic1^1 and Sic1^2.


S6
R10
S11
R12
T14
R15
Y16
L17
A18
S21
N23
T24
150
50
−50
−150
−150
−50
0
50
150


S25
S26
S27
A28
L29
M30
Q31
Q33
K34
Q37
S40
Q41
150
50
−50
−150
−150
−50
0
50
150

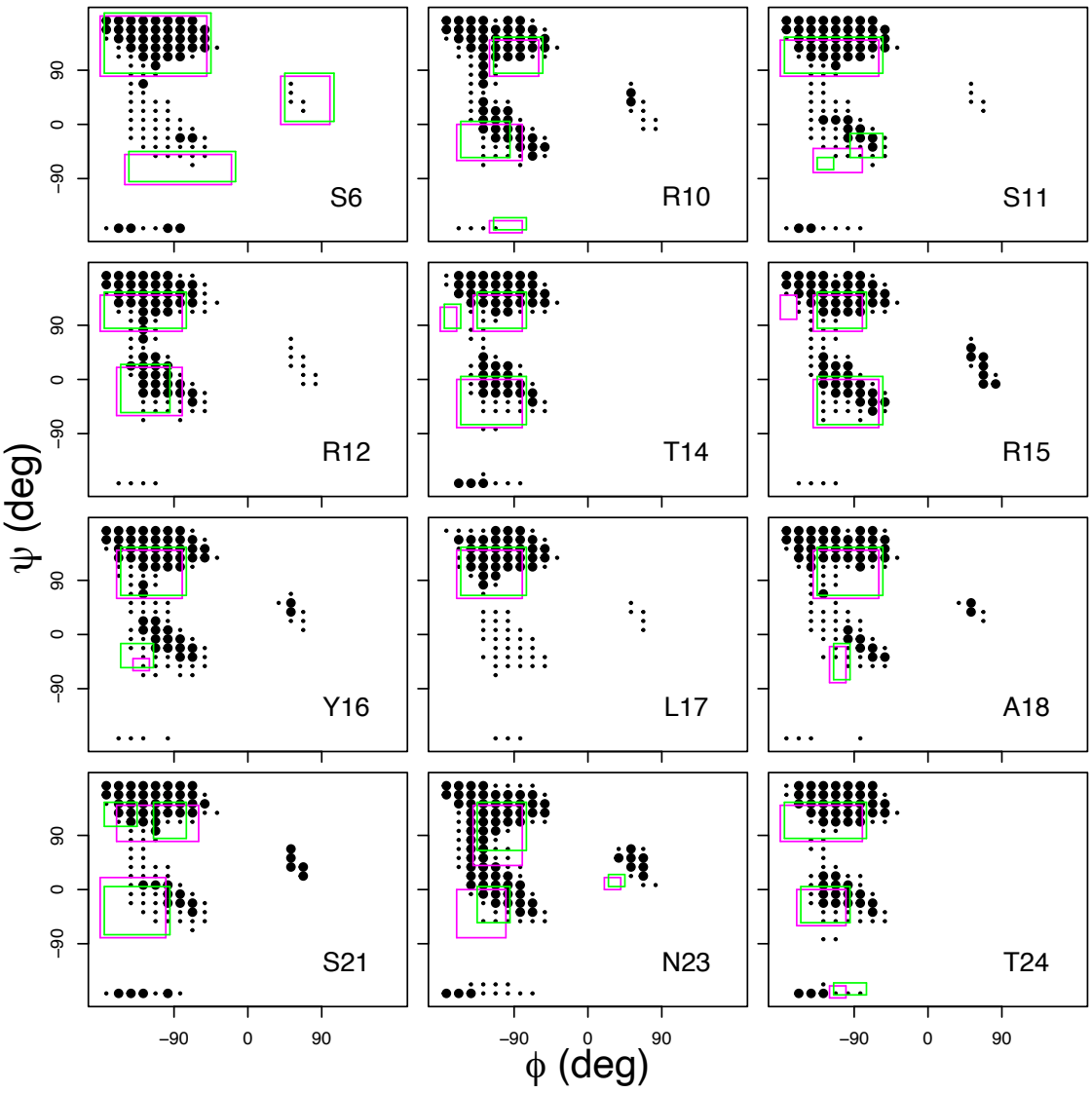

N42
L43
V46
S49
T50
T51
K52
S53
F54
K55
N56
L59
150
50
−50
−150
−150
−50
0
50
150

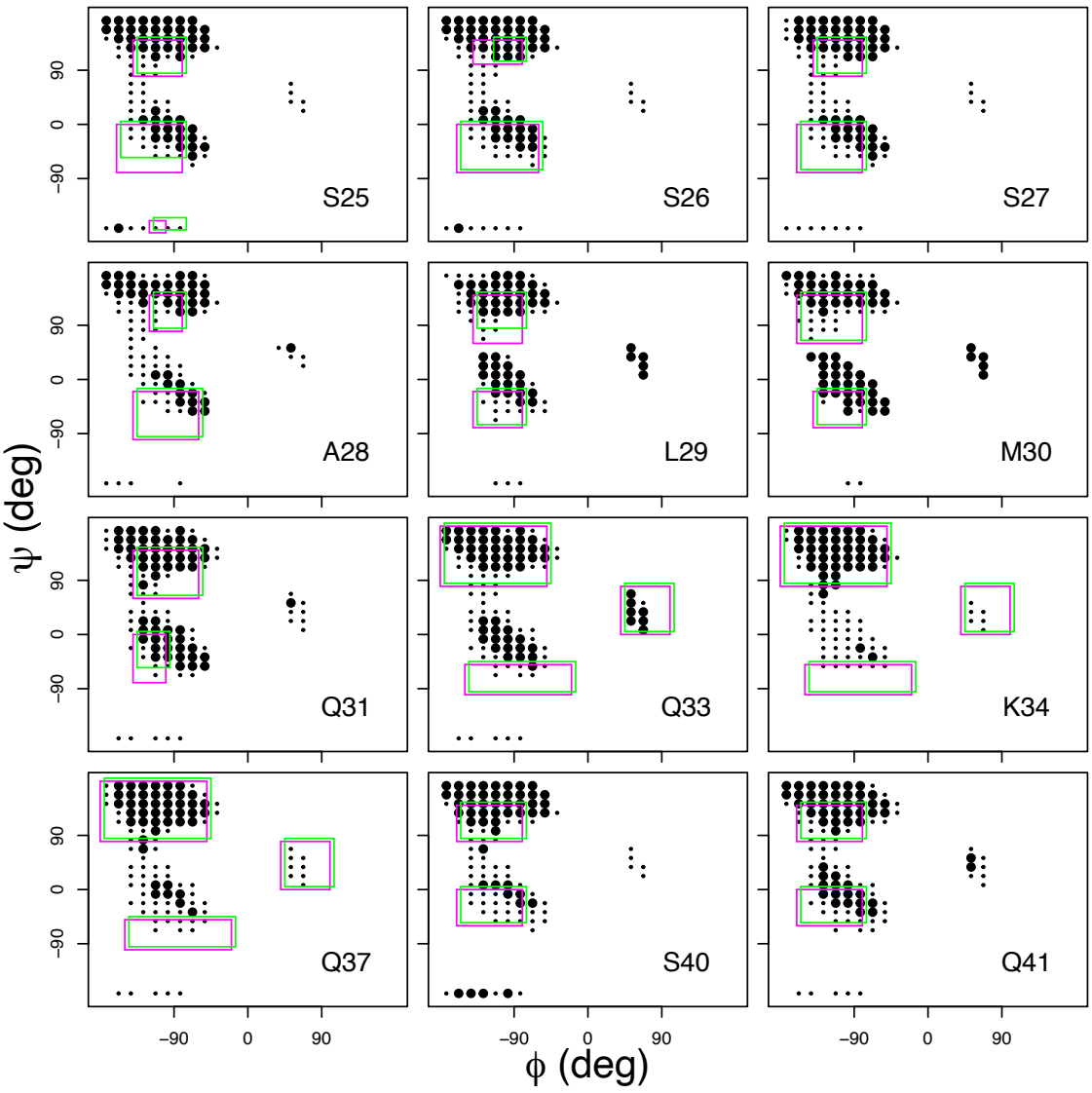

L60
N64
S65
N66
M67
M69
T70
F73
N74
L76
T77
Q80
150
50
−50
−150
−150
−50
0
50
150

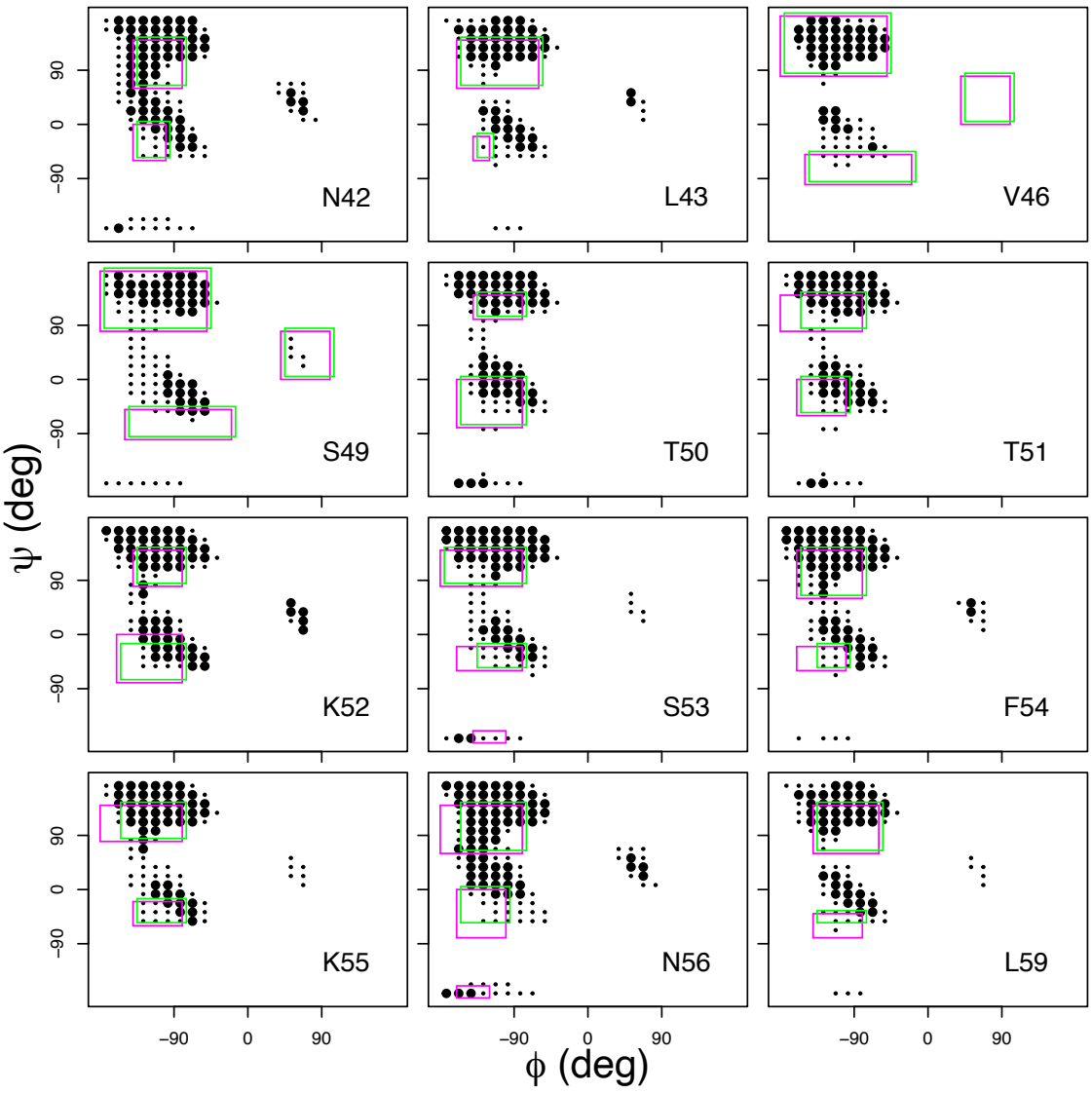

R81
K86
S87
S88
V89
K90
R91
150
50
−50
−150
−150
−50
0
50
150

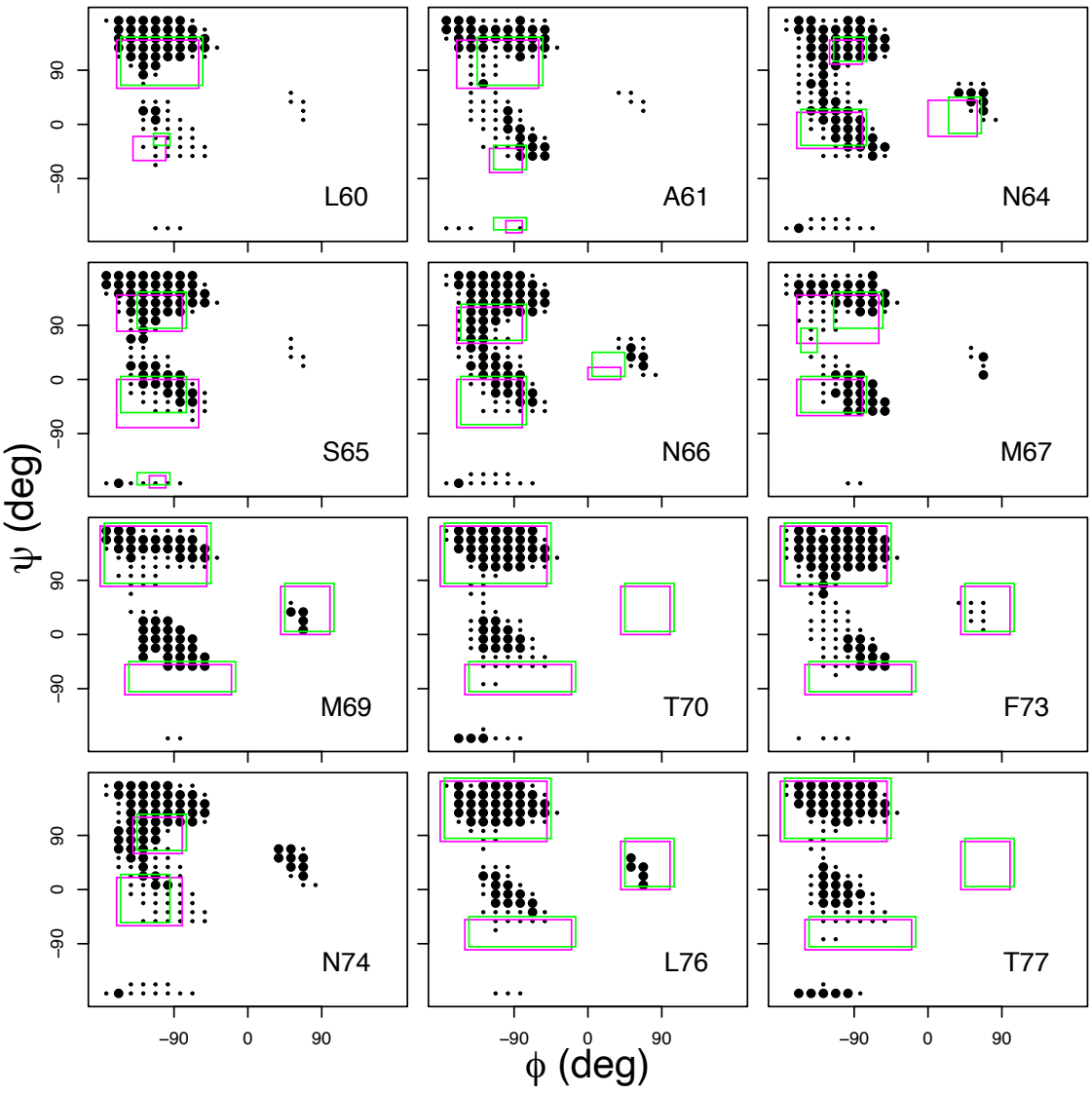
243 Figure S11 (5 next pages)

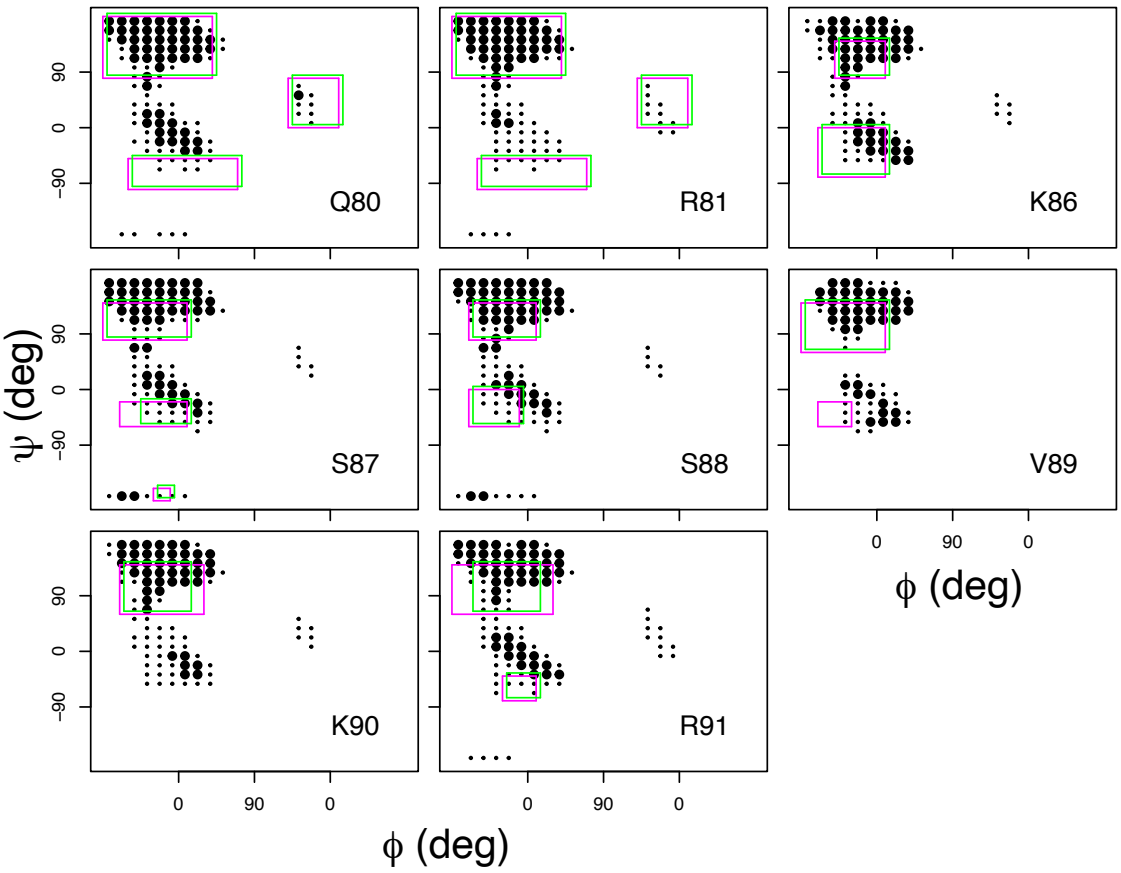
Figure S11: Superimposition of the MERA ϕ, ψ distributions obtained on residues of pSic1 with the (ϕ, ψ) input boxes for TAiBP. The size of the points on MERA distribution is large for predicted probability values larger than 0.005 and small for the other probability values. The TAiBP input boxes are colored in magenta and green for the duplicated TAiBP runs: pSic1^1 and pSic1^2.


S6
R10
S11
R12
T14
R15
Y16
L17
A18
S21
N23
T24
ψ (deg)
ϕ (deg)
−90
0
90


S25
S26
S27
A28
L29
M30
Q31
Q33
K34
Q37
S40
Q41
ψ (deg)
ϕ (deg)
−90
0
90


N42
L43
V46
S49
T50
T51
K52
S53
F54
K55
N56
L59
ψ (deg)
ϕ (deg)
−90
0
90


L60
A61
N64
S65
N66
M67
M69
T70
F73
N74
L76
T77
ψ (deg)
ϕ (deg)
−90
0
90


Q80
R81
K86
S87
S88
V89
K90
R91
ψ (deg)
ϕ (deg)
90
0
−90

249 Figure S12

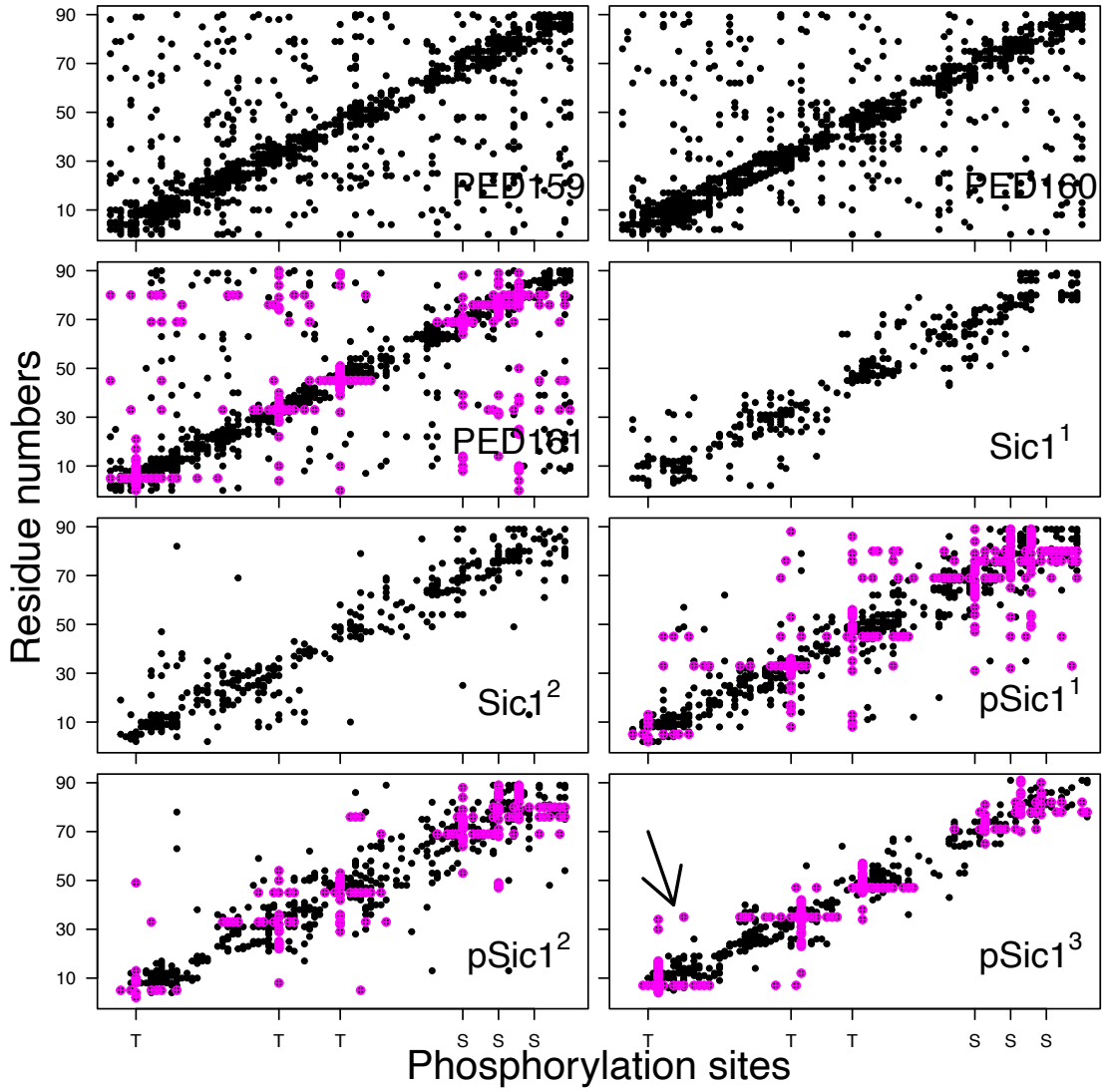


Figure S12: Contact maps displaying cumulative hydrogen bonds observed for the conformation sets PED159(Sic1), PED160(Sic1), PED161(pSic1) refined using MD simulations, as well as for the TAiBP conformational sets Sic1^1, Sic1^2, pSic1^1, pSic1^2 and pSic1^3. The hydrogen bonds involving sidechains of phosphorylated residues are plotted in magenta, whereas the other hydrogen bonds are plotted in black. An arrow on the contact map of pSic1^3 indicates the presence of few long-range hydrogen bonds involving phosphorylated residues.

Fragment	Residue range Sic1[1]	Residue range Sic1[2]	Fragment	Residue range pSic1[1]	Residue range pSic1[2]	Fragment	Residue range pSic1[3]
Pept1	4-12	4-12	Pept1 (pT7)	4-11	4-11	Pept1 (pT7)	4-11
Pept2	10-19	10-19	Pept2	9-15	9-15	Pept2	9-15
Pept3	17-25	17-25	Pept3	13-21	13-21	Pept3	13-21
Pept4	23-32	23-31	Pept4	19-26	19-26	Pept4	19-26
Pept5	30-39	29-38	Pept5	24-32	24-32	Pept5	24-32
Pept6	37-48	36-47	Pept6 (pT35)	30-37	30-37	Pept6 (pT35)	30-37
Pept7	46-56	45-56	Pept7 (pT35)	35-44	35-44	Pept7 (pT35)	35-42
Pept8	54-63	54-63	Pept8 (pT47)	42-51	42-51	Pept8 (pT47)	40-47
Pept9	61-67	61-68	Pept9	49-57	49-58	Pept9	45-51
Pept10	65-71	66-73	Pept10	55-62	56-63	Pept10	49-58
Pept11	69-76	71-79	Pept11	60-66	61-67	Pept11	56-63
Pept12	74-81	77-85	Pept12	64-70	65-71	Pept12	61-67
Pept13	79-87	83-91	Pept13 (pS71)	68-74	69-75	Pept13 (pS71)	65-71
Pept14	85-91	-	Pept14 (pS78)	72-78	73-79	Pept14 (pS78)	69-75
			Pept15 (pS78)	76-82	77-83	Pept15 (pS78)	73-79
			Pept16 (pS82)	80-86	81-87	Pept16 (pS82)	77-83
			Pept17	84-91	85-91	Pept17 (pS82)	81-87
						Pept18	85-91

Table S1: Peptide fragments used for TAiBP runs. The phosphorylated residues in pSic1 are indicated as pS and pT.

ϕ interval	ψ interval
-150 -20	-100 -50
-180 -50	80 180
40 100	0 80

Table S2: Definition of backbone angle generic boxes used for residues of pSic1 on which TALOS-N [1] does not produce a prediction.

Conformations	BioEn1	BioEn2	BioEn3
$\text{Sic1}^1/BioEn1$	0.74	3.74	1.45
$\text{Sic1}^1/BioEn2$	3.64	0.73	1.5
$\text{Sic1}^1/BioEn3$	1.27	1.57	0.65
$\text{Sic1}^2/BioEn1$	0.86	4.71	1.98
$\text{Sic1}^2/BioEn2$	1.97	1.01	0.76
$\text{Sic1}^2/BioEn3$	0.96	2.13	0.75
$\text{pSic1}^1/BioEn1$	1.78	4.21	1.59
$\text{pSic1}^1/BioEn2$	3.65	2.06	2.53
$\text{pSic1}^1/BioEn3$	1.8	3.41	1.46
$\text{pSic1}^2/BioEn1$	1.68	4.42	1.51
$\text{pSic1}^2/BioEn2$	2.24	2.06	1.59
$\text{pSic1}^2/BioEn3$	1.69	3.34	1.3

Table S3: Values of χ^2 between experimental and reconstructed SAXS curves obtained for the various sets of conformations selected by BioEn on Sic1 and pSic1. The Table columns are labeled with experimental SAXS curves, and the Table rows are labeled with the sets of conformations selected from the fitting of SAXS curves.

Residue position	order
first	N, H1, H2, CA, N, HA, CA, C
inner	N, -O, -CA, -C, N, CA, C, +N, -C, N, CA, H1, N, CA, C, HA, C, CA
last	N, -O, -CA, -C, N, CA, C, -C, N, CA, H1, N, CA, C, HA, C, CA, O, C, O2

Table S4: Atom re-ordering used during the iBP calculation step within the first, the last and the inner residues of the peptide fragment. The order is described by the list of atoms names, the signs "-" and "+" describing atoms located in the previous and the next residues in the primary sequence.

A. pSic1[3]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	70	51.6 ± 1.3e-3	**70**	42.9 ± 1.1e-3	**70**	40.2 ± 0.3
	40	39.6 ± 1.2e-3	**40**	45.7 ± 3.6e-4	**40**	43.6 ± 7.2e-2
	49	8.8 ± 3.9e-4	**49**	11.4 ± 2.0e-4	**49**	15.2 ± 6.1e-2
Average final $\chi 2$		0.9		1.2		0.7
Average final S_{KL}		-2.1e-9		-3.4e-8		-2.9e-10
B. pSic1[13]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	239	47.2 ± 2.3e-3	**239**	44.1 ± 8.6e-4	**239**	50.5 ± 1.8e-3
	249	8.6 ± 6.4e-4	**249**	12.0 ± 4.2e-4	**249**	14.1 ± 1.9e-4
	52	29.0 ± 5.4e-4	**52**	26.3 ± 9.3e-4	**52**	19.4 ± 4.8e-4
	54	15.2 ± 1.1e-3	**54**	17.6 ± 7.2e-4	**54**	15.9 ± 7.4e-4
Average final $\chi 2$		0.8		0.9		0.7
Average final S_{KL}		-1.5e-9		-5.4e-9		-2.6e-8
C. pSic1[23]	conformation numbers	populations percentages	conformation numbers	populations percentages	conformation numbers	populations percentages
	239	31.6 ± 2.1e-3	**239**	25.7 ± 9.0	**239**	39.6 ± 6.1e-4
	249	28.5 ± 8.5e-4	**249**	30.3 ± 5.9	**249**	31.2 ± 3.3e-4
			240	2.1 ± 6.3		
			316	3.1 ± 9.1		
			47	1.6 ± 4.9		
	74	16.0 ± 5.4e-4	**74**	16.5 ± 5.5	**74**	16.6 ± 2.3e-4
	99	23.9 ± 5.4e-4	**99**	20.7 ± 3.2	**99**	12.6 ± 6.9e-4
Average final $\chi 2$		0.8		0.9		0.7
Average final S_{KL}		-3.2e-9		-4.6e-9		-2.9e-10

Table S5: Conformations and populations selected using BioEn 0.1.1 [31] on the three sets of SAXS curves. The conformations were generated by the runs pSic1[3] and then pooled with pSic1[1] and pSic1[2] to produce pSic1[13] pSic1[23]. For each SAXS curve and set of protein conformations, after ten runs starting from random values of populations and performed on the whole set of conformations, all conformations for which the sum of populations over the ten runs was larger than 0.01 were gathered, and a second run of ten additional BioEn calculations was performed on this reduced set of conformations. The average and standard deviation values of populations obtained for each selected conformation from the second set of BioEn runs, are given in the Table, along with the final average values of reduced χ^2 and of entropy S_{KL}. The labels of conformations selected in at least two runs are written in bold. Numbers larger than 200 in pSic1[13] and pSic1[23] were assigned to the conformations from pSic1[3].

A. pSic1^{3}	conformation numbers	populations percentages
	40	28.5 ± 3.2e-5
	47	39.5 ± 3.8e-5
	49	32.0 ± 3.1e-5
B. pSic1^{13}	conformation numbers	populations percentages
	247	39.5 ± 2.7
	249	31.7 ± 3.0
	240	28.8 ± 0.9
C. pSic1^{23}	conformation numbers	populations percentages
	240	28.8 ± 1.0
	247	39.4 ± 2.6
	249	31.7 ± 2.8

Table S6: Conformations and populations selected by fitting of the Ramachandran maps using RamaMix. For each set of protein conformations, 100 runs were performed starting from random values for the populations. The few optimizations which did not converge, were discarded: 2 for pSic1^{13} and for pSic1^{23}. The backbone angles ϕ and ψ were allowed to move up to 15°. The populations of conformations for the converged runs were averaged and these mean values are given as percentages in the Table along with the corresponding standard deviation values. The labels of conformations also selected by BioEn are written in bold.Numbers larger than 200 in pSic1^{13} and pSic1^{23} were assigned to the conformations from pSic1^{3}.

References

[1] Y. Shen and A. Bax. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol*, 1260:17–32, 2015.

[2] L. Liberti, C. Lavor, and A. Mucherino. The discretizable molecular distance geometry problem seems easier on proteins. *Distance Geometry: Theory, Methods and Applications. Mucherino, Lavor, Liberti, Maculan (eds.)*, pages 47–60, 2014.

[3] C Lavor, R Alves, W Figueiredo, A Petraglia, and N Maculan. Clifford Algebra and the Discretizable Molecular Distance Geometry Problem. *Adv. Appl. Clifford Algebras*, 25:925–942, 2015.

[4] B Worley, F Delhommel, F Cordier, TE Malliavin, B Bardiaux, N Wolff, M Nilges, C Lavor, and L Liberti. Tuning interval Branch-and-Prune for protein structure determination. *Journal of Global Optimization*, 72:109–127, 2018.

[5] R Engh and R Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A*, 47:392–400, 1991.

[6] T. E. Malliavin. Tandem domain structure determination based on a systematic enumeration of conformations. *Sci Rep*, 11:16925, 2021.

[7] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[8] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol Cybern*, 43:59–69, 1982.

[9] T. Kohonen. Self-organizing maps. *Springer Series in Information Sciences, Heidelberg, Germany.*, 2001.

[10] L. Miri, G. Bouvier, A. Kettani, A. Mikou, L. Wakrim, M. Nilges, and T. E. Malliavin. Stabilization of the integrase-DNA complex by Mg^{2+} ions and prediction of key residues for binding HIV-1 integrase inhibitors. *Proteins*, 82(3):466–478, Mar 2014.

[11] G Bouvier, N Duclert-Savatier, N Desdouits, D Meziane-Cherif, A Blondel, P Courvalin, M Nilges, and TE Malliavin. Functional motions modulating VanA ligand binding unraveled by self-organizing maps. *J Chem Inf Model*, 54:289–301, 2014.

[12] YG Spill, G Bouvier, and M Nilges. A convective replica-exchange method for sampling new energy basins. *J Comput Chem*, 34:132–140, 2013.

[13] G. W. Gomes, M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradinaru. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc*, 142:15697–15710, 2020.

[14] JC Phillips, R Braun, W Wang, J Gumbart, E Tajkhorshid, E Villa, C Chipot, RD Skeel, L Kale, and K Schulten. Scalable molecular dynamics with NAMD. *J Comput Chem*, 26:1781–1802, 2005.

[15] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Mackerell. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ and ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput*, 8:3257–3273, 2012.

[16] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*, 14:71–73, 2017.

[17] D. E. Tanner, K. Y. Chan, J. C. Phillips, and K. Schulten. Parallel Generalized Born Implicit Solvent Calculations with NAMD. *J Chem Theory Comput*, 7(11):3635–3642, Nov 2011.

[18] J.P. Ryckaert, G. Ciccotti, and HJC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints and Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[19] HC Andersen. Rattle: a "Velocity" Version of the Shake Algorithm for Molecular Dynamics Calculations. *J Comp Phys*, 52:24–34, 1983.

[20] D Frenkel and B Smit. *Understanding molecular simulation: from algorithms to applications*. Academic press, San Diego, California, 2002.

[21] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, NY, 2nd edition, 1998.

[22] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley and Sons, Inc., 1997.

[23] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[24] Kanti V. Mardia, Gareth Hughes, Charles C. Taylor, and Harshinder Singh. A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, 36(1):99–109, 2008.

[25] Harshinder Singh, Vladimir Hnizdo, and Eugene Demchuk. Probabilistic model for two dependent circular variables. *Biometrika*, 89(3):719–723, 2002.

[26] Kanti V. Mardia, Charles C. Taylor, and Ganesh K. Subramaniam. Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data. *Biometrics*, 63(2):505–512, 2007.

[27] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci U S A*, 105:8932–8937, 2008.

[28] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, September 1995.

[29] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Verlag, New York, NY, 2005.

[30] D. E. Amos. Computation of modified Bessel functions and their ratios. *Mathematics of Computation*, 28(125):239–251, 1974.

[31] J. Köfinger, L. S. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer. Efficient Ensemble Refinement by Reweighting. *J Chem Theory Comput*, 15(5):3390–3401, May 2019.