



HAL
open science

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

► **To cite this version:**

Daniel Förster, Leo Liberti, Antonio Mucherino, Jung-Hsin Lin, Thérèse E Malliavin, et al.. Low-resolution description of the conformational space for intrinsically disordered proteins. *Scientific Reports*, In press, 10.1038/s41598-022-21648-9 . hal-03796134

HAL Id: hal-03796134

<https://hal.science/hal-03796134v1>

Submitted on 4 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

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

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

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

10 (4) IRISA, University of Rennes 1, France

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

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

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

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

18 Corresponding authors:

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

20 Jérôme Idier, jerome.idier@ls2n.fr

21 Short title

22 Conformational space of IDPs Sic1 and pSic1

23 August 29, 2022

24 Abstract

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

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

40 the human proteome and, depending on the organism type, the overall percentage of amino
41 acids predicted to be disordered ranges from about 12% up to 50%.¹ In addition, the
42 conformational plasticity of the disordered regions of proteins allows them to interact with
43 numerous partners in the cell, as for example for the three intrinsically disordered domains of
44 the tumor protein P53.² This moonlighting³ behavior explains the strong impact of IDPs in
45 cellular signaling, regulation, and control, and the differences observed in their interactomes
46 with respect to globular proteins.⁴

47 Intrinsically disordered proteins represent a challenge for structural biology for several
48 reasons. In solution, the nuclear Overhauser effects measuring distance between hydrogens
49 is usually not available. On the other hand, crystallization processes are hampered by the
50 conformational disorder, or the variability of conformations in the crystal or in the elec-
51 tron cryogenic maps makes impossible the observation of electronic density for disordered
52 regions. Numerous approaches have been proposed⁵⁻⁸ for the calculation of protein confor-
53 mations, based on molecular dynamics or Monte Carlo simulations for generating molecular
54 conformations.

55 We propose here to explore a new approach for the exploration of the conformational
56 space of IDPs, based on a systematic enumeration of conformations in the frame of the dis-
57 tance geometry problem. We amend here our previous work introducing TAIiBP as a new
58 tool to investigate structural ensembles of IDPs in a systematic way, by predicting popu-
59 lations and consequently selecting pools of representative conformations. This approach,
60 initiated as the interval Branch-and-Prune (iBP) algorithm by Mucherino and coworkers,⁹
61 was adapted to the protein molecular modeling as threading-augmented interval Branch-and-
62 Prune (TAiBP).^{10,11} Based on distance geometry, TAIiBP, explores the entire conformational

63 space compatible with NMR chemical shifts retaining conformations that are most different
64 from one another yielding thus a diverse set of conformations to be analyzed further. This
65 is in contrast to Monte Carlo methods which are informed by force fields and explore the
66 part of the configurational space that is thermodynamically relevant in more detail. TAI BP
67 was shown recently¹² to allow the analysis of the conformational space of a tandem domain
68 of protein whirlin, in which a disordered linker induces a large orientation variability of two
69 PDZ domains.¹³ The application of TAI BP to the tandem domain was made possible by
70 the analysis of unprocessed output of the neural network TALOS-N,¹⁴ the Ramachandran
71 likelihood maps. Indeed, drawing boxes on the most probable regions of these maps, allowed
72 the determination of intervals on backbone angles, which serve as inputs for the TAI BP
73 algorithm. It should be noticed that the approach MERA has been developed¹⁵ for the
74 prediction of the ϕ , ψ distributions for IDPs.

75 In the present work, we apply TAI BP to a well-know example of IDP.^{16,17} The obtained
76 IDP conformations will be filtered and their relative populations determined by BioEn¹⁸
77 using SAXS data. In parallel, we propose an original method, RamaMix, to select the main
78 conformations, as well as their populations, according the Ramachandran likelihood maps
79 predicted by TALOS-N.¹⁴ The principle of RamaMix is to fit a bivariate, periodic, finite
80 mixture model to the output of TALOS-N. The N terminal fragment of the intrinsically
81 disordered protein Sic1, as well as its phosphorylated form pSic1, each one spanning 90
82 residues, will be studied.

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

86 78, and 82) in order to be degraded by the proteasome. Sic1 as well as pSic1 were shown^{16,19}
87 to contain significant amount of transient secondary structures.

88 The comparison of repeated runs of TAI_{BP} on Sic1 and pSic1 reveals a good repro-
89 ducibility of global conformational shape. Qualitatively similar but quantitatively different
90 populations are obtained either by fitting distinct SAXS curves or Ramachandran maps. The
91 sets of individual conformations selected from the fitting of various data are partially distinct,
92 but better convergence is observed for the profiles of local gyration radius. These profiles
93 could be proposed as a low resolution description of the IDP conformational space. **Depend-**
94 **ing on the way the TAI_{BP} conformations are generated, and on the processing method to**
95 **obtain the populations, different patterns of variations are observed for the resulting gyration**
96 **radius of Sic1 and pSic1.**

97 **Results**

98 **Enumeration of protein conformations**

99 The TALOS-N¹⁴ prediction was obtained using the chemical shifts measured for the nuclei
100 H α , HN, ¹⁵N, ¹³C α , ¹³C β of Sic1 and pSic1 residues, and was used to determine boxes of ϕ
101 and ψ values, giving the limits in which the conformations will be enumerated. Indeed, from
102 the NMR chemical shifts and the protein sequence information, the TALOS-N neural network
103 predicts the likelihood that a given residue n has backbone torsion angles that fall in any of
104 the 324 voxels, of $20^\circ \times 20^\circ$ each, that make up the Ramachandran map.¹⁴ Following the
105 approach proposed in Ref. 12, we define boxes (Figures S1-S4) using Ramachandran regions

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

108 **rev1mera** In order to probe the reliability of the (ϕ, ψ) boxes obtained from the TALOS-N
109 likelihood maps, these boxes were compared to the predictions performed using the approach
110 MERA,¹⁵ which predicts the residue-by-residue Ramachandran map distributions for disor-
111 dered proteins using short-range NOEs, chemical shifts, J couplings and spectral density
112 derived from the N¹⁵ relaxation measurement. As only chemical shifts were available for
113 Sic1 and pSic1, the MERA prediction was performed putting to zeros all other possible in-
114 puts. The MERA Ramachandran map distributions are plotted for all successful predicted
115 residues, along with the input boxes derived from the TALOS-N prediction (Figures S10 and
116 S11), showing a reasonable agreement between the two methods.

117 Two replicates of boxes were generated for Sic1 and pSic1, using threshold values of
118 0.01 and 0.011 on the Ramachandran probability maps as described in section “Extraction
119 of boxes from Ramachandran likelihood” in the Supplementary Material. Using these sets
120 of input boxes, five TAI_{BP} runs were performed, named Sic1¹, Sic1², pSic1¹, pSic1² and
121 pSic1³. **pSic1³** The run pSic1³ differs from the others by the procedure for selecting more
122 extended representative conformations after the SOM clustering, as described in the section
123 ”Clustering of generated conformations” in the Supplementary Information.

124 The two replicates of TAI_{BP} calculations introduced in the previous subsection were
125 based on similar numbers of fragments: 14 and 13 for Sic1¹ and Sic1², 17 for for pSic1¹
126 and pSic1² and 18 for pSic1³ (Table S1). The larger number of fragments used for pSic1
127 arises from the regions of residues 5-9, 33-37, 45-49, 69-73, 76-84 for which TALOS-N was
128 unable to give a prediction due to the phosphorylated residues and for which generic boxes

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

132 The boxes used as inputs for the TAIiBP runs (Figures S1-S4) are quite similar. The loop
133 region (positive ϕ) is slightly more populated for runs pSic1¹ and pSic1². For the iBP and
134 assembly steps forming the TAIiBP approach, the duplicate runs, marked in colors red and
135 green in Figure 1, produces parameter values similar in most of the protein sequence.

136 For the iBP steps, three parameters were compared (Figure 1, first and second lines)
137 along the residue number located at the middle of each fragment: the number of individual
138 iBP runs (N_{iBPrun}), the number of saved conformations ($N_{iBPconf}$) and the number of ob-
139 tained conformations after clustering ($N_{clustiBP}$). The three analyzed parameters are located
140 in similar ranges for all calculations. Nevertheless, N_{iBPrun} displays the largest observed
141 values (3888) around the positions of phosphorylated Threonines in agreement with the
142 larger generic boxes used in these protein regions (Table S2). Such increase is not observed
143 for phosphorylated Serines due to shorter fragments used in the region 50-90 (Table S1).
144 For every calculation, $N_{iBPconf}$ is smaller than 10^9 , which is the input given for the maxi-
145 mum number of solutions: all individual iBP trees have been thus completely parsed. The
146 $N_{iBPconf}$ profiles display smaller values, mostly in the range 10^6 - 10^7 , for all calculations in
147 the region of residues 60-90. At the contrary of $N_{iBPconf}$, the numbers of clustered confor-
148 mations ($N_{clustiBP}$) display relatively flat profiles for Sic1, but a decrease in the number of
149 conformations of pSic1 in the region of residues 60-90. This larger reduction of conforma-
150 tions due to the clustering is the sign that the conformations generated by iBP in the region
151 60-90 are more diverse in Sic1 than in pSic1. In all calculations, the C terminal fragments

152 which are smaller than the others (Table S1), display smaller N_{iBP} , $N_{iBPconf}$ and $N_{clustiBP}$.
153 The results obtained for the run pSic1³ (blue crosses) are quite similar to those of the run
154 pSic², which is not surprising as the fragment definition are the same, except around residues
155 40-60 (Table S1).

156 Three parameters are plotted (Figure 1, third and fourth lines) along the assembled
157 fragments: the number of conformations rejected due to C α atoms closer than 1Å ($N_{clashes}$),
158 the number of saved conformations (N_{saved}) and the number of clustered conformations
159 (N_{clust}). Looking at the relative ranges of values of $N_{clashes}$ and N_{saved} , between 10% and
160 15% of the assembled fragments are rejected due to the steric clashes. The profiles of N_{clust}
161 are different for Sic1 and pSic1, as the number of clustered conformations increases up to
162 the last fragment, whereas this number already starts to decrease in the region of residues
163 60-90 in pSic1. This effect can be put in parallel with the decrease of $N_{clustiBP}$ in the same
164 region during the iBP step. The last fragments of proteins have strong decreasing effects
165 on N_{clust} , due to their smaller size (Table S1) which induces probably less variability in the
166 generated conformations. Smaller numbers of clashes are mostly obtained for run pSic1³
167 (blue crosses), which is probably due to the larger extension of conformations. The number
168 of saved conformations is often larger than in other runs, which may be a consequence of the
169 smaller numbers of clashes. Unsurprisingly, the number of clustered conformations increases
170 along the number of saved conformations.

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

175 and final values of backbone angles produce values in the range 4.2-4.9° for ϕ and in the range
176 0.04-2.4° for ψ . Similarly, the average coordinate RMSD between the initial and final frames
177 of the refinement trajectories are 0.6 Å for the four runs Sic1¹, Sic1², pSic1¹ and pSic1². The
178 drift is larger for pSic1³, with backbone angle values in the ranges -24 to 6° for ϕ and -40 to
179 -1° for ψ , and an average coordinate RMSD of 0.7 Å. The conformations displaying potential
180 energy smaller than -50 kcal/mol for the runs Sic1¹ and Sic1² and smaller than -600 kcal/mol
181 for the runs pSic1¹ and pSic1², were selected for further analyses. This selection produces
182 sets of 98 (Sic1¹), 133 (Sic1²), 161 (pSic1¹), 121 (pSic1²) and 148 (pSic1³) conformations.

183 **Comparison of the conformations between duplicate TAI BP runs**

184 The distributions of gyration radii R_g and maximal diameters D_{max} (Figure 2 top) are
185 quite similar for the duplicate runs on Sic1 and pSic1. The global envelope of generated
186 conformation is thus reproducible between the replicated TAI BP runs. The distribution of
187 gyration radii R_g and maximal diameters D_{max} have been plotted in magenta for the run
188 pSic1³ to display the larger extension of the obtained conformations.

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

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

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

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

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

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

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

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

226 **Validation of the finite mixture model on synthetic data**

227 Once a set of conformations have been selected using TAIiBP, one needs to detect the con-
228 formations significantly populated and to evaluate their relative populations. Indeed, the
229 systematic enumeration along all possible combination of the ϕ/ψ boxes induces the gener-
230 ation of conformations spanning a space possibly larger than the conformations effectively
231 populated. The populations were determined, from one side, using BioEn¹⁸ on SAXS data,
232 and on the other side, using on the Ramachandran maps, a finite mixture model, RamaMix,
233 specially developed for this purpose. We first present in this section a validation of RamaMix
234 on synthetic data.

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

238 to investigate the effect of conformational superimposition on the population determination.
239 Corresponding populations were also chosen randomly (see caption of Figure S5). Noise levels
240 of 0.2, 1, 2, 3, 5 and 10 were added to the histogram obtained from the pseudo Ramachandran
241 map, the maximum value of the histogram being around 15. The starting points for each
242 RamaMix run was the ϕ_0 , ψ_0 values from the synthetic Ramachandran plot, and random
243 population values. During each RamaMix run, several upper limits were imposed to the
244 drift of the backbone angles during the optimization, with values of: 1° , 10° , 20° , 30° , 40°
245 and 50° . For each Ramachandran synthetic map, each noise level and each drifting limit
246 value, one hundred runs are performed producing sets of backbone angles (ϕ_0 and ψ_0) (Eq.
247 7), von Mises parameters (Eq. 8) (κ_1 , κ_2 and ρ) and populations γ_q (Eq. 2). Over the
248 12600 individual RamaMix runs, only 275 runs were terminated without convergence of the
249 optimization. Averages and standard deviations were calculated from the sets of obtained
250 parameters. The differences between the averaged and the input values, as well as the
251 standard deviations (Figure 3) are used to evaluate RamaMix.

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

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

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

265 The parameters describing the von Mises distribution (Figure 3G-L) display contrasted
266 results: the differences are larger for ρ than for κ_1 and κ_2 . For κ_1 and κ_2 , the standard
267 deviations are much larger than the differences whereas they are similar for ρ . The differences
268 between ρ and κ_1 and κ_2 , arise from the definition of these parameters (Eq. 8) in which ρ
269 occupies a different place than κ_1 and κ_2 .

270 **Determination of populations**

271 The TAI_{BP} conformations were fitted to the SAXS curves and Ramachandran probability
272 maps using BioEn¹⁸ and RamaMix.

273 The following sets of conformations were processed: the conformations obtained from
274 runs Sic1¹, Sic1², pSic1¹, pSic1² and pSic1³, as well as two mixed sets of conformations
275 obtained by pooling the conformations from pSic1¹ and pSic1³ and the conformations from
276 pSic1² and pSic1³. These mixed sets of conformations will be denoted pSic1¹³ and pSic1²³
277 and encompass respectively 309 and 269 conformations.

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

282 Most of the conformations selected only once, display populations smaller than 15%. But
283 the populations vary significantly from one analysis to another as for example for the confor-
284 mation 109 from the run Sic1² (Table 1B) which display populations of 26.6, 40.9 and 43.8%
285 for the three SAXS curve processing. Normalized χ^2 values smaller than one are found for
286 each calculation along with null final S_{KL} values, in agreement with the definition of S_{KL} as
287 the Kullback-Leibler divergence.^{18,20}

288 Tables 3 and S6 present the populations obtained by RamaMix from the fitting of the
289 Ramachandran probability maps on the same sets of conformations. The variations of back-
290 bone angles ϕ and ψ during the RamaMix optimization are smaller than 0.25° for ϕ and 0.1°
291 for ψ during all considered calculations. These variations are smaller for pSic1³ with 0.12°
292 and 0.03° for ϕ et ψ , and even smaller for the mixed pools of conformations with 0.06 and
293 0.02°. Among six of the seven sets of TAI_{BP} conformations, conformations (marked in bold)
294 already repeatedly selected by BioEn, were also selected by RamaMix (Tables 1, 2 and S5).

295 Similarly to the populations obtained by BioEn between the different SAXS data, the
296 populations found using RamaMix are quite different than the ones determined by BioEn.
297 Another difference between BioEn and RamaMix processing is the smaller number of con-
298 formations selected by RamaMix, it can arise from the essential difference between the data,
299 as the SAXS curves describe a global picture of the conformations whereas the Ramachan-
300 dran maps give a local information. A smaller number of conformations are selected from
301 the sets where more extended conformations were included: this may be due to the impor-
302 tant conformational drift induced by the systematic choice of extended conformations during
303 the clustering step (Supplementary information section "Clustering of generated conforma-
304 tions").

305 In order to compare the conformations selected by BioEn on the three SAXS curves,
306 several curves superimpositions have been realized. The superimposition of SAXS curves
307 reconstructed from the conformations selected from Tables 1 and 2 to the corresponding
308 fitted SAXS curves (Figure S6) displays a reasonable agreement with χ^2 in the range 0.6-
309 2.06. These values are larger than the ones given in the Tables 1 and 2, due to the fact
310 that conformations displaying populations smaller than 1% have been removed. Besides, a
311 comparison of all sets of BioEn conformations with all SAXS curves (Table S3) reveals that
312 the conformations and populations determined from the fit of one SAXS curve display χ^2
313 values with another SAXS curve going up to 4.42. The variability between the three SAXS
314 curves induces thus a drift between conformations and populations selected from the fit of
315 each curve.

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

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

328 The comparison of profiles P_q (Eq. 1) between conformations selected by RamaMix (Figure
329 4, right column) reveals two features. When few conformations have been selected (as for
330 Sic1² and pSic1²), the distances between their profiles P_q are larger than 8 Å. When more
331 conformations are selected (as for Sic1¹ and pSic1¹), profile distances smaller than 8 Å are
332 observed. The small P_q distances reveal a certain convergence of the profiles P_q .

333 The comparison of conformations selected by BioEn and RamaMix as well as the com-
334 parison between conformations selected from the fit of the various SAXS curves is displayed
335 in Figures S8 and S9. A close inspection of these distance matrices for BioEn conformations
336 (Figure S8) shows that, if one excludes the conformations populated less than 10%, there
337 are only three conformations displaying profile distances larger than 8 Å and selected in two
338 distinct BioEn runs: (i) for Sic1², the conformation 106 selected on the SAXS curve BioEn1
339 compared to the conformations selected on the two other SAXS curves; (ii) for pSic1², the
340 conformation 74, selected in the runs BioEn1 and BioEn2, and compared to the conforma-
341 tions selected from the run BioEn3; (iii) for pSic1², the conformation 139 selected from the
342 run BioEn3, and compared to the conformations from the run BioEn1. Overall, most of the
343 conformations populated more than 10% from the fitting of different SAXS curves display
344 profile distances smaller than 8 Å, supporting a convergence of the profiles in the different
345 fits.

346 On the other hand, the comparison between BioEn and RamaMix fitting (Figure S9)
347 displays contrasted behaviors between the duplicated TAiBP runs. For Sic1² and pSic1², all
348 RamaMix conformations display profiles closer than 8 Å to the profiles of BioEn conforma-
349 tions. For pSic1¹, this is also the case for three RamaMix conformations (16, 98, 101) over
350 five. For Sic1¹, only the conformation 79 displays profile distances smaller than 8 Å for the

351 three comparisons.

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

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

365 **Comparison with PED conformations and link with biological ac-** 366 **tivity**

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

370 **rgyr** The values of the resulting gyration radii were calculated (Table 4) from the popula-
371 tions determined by BioEn and RamaMix, and using the individual gyration radii of selected

372 conformations. Globally, the resulting gyration radii display orders of values agreeing with
373 the measurements reported in Figure 2E of Ref. 17. For the conformations extracted from
374 the data-sets Sic1¹ and Sic1², the resulting gyration radii agree with the measurement of 3.0
375 \pm 4.1 Å given in Figure 2E of Ref. 17. But, for pSic1¹ and pSic1², the resulting gyration
376 radii are smaller than the measurements of Ref. 17: this is particularly true for the BioEn
377 processing whereas the RamaMix processing displays values closer to those of Gomes et al.¹⁷
378 On the more extended conformations (pSic1³), all resulting gyration radii are significantly
379 closer to the Gomes et al¹⁷ measurements, for BioEn and RamaMix processing. Pooling
380 pSic1³ with the conformations of pSic1¹ or pSic1² (sets pSic1¹³ and pSic1²³) produces dif-
381 ferent effects for BioEn and RamaMix processing. For these mixed data-sets, the resulting
382 gyration radii obtained from the BioEn processing (range 27.2-28.1 Å) decrease to reach a
383 level just slightly larger than the one obtained for pSic1¹ and pSic1² (range 26.1-27.9 Å). At
384 the contrary, the gyration radii obtained by RamaMix processing are the same than the ones
385 obtained for pSic1³. Overall, the BioEn processing is more sensitive than RamaMix to the
386 presence of conformation with lower gyration radii. The discrepancy of the results obtained
387 here on pSic1 with those shown in Ref. 17 arises in part from the tendency to obtain smaller
388 gyration radii by processing of the whole SAXS curve with respect to the larger gyration
389 radii obtained by using the Guinier approximation within the low-q region of the SAXS data.

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

395 `rev3hbond` The presence of phosphorylated residues decreases obviously the global charge
396 of pSic1 with respect to Sic1. It was pointed out that the induced variation in long-range
397 electrostatic interactions plays a role in the electrostatic interaction of pSic1 with its target
398 Cdc14.²² But, the variations of charges gives also various opportunities for the formation of
399 hydrogen bonds, which were analyzed for the whole set of conformations from the TAIiBP
400 runs as well as for the PED sets of conformations. All PED conformations were submitted
401 to the same refinement than the one used on TAIiBP conformations described in the section
402 "Molecular dynamics refinement in implicit solvent" of the Supplementary Material, using
403 positional restraints on protein backbone atoms with a constant force of 50 kcal/mol. The
404 cumulative variations of ψ and ϕ angles during the refinement were in the range 0.8-1.2° for
405 ϕ and in the range 0.3-0.8° for ψ , and the coordinate RMSD around 0.1 Å. All hydrogen
406 bonds were detected on the refined PED conformations as well as on the TAIiBP conforma-
407 tions. Cumulative contact maps (Figure S12) display these hydrogen bonds, according to
408 the involved residues, the hydrogen bonds involving phosphorylated residues being colored
409 in magenta. The inspection of these contact maps reveals that the PED and TAIiBP confor-
410 mations display distinct tendencies. Long range hydrogen bonds involving phosphorylated
411 residues are more present in the less extended set of TAIiBP conformations pSic¹ and pSic².
412 On the other hand, the conformation set PED161 of pSic1 displays the largest number of
413 long-range hydrogen bonds involving the sidechains of phosphorylated residues. Thus, the
414 presence of phosphorylated residues can induce the appearance of long-range hydrogen bonds
415 whatever is the variation of resulting gyration radius.

416 Discussion

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

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

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

438 should provide geometrical restraints allowing a more precise exploration of the conforma-
439 tional space. [labellreviewer1](#) The profiles are closer between the conformations selected by
440 the various fits of SAXS curves, than between the conformations selected by BioEn and
441 RamaMix. This is expected as the various fits of SAXS curves use an homogeneous infor-
442 mation. More surprisingly, similar profiles are observed between conformations selected by
443 RamaMix and BioEn, for the runs Sic1² and pSic1², and for many conformations of the runs
444 Sic1¹ and pSic1¹s.

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

452 [rgyr](#) The comparison of the resulting gyration radii obtained from the BioEn and Ra-
453 maMix processing with the values measured in Ref. 17 showed (Table 4) that various ranges
454 of gyration radii are obtained, depending on the clustering procedure in TAI_{BP}, as well as
455 on the method for SAXS processing. In particular, the processing of the whole SAXS curve
456 with BioEn displays a tendency to underestimate the gyration value with respect to the
457 processing of the Guinier curve. The determination of populations from the Ramachandran
458 probability maps, using RamaMix, seems to be less prone to the underestimation of the
459 gyration radius.

460 [song](#) The discrepancy between resulting gyration radii obtained by processing the whole

461 SAXS curve (BioEn) or restricting the analysis to the low-q region (Guinier approximation¹⁷)
462 agrees with independent calculations performed using coarse-grained protein model,²⁶ in
463 which various distribution of gyration values produce very similar SAXS spectra (Figure 10a
464 of²⁶) or different disordered ensemble produce similar Kratky plots (Figure 13 of²⁶).

465 **Methods**

466 **Origins of the data**

467 Three sets of conformations for Sic1 and pSic1 were available from the Protein Ensemble
468 Database (PED) proteinensemble.org:²¹ PED159 and PED160 for pSic1 and PED161 for
469 Sic1.¹⁷ The residue numbering used here is the one proposed in the PED. The NMR chemical
470 shifts were downloaded from the Biological Magnetic Resonance Data Bank (BMRB)²⁷ as
471 entries: 16657 for Sic1¹⁷ and 16659 for pSic1.¹⁹ The SAXS data-sets recorded as triplicate
472 sets in the conditions described in [Ref.](#)¹⁷ were provided by Tanja Mittag.

473 **Enumeration of conformations using TAIiBP**

474 The protein conformations have been enumerated using the recently proposed TAIiBP ap-
475 proach,^{10–12} which generalizes the interval branch-and-prune (iBP) algorithm^{9,28–31} so as to
476 overcome the combinatorial barrier arising from the enormous space of IDP conformations.³²
477 TAIiBP is composed of two steps: (i) the enumeration of conformations for peptide fragments
478 (Table S1) spanning the studied protein using individual iBP calculations; (ii) the enumer-
479 ation of Sic1 and pSic1 conformations by systematic assembly of fragment conformations in

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

481 The boxes of backbone angles ϕ and ψ used as inputs for the iBP step were determined
482 from the Ramachandran likelihood maps predicted by TALOS-N¹⁴ (see section “Extraction of
483 boxes from Ramachandran likelihood maps” and Figures S1-S4 in Supplementary Material).
484 The ϕ/ψ boxes were systematically combined by permutation to prepare individual iBP
485 calculations as in Ref 11. The enumeration of conformations is realized by the building of
486 a tree, each node of the tree corresponding to an atomic position. The tree building allows
487 the enumeration of the various possibilities for atom positions (branching step) whereas
488 additional geometric information is used to accept or reject a newly built branch (pruning
489 step). As the angles ϕ and ψ are straightforwardly related to distances between atoms C
490 and N of residues successive in the sequence,^{11,12} the discretization of intervals of these
491 angles is used in the branching step. In the iBP step, the pruning was applied by preventing
492 atoms to be closer than the sum of their van der Waals radii and by checking that the
493 improper angle values are correct. In addition, each solution displaying a coordinate root-
494 mean-square deviation (RMSD) smaller than 2 Å with the previously stored solution, is
495 rejected. The details of the iBP step calculation are described in the section “Enumeration
496 of conformations” of the Supplementary Material.

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

503 peptide and to continue with the second one. The assembled conformations in which $C\alpha$
504 atoms closer than 1 Å are observed, were pruned from the calculation. The fragment as-
505 sembly was implemented using python scripting based on the MDAnalysis^{34,35} and numpy
506 1.7.1³⁶ packages.

507 To scale down the combinatorial explosion of the calculation, a clustering approach based
508 on **Self-Organizing Maps (SOM)**³⁷⁻⁴⁰ was systematically applied to the generated sets of con-
509 formations larger than 100 during the iBP and assembly steps. The details of this approach
510 are described in the section “Clustering of generated conformations” of the Supplementary
511 Material.

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

515 **Determining the population from Ramachandran maps**

516 The approach RamaMix, based on a finite mixture model, was designed to determine the pop-
517 ulations of conformations by fitting on the Ramachandran probability maps. The setting-up
518 of this approach is based on the hypothesis that the likelihood maps describing the likeli-
519 hood of the TALOS-N prediction¹⁴ can be transformed by normalization into the probability
520 density of the presence of ϕ and ψ values in the set of conformations populated in solution.

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

523 mined on each conformation q :

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

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

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

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

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

534 Calculations detailed in the Supplementary material (sections "Determination of the popu-
 535 lations from the Ramachandran maps" and "Maximum likelihood estimation for bivariate
 536 sine mixtures") show that using the Kullback-Leibler divergence is equivalent to the maxi-
 537 mization of the log-likelihood of the data:⁴¹

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

538 For the sake of clarity, let us first introduce a standard, non-periodized Gaussian density
 539 $p_q^n(\phi, \psi)$ for the residue n in conformation q :

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

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

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

543 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
 544 n in the conformation q and C_q^n is the corresponding covariance. The software IMOD⁴²
 545 was used for determining the full Hessian (N, N) (N is the total number of residues in the
 546 protein) matrix H_q along the backbone dihedral angles. The Hessian matrix is then inverted
 547 to produce: $C_q = H_q^{-1}$. The covariance matrix C_q^n of the angles ϕ and ψ of the considered
 548 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
 549 of this matrix $[C_q^n]^{-1}$ is used in Eq. 6.

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

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

555 with

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

556 $\kappa_1, \kappa_2 \geq 0$ and $\lambda^2 < \kappa_1 \kappa_2$. According to Ref. 46, the integration constant T is expressed as

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

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

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

559 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
 560 bivariate Gaussian where $\rho \in (-1, 1)$ denotes the normalized correlation coefficient between
 561 the two components of the bivariate Gaussian:

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

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

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

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

568 A well-known local optimization scheme to identify finite mixture models by maximum
 569 likelihood is the Expectation-Maximization (EM) algorithm.^{50,51} Unfortunately, the M step
 570 of the EM has no analytical expression in the case of mixtures of bivariate Von-Mises densi-
 571 ties. Therefore, we have performed local optimization based on L-BFGS-B⁵² instead, given
 572 that both the likelihood and its gradient can be evaluated efficiently, and that some param-
 573 eters are subject to box constraints. The implementation details and equations are given in
 574 the sections "Determination of the populations from the Ramachandran maps" and "Maxi-
 575 mum likelihood estimation for bivariate sine mixtures" of the Supplementary Material.

576 By optimization of the log-likelihood, the RamaMix approach will thus produce the Q
577 normalized populations γ_q , the $Q \times N$ couples of backbone angles ϕ_q^n and ψ_q^n , as well as the
578 $Q \times N$ triplets $(\kappa_1^n, \kappa_2, \rho_q^n)$ describing the von Mises distributions. The calculations were
579 performed starting from the ϕ and ψ values observed in the set of TAI BP conformations,
580 complemented by von Mises parameters allowing us to approximate the Gaussian distribu-
581 tions determined by IMOD. Moreover, the variation of ϕ and ψ values was limited by a
582 threshold of 15° during the optimization in order to avoid inappropriate drift.

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

585 **Determining the populations from SAXS data**

586 The software BioEn 0.1.1¹⁸ was used in order to determine the populations from SAXS data.
587 On each considered conformation, theoretical SAXS curves were calculated using CRY SOL⁵³
588 available in the package ATSAS 3.0.3⁵⁴ with 847 points, a maximum scattering vector of 0.503
589 nm^{-1} and a maximum order of harmonics of 18. A 1D cubic interpolation⁵⁵ was used to
590 obtain the theoretical SAXS values at the same sets of scattering vectors q than the ones
591 at which the experimental SAXS curve was recorded.

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

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

599 **Acknowledgments**

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

604 **Availability of Data and Materials**

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

607 **References**

- 608 [1] Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically dis-
609 ordered protein regions. *Annu Rev Biochem* **83**, 553–584 (2014).
- 610 [2] Kumar, A., Kumar, P., Kumari, S., Uversky, V. N. & Giri, R. Folding and structural
611 polymorphism of p53 C-terminal domain: One peptide with many conformations. *Arch*
612 *Biochem Biophys* **684**, 108342 (2020).

- 613 [3] Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic Protein
614 Interaction Networks and New Structural Paradigms in Signaling. *Chem Rev* **116**,
615 6424–6462 (2016).
- 616 [4] Teilum, K., Olsen, J. G. & Kragelund, B. B. On the specificity of protein–protein
617 interactions in the context of disorder. *Biochemical Journal* **478**, 2035–2050 (2021).
- 618 [5] Bernadó, P. *et al.* A structural model for unfolded proteins from residual dipolar cou-
619 plings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* **102**, 17002–17007
620 (2005).
- 621 [6] Allison, J. R., Varnai, P., Dobson, C. M. & Vendruscolo, M. Determination of the
622 free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance
623 measurements. *J Am Chem Soc* **131**, 18314–18326 (2009).
- 624 [7] Fisher, C. K., Huang, A. & Stultz, C. M. Modeling intrinsically disordered proteins
625 with bayesian statistics. *J Am Chem Soc* **132**, 14919–14927 (2010).
- 626 [8] Krzeminski, M., Marsh, J. A., Neale, C., Choy, W. Y. & Forman-Kay, J. D. Character-
627 ization of disordered proteins with ENSEMBLE. *Bioinformatics* **29**, 398–399 (2013).
- 628 [9] Lavor, C., Liberti, L. & Mucherino, A. The interval Branch-and-Prune algorithm for
629 the discretizable molecular distance geometry problem with inexact distances. *J Glob*
630 *Optim* **56**, 855–871 (2013).
- 631 [10] Worley, B. *et al.* Tuning interval Branch-and-Prune for protein structure determination.
632 *Journal of Global Optimization* **72**, 109–127 (2018).

- 633 [11] Malliavin, T. E., Mucherino, A., Lavor, C. & Liberti, L. Systematic Exploration of
634 Protein Conformational Space Using a Distance Geometry Approach. *J Chem Inf Model*
635 **59**, 4486–4503 (2019).
- 636 [12] Malliavin, T. E. Tandem domain structure determination based on a systematic enu-
637 meration of conformations. *Sci Rep* **11**, 16925 (2021).
- 638 [13] Delhommel, F. *et al.* Structural Characterization of Whirlin Reveals an Unexpected
639 and Dynamic Supramodule Conformation of Its PDZ Tandem. *Structure* **25**, 1645–
640 1656 (2017).
- 641 [14] Shen, Y. & Bax, A. Protein structural information derived from NMR chemical shift
642 with the neural network program TALOS-N. *Methods Mol Biol* **1260**, 17–32 (2015).
- 643 [15] Mantsyzov, A. B. *et al.* A maximum entropy approach to the study of residue-specific
644 backbone angle distributions in α -synuclein, an intrinsically disordered protein. *Protein*
645 *Sci* **23**, 1275–1290 (2014).
- 646 [16] Mittag, T. *et al.* Structure/function implications in a dynamic complex of the intrinsi-
647 cally disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**,
648 494–506 (2010).
- 649 [17] Gomes, G. W. *et al.* Conformational Ensembles of an Intrinsically Disordered Protein
650 Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc* **142**, 15697–
651 15710 (2020).
- 652 [18] Köfinger, J. *et al.* Efficient Ensemble Refinement by Reweighting. *J Chem Theory*
653 *Comput* **15**, 3390–3401 (2019).

- 654 [19] Mittag, T. *et al.* Dynamic equilibrium engagement of a polyvalent ligand with a single-
655 site receptor. *Proc Natl Acad Sci U S A* **105**, 17772–17777 (2008).
- 656 [20] Różycki, B., Kim, Y. C. & Hummer, G. SAXS ensemble refinement of ESCRT-III
657 CHMP3 conformational transitions. *Structure* **19**, 109–116 (2011).
- 658 [21] Lazar, T. *et al.* PED in 2021: a major update of the protein ensemble database for
659 intrinsically disordered proteins. *Nucleic Acids Res* **49**, D404–D411 (2021).
- 660 [22] Borg, M. *et al.* Polyelectrostatic interactions of disordered ligands suggest a physical
661 basis for ultrasensitivity. *Proc Natl Acad Sci U S A* **104**, 9650–9655 (2007).
- 662 [23] Bernadó, P. *et al.* A structural model for unfolded proteins from residual dipolar cou-
663 plings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* **102**, 17002–17007
664 (2005).
- 665 [24] Ozenne, V. *et al.* Flexible-meccano: a tool for the generation of explicit ensemble
666 descriptions of intrinsically disordered proteins and their associated experimental ob-
667 servables. *Bioinformatics* **28**, 1463–1470 (2012).
- 668 [25] Bondarenko, V. *et al.* Structures of highly flexible intracellular domain of human $\alpha 7$
669 nicotinic acetylcholine receptor. *Nat Commun* **13**, 793 (2022).
- 670 [26] Song, J., Li, J. & Chan, H. S. Small-Angle X-ray Scattering Signatures of Conforma-
671 tional Heterogeneity and Homogeneity of Disordered Protein Ensembles. *J Phys Chem*
672 *B* **125**, 6451–6478 (2021).
- 673 [27] Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res* **36**, D402–408 (2008).

- 674 [28] Lavor, C., Liberti, L., Maculan, N. & Mucherino, A. The Discretizable Molecular
675 Distance Geometry Problem. *Computational Optimization and Applications* **52**, 115–
676 146 (2012).
- 677 [29] Liberti, L., Lavor, C. & Mucherino, A. The discretizable molecular distance geometry
678 problem seems easier on proteins. *Distance Geometry: Theory, Methods and Applica-*
679 *tions. Mucherino, Lavor, Liberti, Maculan (eds.)* 47–60 (2014).
- 680 [30] Liberti, L., Lavor, C., Maculan, N. & Mucherino, A. Euclidean Distance Geometry and
681 Applications. *SIAM Rev* **56**, 3–69 (2014).
- 682 [31] Lavor, C., Alves, R., Figueiredo, W., Petraglia, A. & Maculan, N. Clifford Algebra and
683 the Discretizable Molecular Distance Geometry Problem. *Adv. Appl. Clifford Algebras*
684 **25**, 925–942 (2015).
- 685 [32] Levinthal, C. Are there pathways for protein folding? *J Chem Phys* **65**, 44–45 (1968).
- 686 [33] Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. & Baker, D. Generalized fragment
687 picking in Rosetta: design, protocols and applications. *PLoS One* **6**, e23294 (2011).
- 688 [34] Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAAnalysis: a
689 toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–
690 2327 (2011).
- 691 [35] Richard J. Gowers *et al.* MDAAnalysis: A Python Package for the Rapid Analysis of
692 Molecular Dynamics Simulations. In *Proceedings of the 15th Python in Science Confer-*
693 *ence*, 98–105 (2016).

- 694 [36] Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
695 URL <https://doi.org/10.1038/s41586-020-2649-2>.
- 696 [37] Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol Cybern*
697 **43**, 59–69 (1982).
- 698 [38] Kohonen, T. Self-organizing maps. *Springer Series in Information Sciences, Heidelberg,*
699 *Germany.* (2001).
- 700 [39] Miri, L. *et al.* Stabilization of the integrase-DNA complex by Mg²⁺ ions and prediction
701 of key residues for binding HIV-1 integrase inhibitors. *Proteins* **82**, 466–478 (2014).
- 702 [40] Bouvier, G. *et al.* Functional motions modulating VanA ligand binding unraveled by
703 self-organizing maps. *J Chem Inf Model* **54**, 289–301 (2014).
- 704 [41] Lehmann, E. L. & Casella, G. *Theory of point estimation.* Springer Texts in Statistics
705 (Springer-Verlag, New York, NY, 1998), 2nd edn.
- 706 [42] Lopéz-Blanco, J. R., Garzón, J. I. & Chacón, P. iMod: multipurpose normal mode
707 analysis in internal coordinates. *Bioinformatics* **27**, 2843–2850 (2011).
- 708 [43] Wako, H. & Endo, S. Normal mode analysis based on an elastic network model for
709 biomolecules in the Protein Data Bank, which uses dihedral angles as independent
710 variables. *Comput Biol Chem* **44**, 22–30 (2013).
- 711 [44] Na, H. & Song, G. Bridging between normal mode analysis and elastic network models.
712 *Proteins* **82**, 2157–2168 (2014).

- 713 [45] Tirion, M. M. & ben Avraham, D. Atomic torsional modal analysis for high-resolution
714 proteins. *Phys Rev E Stat Nonlin Soft Matter Phys* **91**, 032712 (2015).
- 715 [46] Singh, H., Hnizdo, V. & Demchuk, E. Probabilistic model for two dependent circular
716 variables. *Biometrika* **89**, 719–723 (2002).
- 717 [47] Mardia, K. V., Hughes, G., Taylor, C. C. & Singh, H. A multivariate von Mises distri-
718 bution with applications to bioinformatics. *Canadian Journal of Statistics* **36**, 99–109
719 (2008).
- 720 [48] Boomsma, W. *et al.* A generative, probabilistic model of local protein structure. *Proc*
721 *Natl Acad Sci U S A* **105**, 8932–8937 (2008).
- 722 [49] Clenshaw, C. Chebyshev series for mathematical functions. *NPL Mathematical Tables*
723 **5** (1962).
- 724 [50] McLachlan, G. J. & Krishnan, T. *The EM Algorithm and Extensions*. Wiley series in
725 probability and statistics (John Wiley and Sons, Inc., 1997).
- 726 [51] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and*
727 *Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).
- 728 [52] Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A limited memory algorithm for bound
729 constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208 (1995).
- 730 [53] Svergun, D. I., Barberato, C. & Koch, M. CRY SOL - a Program to Evaluate X-ray
731 Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl.*
732 *Cryst.* **28**, 768–773 (1995).

- 733 [54] Manalastas-Cantos, K. *et al.* ATLAS 3.0: expanded functionality and new tools for
734 small-angle scattering data analysis. *J Appl Crystallogr* **54**, 343–355 (2021).
- 735 [55] Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in
736 Python. *Nature Methods* **17**, 261–272 (2020).
- 737 [56] Galassi, M. *GNU Scientific Library Reference Manual (3rd Ed.)* (Network Theory Ltd.,
738 2009).

740 Figure 1

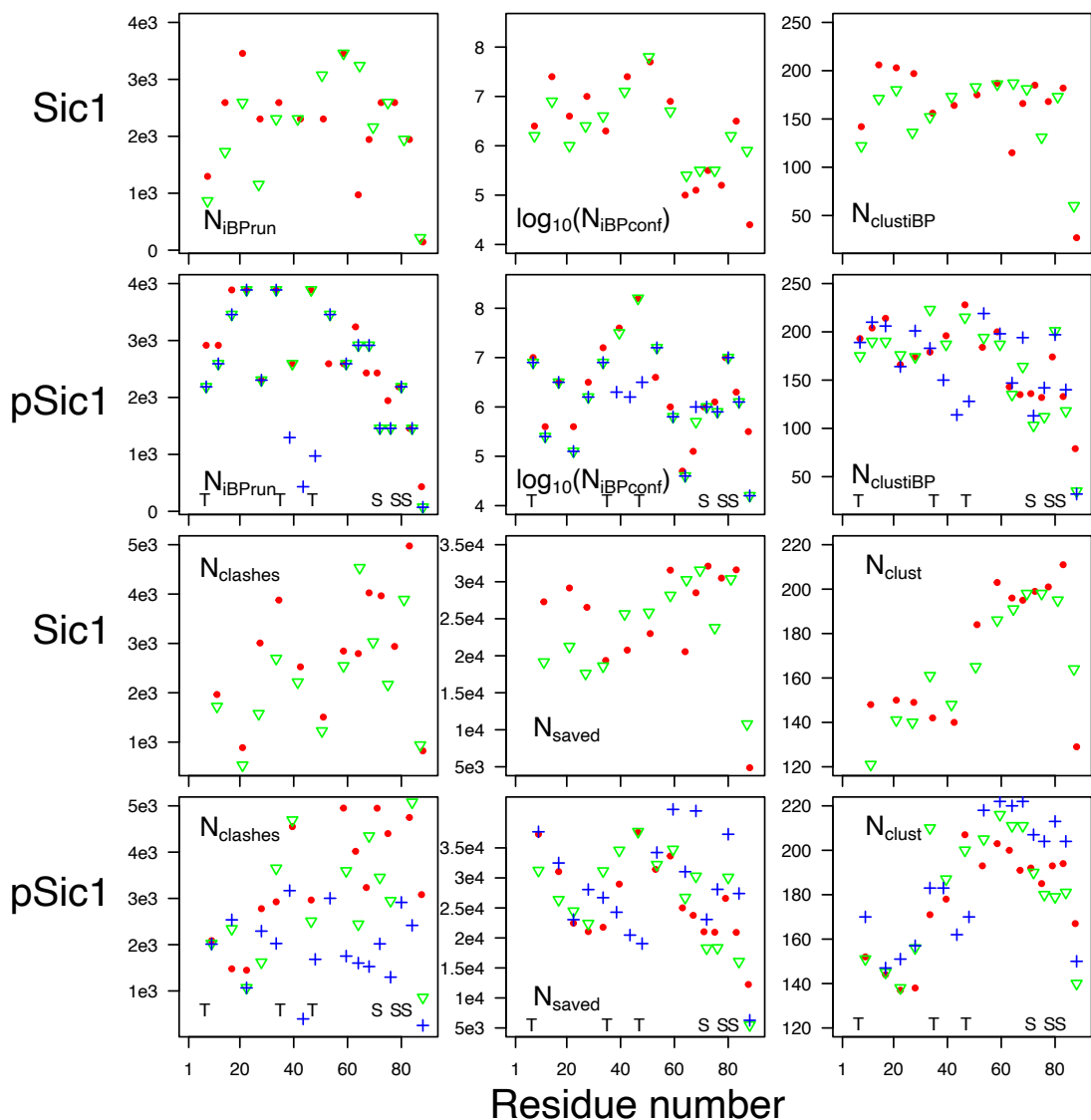


Figure 1: Parameters of the iBP and assembly steps of the TAiBP procedure. The signs red and green correspond respectively to the duplicated runs in which thresholds of 0.01 and 0.011 have been applied on the probability Ramachandran map. **The blue crosses correspond to the run pSic1³ producing more 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).

Figure 2

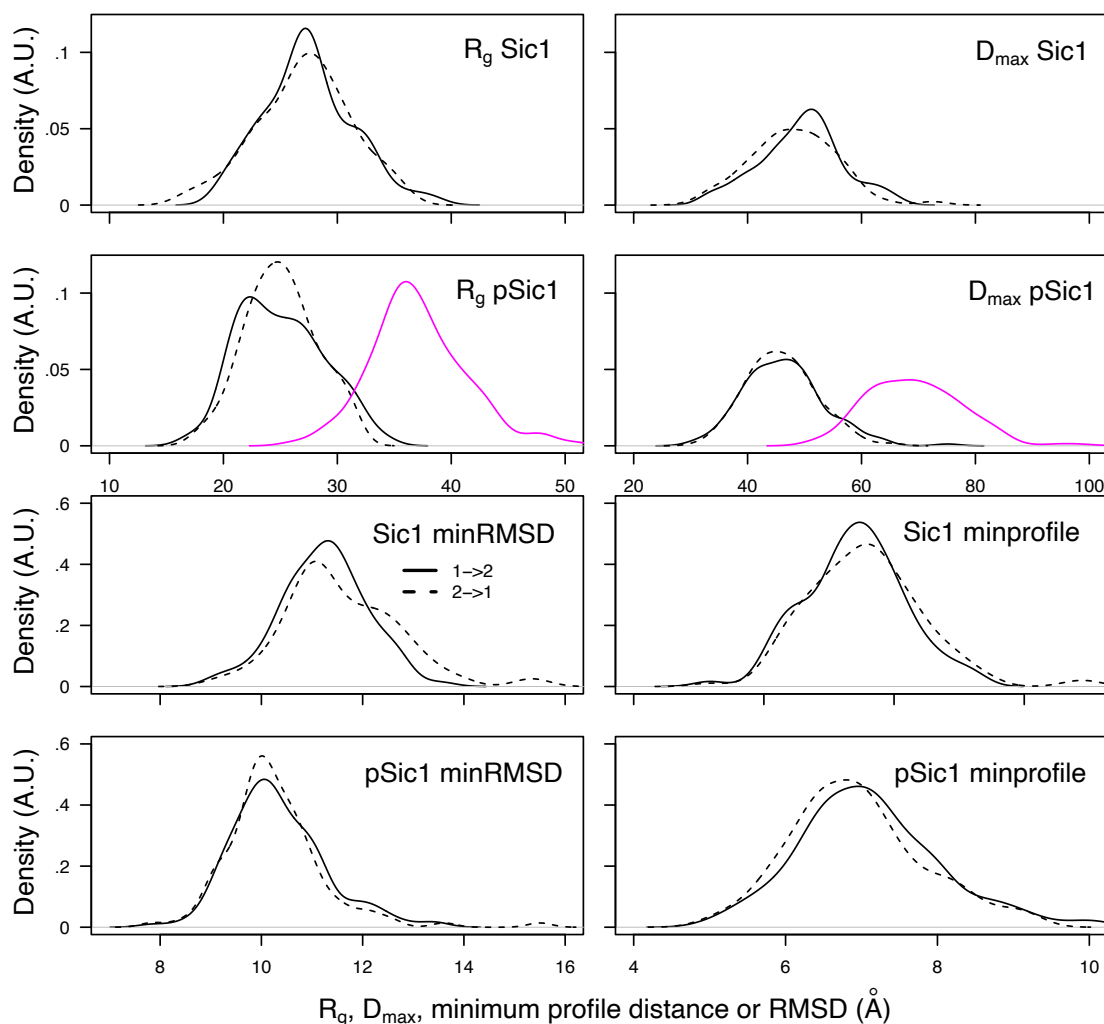


Figure 2: **Four panels on top**: Distribution of the gyration radii R_g and of maximal diameters D_{max} values in the two sets of TAiBP obtained during the first runs Sic1¹ and pSic1¹ (solid line) and the second runs Sic1² and pSic1² (dashed line) runs. **The R_g and D_{max} distribution obtained for the run pSic1³ 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 Sic1¹ and Sic1² and for pSic1¹ and pSic1². full line: first run with respect to the second one, dashed line: second run with respect to the first one.

Figure 3

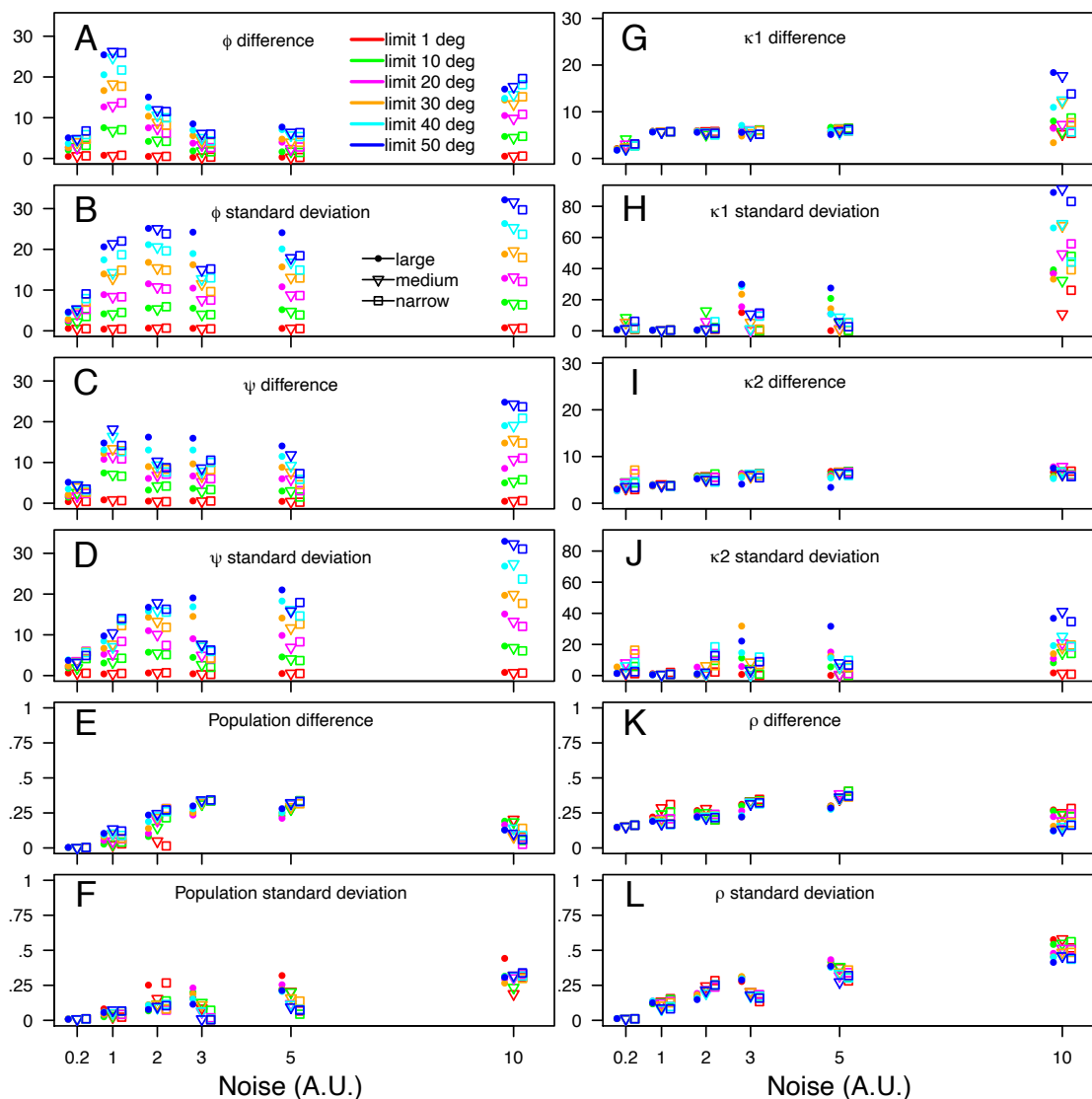


Figure 3: Efficiency of RamaMix for determining the ϕ_0 , ψ_0 positions (A-D: Eq. 7), the von Mises shape parameters κ_1 , κ_2 and ρ (G-L: Eq. 8), and the populations γ_q (E-F: Eq. 2) using synthetic data and various noise levels described in Figure S5. The results obtained for large, medium and narrow scattered synthetic Ramachandran maps are drawn as bullets, triangles and squares.

Figure 4

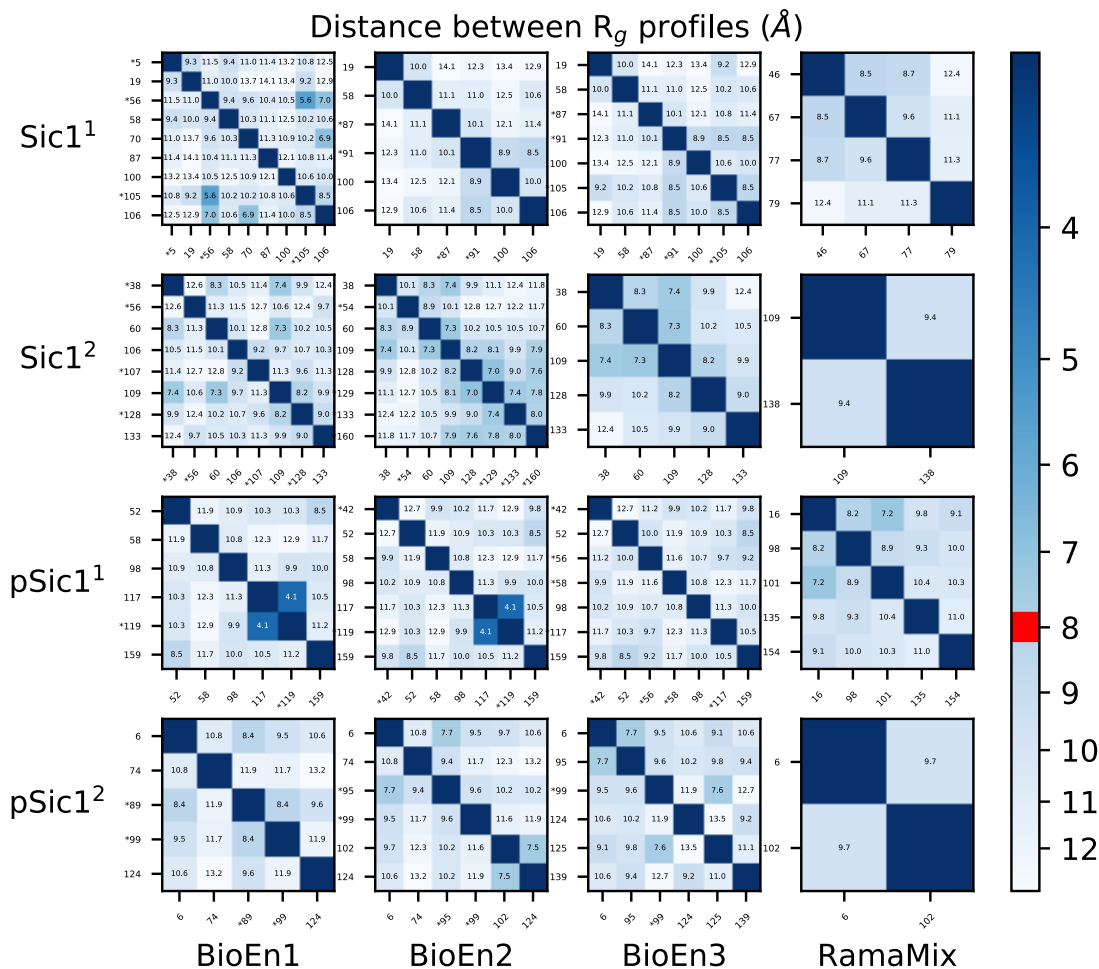


Figure 4: Distances between the profiles P_q (Eq. 1) of local gyration radii between the conformations selected from the fit of SAXS curves (BioEn1, BioEn2, BioEn3) or Ramachandran maps (RamaMix). The conformations for which populations smaller than 10% were calculated, are labeled with an asterisk.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

Figure 5

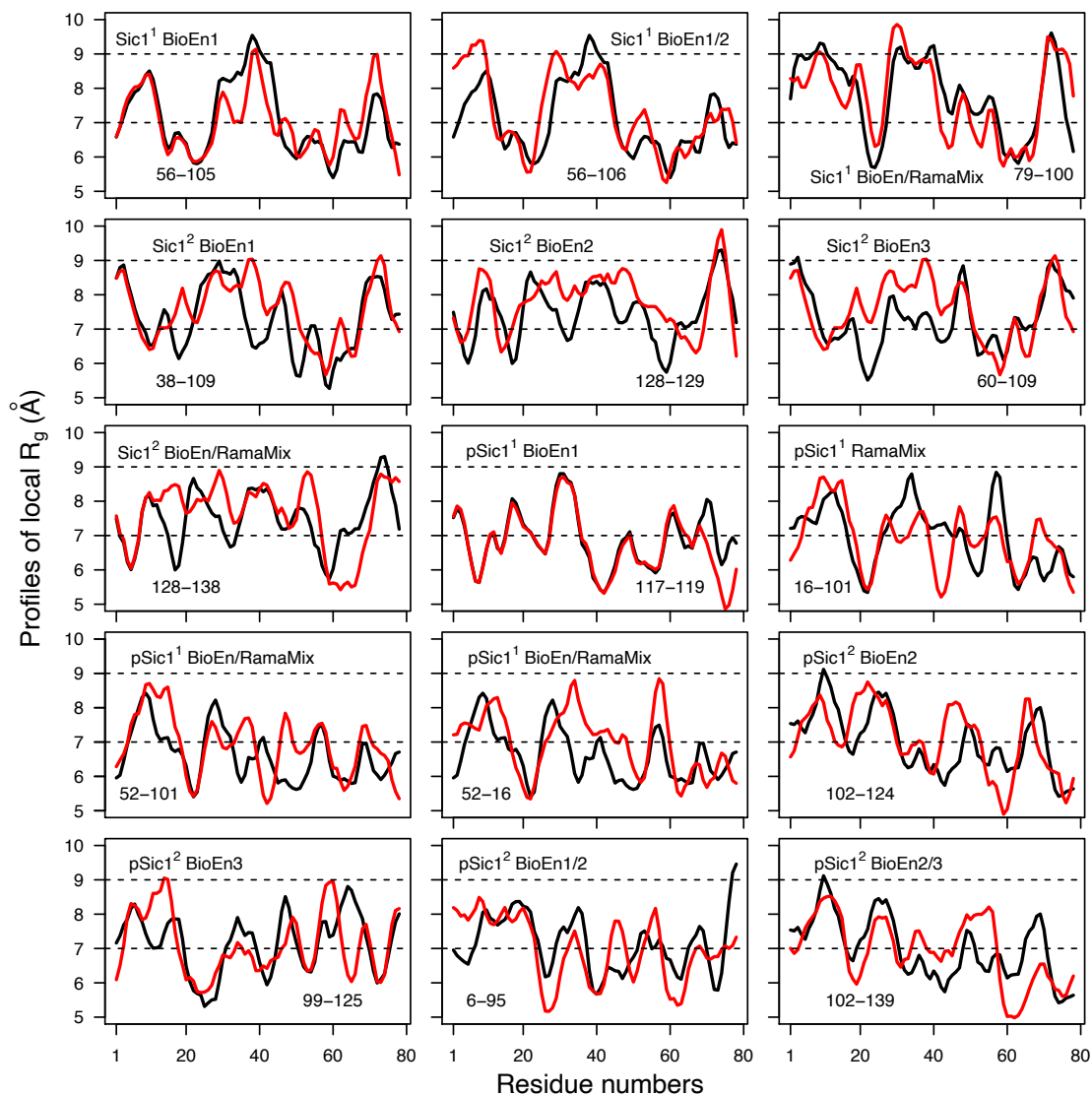


Figure 5: Superposition of profiles P_q (Eq. 1) displaying distances smaller than 8 Å extracted from Figures 4 and S8, S9. The name of the run (Sic1¹, Sic1², pSic1¹, pSic1²) is given, along with the name of the considered fits (BioEn1, BioEn2, BioEn3, RamaMix, RamaMix/BioEn) and the conformations numbers. 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.

Figure 6

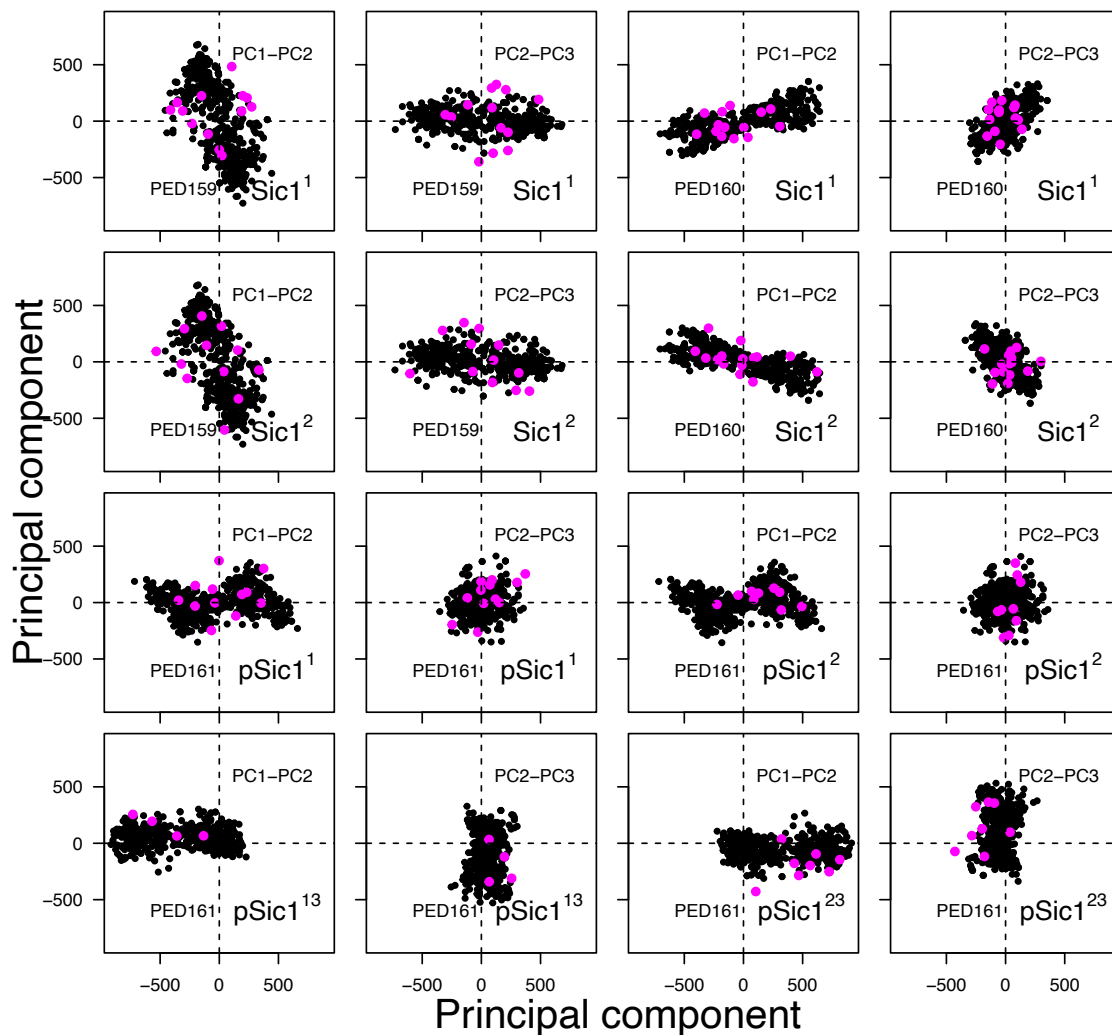


Figure 6: Projections of the Sic1 and pSic1 conformations along the three largest components of their principal component analysis (PCA). On these projections, the TAIbP conformations selected by BioEn or RamaMix are colored in magenta and the conformations stored in PED²¹ are colored in black.

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 χ^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 χ^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 0.1.1¹⁸ on the three sets of SAXS curves. The conformations were generated by the runs Sic1¹ and Sic1². 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 χ^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 χ^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 0.1.1¹⁸ on the three sets of SAXS curves. The conformations were generated by the runs pSic1¹ and pSic1². The Table caption is the same than for Table 1

A. Sic1 ¹	conformation	populations
	numbers	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	populations
	numbers	percentages
	109	67.8 ± 2.9
	138	32.1 ± 0.8
C. pSic1 ¹	conformation	populations
	numbers	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	populations
	numbers	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 Sic1¹, 2 for Sic1², 3 for pSic1¹ and 3 for pSic1². 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 (\AA) calculated from the individual gyration radii of the conformations selected by the BioEn and RamaMix analyses. The data-sets Sic1¹, Sic1² and pSic1¹, pSic1², pSic1³ were obtained using the approach TAiBP on the proteins Sic1 and pSic1. The data-sets pSic1¹³ and pSic1²³ were obtained by pooling together the conformations of pSic1³ and pSic1¹ or the conformations of pSic1³ and pSic1².

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

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

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

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

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

10 (4) IRISA, University of Rennes 1, France

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

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

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

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

18 Corresponding authors:

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

20 Jérôme Idier, jerome.idier@ls2n.fr

21 **Short title**

22 Conformational space of IDPs Sic1 and pSic1

23 August 29, 2022

24 **Extraction of boxes from Ramachandran likelihood**

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

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

40 45-49, 69-73, 76-84, due to the lack of phosphorylated proteins in the learning set of the
41 neural network. Thus, for these residues, generic boxes (Table S2) have been used as input
42 of TAI_{BP}, in order to cover the Ramachandran regions corresponding to α -helix, extended,
43 β strand and loop structures.

44 Enumeration of conformations

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

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

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

61 were scaled by a factor of 0.7.

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

71 Clustering of generated conformations

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

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

79 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
80 four largest eigenvalues along with the corresponding eigenvectors, and is formatted as an

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

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

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

93 where $N(\nu)$ is the set of neighbors, and $d(\nu, \mu)$ is the Euclidean distance between the neurons
94 μ and ν . **pSic1³** The neurons corresponding to local minima of the U-matrix, and thus to
95 local maxima of conformational homogeneity, are extracted and for all performed runs except
96 pSic1³, the protein conformation displaying the closest distance to this neuron is saved. In the
97 case of pSic1³, among the conformations saved in the neurons, the one displaying the longest
98 distance between the Carbons α of the first and the last residues is saved, in order to obtain
99 more extended conformations in agreement with the values of gyration radii measured in Ref.
100 [13]. The conformations generated during the iBP or assembly steps are finally replaced by
101 the sets of representative conformations extracted from local minima of U-matrix.

102 **Molecular dynamics refinement in implicit solvent**

103 Molecular dynamics (MD) trajectories were used to relax the Sic1 and pSic1 conformations
104 obtained from the TAIbP approach. The MD trajectories were recorded using NAMD 2.13
105 [14]. Topology parameters were taken from the force fields c36 [15] and c36m [16]. The
106 simulations were performed at a temperature of 300 K. A Generalized Born implicit solvent
107 (GBIS) [17] model was used with an ion concentration of 0.3M, and a cutoff of 12 Å for
108 calculating Born radius. A cutoff of 14 Å and a switching distance of 13 Å were defined
109 for non-bonded interactions. The RATTLE algorithm [18, 19] was used to keep all covalent
110 bonds involving hydrogens rigid, enabling a time step of 2 fs. Temperature was regulated
111 according to a Langevin thermostat [20]. At the beginning of each trajectory, the system
112 was first minimized for 1,000 steps, then heated up gradually from 0 K to 300 K in 30,000
113 integration steps. Finally, the system was equilibrated for 5,000 steps. During all steps, from
114 minimization to production, positional restraints were applied on protein backbone atoms
115 with a constant force of 1 kcal/mol. A production run of 100ps was then performed and the
116 conformation of the final frame was saved as the relaxed conformation.

117 **Determination of the populations from the Ramachan-** 118 **dran maps**

119 Using the neural network TALOS-N [1], it is possible, starting from the NMR chemical
120 shifts measured on protein atoms, to determine for each residue n a 2D probability density
121 $p_{\text{TALOS-N}}^n(\phi, \psi)$. As NMR analyses a sample containing a mixture of conformations, we

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

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

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

$$\sum_{q=1}^Q \gamma_q = 1. \quad (4)$$

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

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

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

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

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

135 In practice, TALOS-N probability densities are available on a finite rectangular grid $\{\phi_1, \dots, \phi_I\} \times$
 136 $\{\psi_1, \dots, \psi_J\}$. Let us modify (6) using a zeroth-order approximation of the integrals:

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

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

$$p_{ij}^n = p^n(\phi_i, \psi_j) = \sum_{q=1}^Q \gamma_q p_q^n(\phi_i, \psi_j). \quad (8)$$

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

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

140 $p_{ij}^n(\phi, \psi)$ being given by (8), under the constraints $\gamma_q \geq 0$ and $\sum_{q=1}^Q \gamma_q = 1$. Let us remark
141 that f is a concave function of γ , so that its local maximization with respect to γ cannot be
142 trapped in a local maximum.

143 In case TALOS-N yielded observations in the form of angle couples $y_m^n = (\phi_m^n, \psi_m^n)$,
144 $m = 1, \dots, M$, instead of a probability map, we would naturally maximize the log-likelihood
145 of the data [21],

$$\mathcal{L}(y; \theta) = \sum_{n=1}^N \sum_{m=1}^M L^n(\phi_m^n, \psi_m^n), \quad (10)$$

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

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

156 [24]. Moreover, multivariate extensions exist for the latter. In particular, in the Ref. [25]
 157 a bivariate version was introduced, motivated by problems of modelling torsional angles
 158 in molecules, and a pseudo-maximum likelihood method was proposed [24] to estimate its
 159 parameters. Moreover, a so-called *cosine* version was investigated [26] and an Expectation-
 160 Maximization (EM) algorithm was used [26, 27] to solve a problem that is almost identical
 161 to ours.

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

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

163 with

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

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

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

166 In Ref. [25], expressions of $\kappa_1, \kappa_2, \lambda$ are given as functions of the parameters $(\sigma_1^2, \sigma_2^2, \rho)$
 167 of a bivariate Gaussian:

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

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

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

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

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

177 **Maximum likelihood estimation for bivariate sine mix-** 178 **tures**

Let $Y = (y_1, \dots, y_D)$ 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 $(1, \dots, Q)$. For each d , we have

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

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

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

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

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

190 **Expectation-Maximization (EM) algorithm**

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

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

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

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

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

with

$$Q_0(\gamma, \theta^{\text{old}}) = \text{E} [\ln \text{Pr}(C; \gamma) | Y; \theta^{\text{old}}], \quad (20)$$

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

On the one hand, along classical derivations, we get

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

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

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

On the other hand,

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

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

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

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

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

201 Gradient-based log-likelihood maximization

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

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

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

206 **Partial derivative w.r.t. the weights γ**

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

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

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

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

207 **Partial derivative w.r.t. the shape parameters ζ^S**

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

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

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

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

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

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

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

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

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

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

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

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

and (30)-(32) by

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

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

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

with

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

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

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

211 given

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

212 **Partial derivative w.r.t. the location parameters ζ^L**

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

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

Moreover,

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

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

213 **Case of multiple datasets**

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

- 217 • Q normalized weights $\gamma = (\gamma_q)$ for the protein conformations (classes),
- 218 • $5NQ = 3NQ + 2NQ$ shape and location parameters specific to each conformation and
 219 each residue, respectively $\zeta_{qn}^S = (\kappa_{1qn}, \kappa_{2qn}, \lambda_{qn})$ and $\zeta_{qn}^L = (\phi_{qn}, \psi_{qn})$.

The log-likelihood then reads

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

220 where the n th residue corresponds to D_n observed pairs of angles y_{dn} .

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

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

225 The scheme developed here has been used to calculate the relative weights of the confor-
226 mations by fitting the probability Ramachandran maps obtained using TALOS-N [1].

228 Figure S1

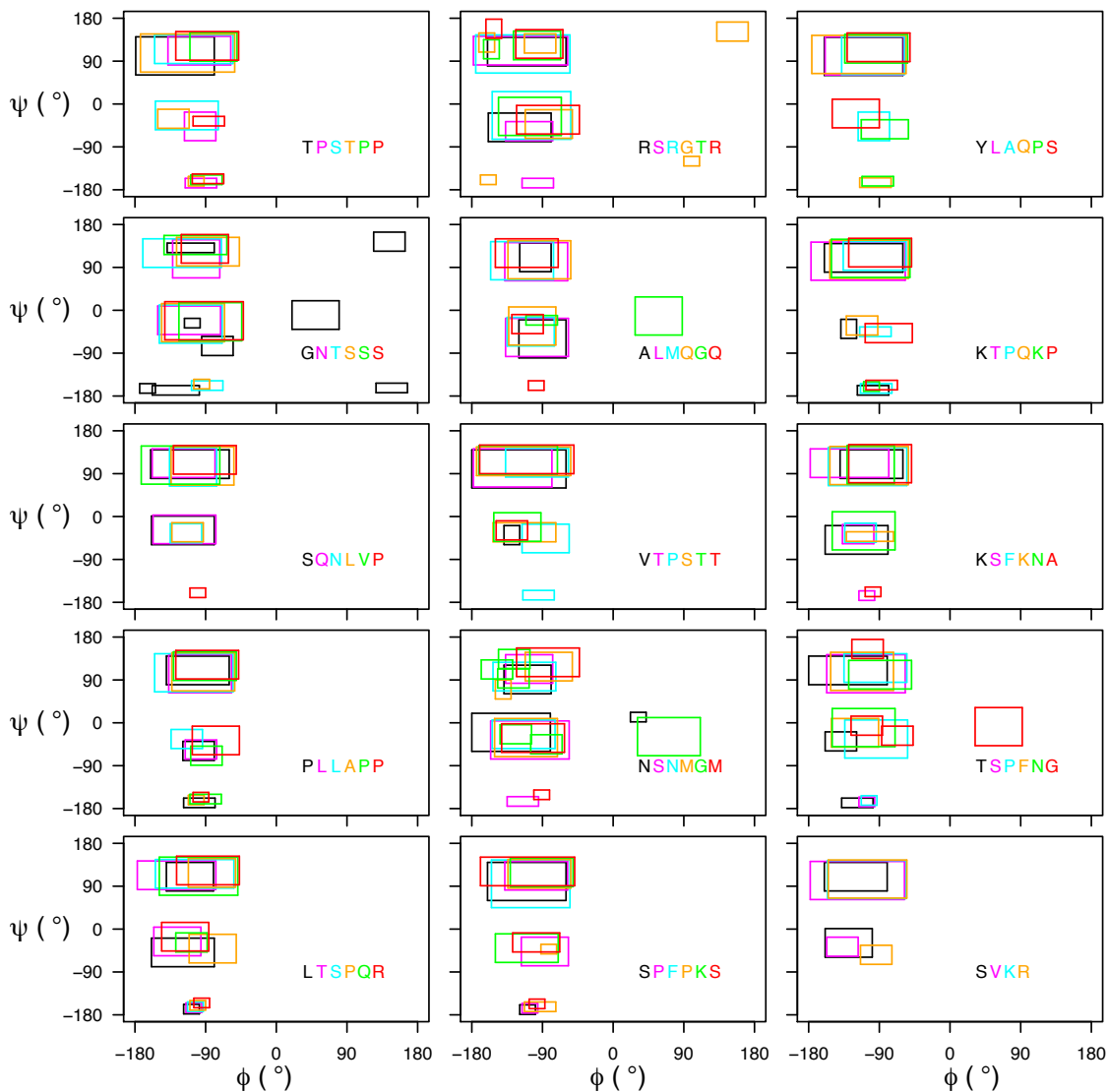


Figure S1: Boxes for backbone angles used as inputs for the run Sic1¹ and obtained from the Ramachandran maps using a threshold of 0.01. The boxes and the corresponding sequence are colored according to the considered residue.

Figure S2

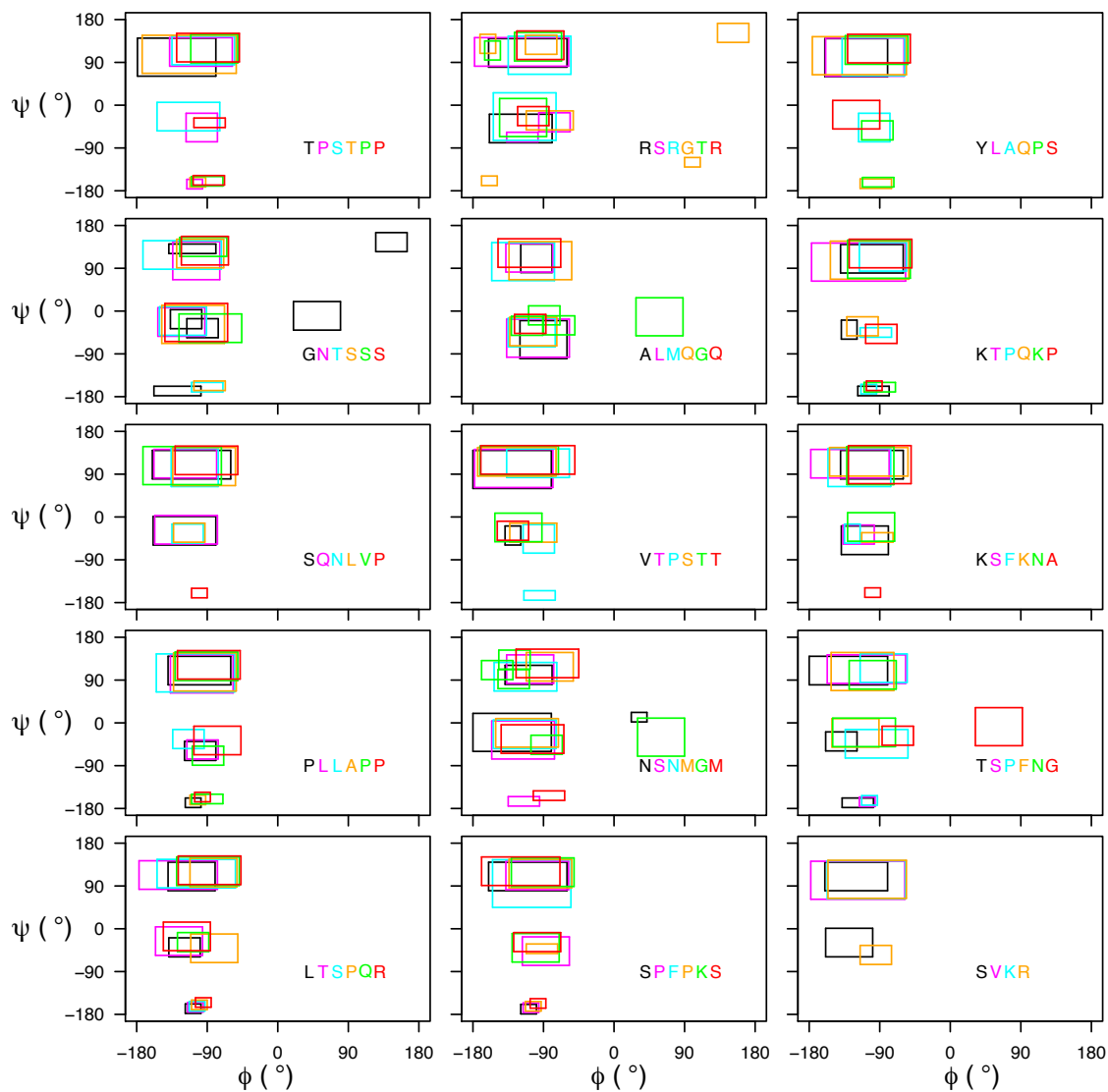


Figure S2: Boxes for backbone angles used as inputs for the run Sic1² and obtained from the Ramachandran maps using a threshold of 0.011. The boxes and the corresponding sequence are colored according to the considered residue.

Figure S3

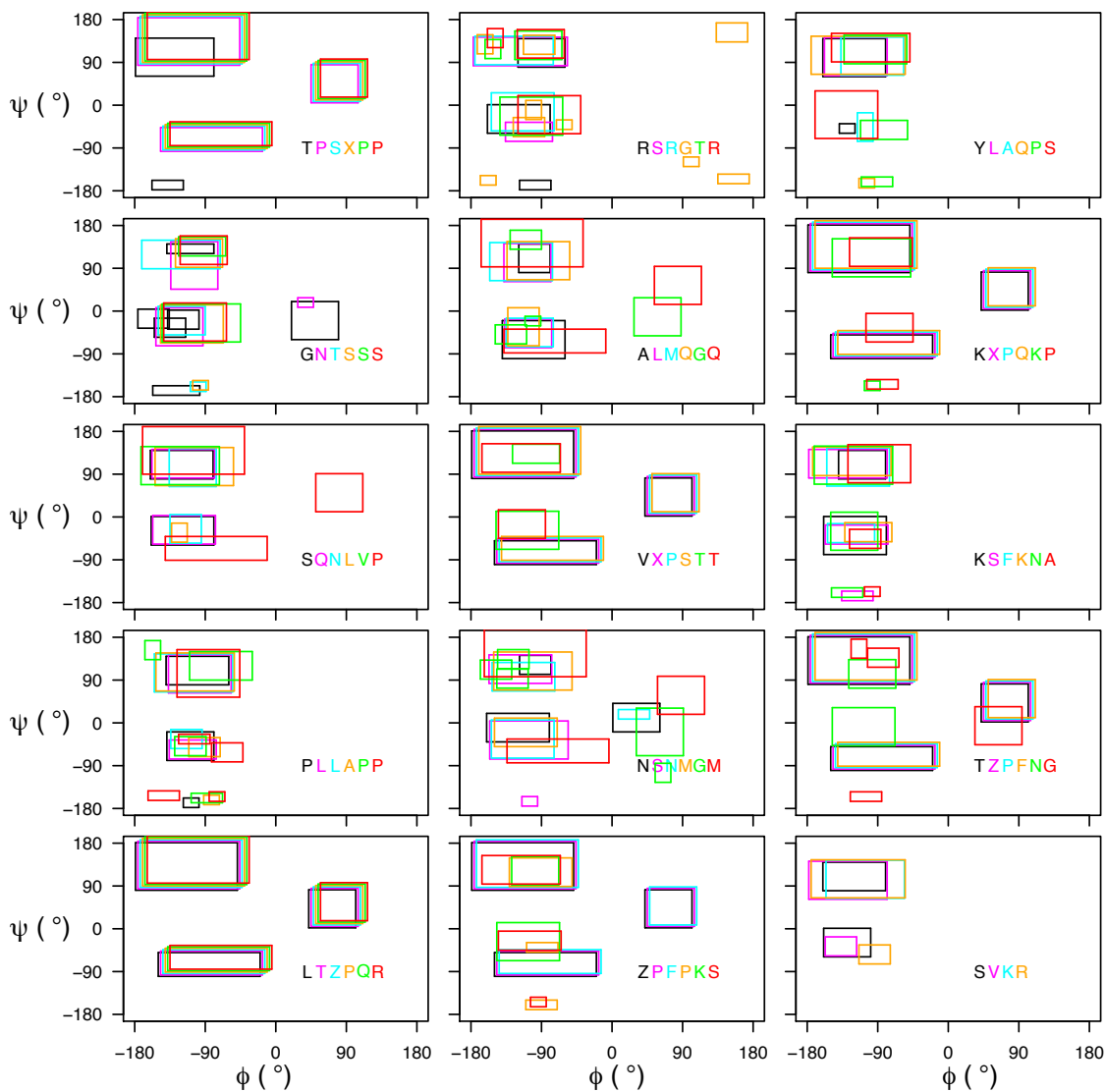


Figure S3: Boxes for backbone angles used as inputs for the run pSic1¹ and obtained from the Ramachandran maps using a threshold of 0.01. The boxes and the corresponding sequence are colored according to the considered residue. The pT and pS residues are marked as X and Z in the sequences.

Figure S4

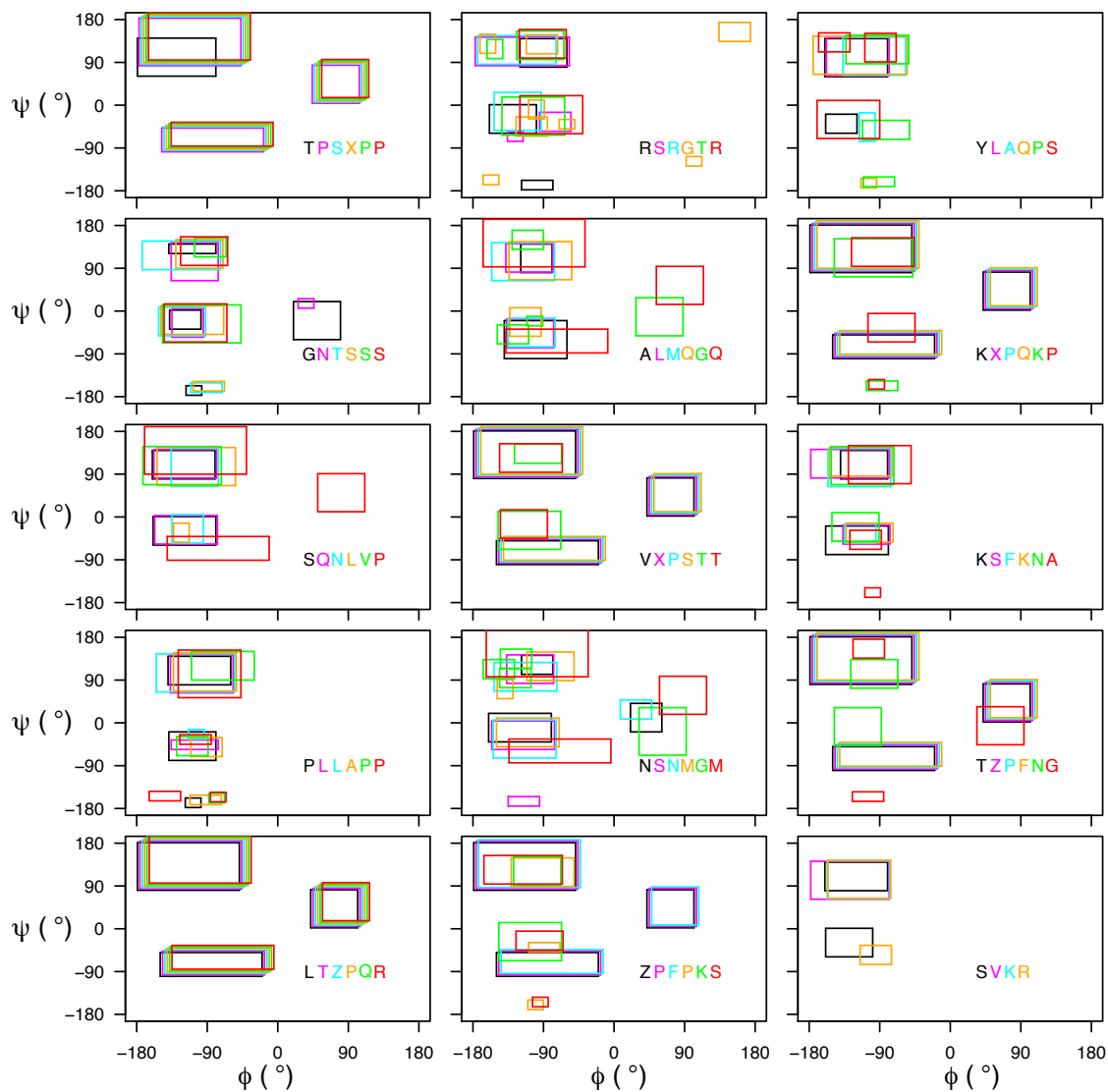


Figure S4: Boxes for backbone angle used as inputs for the run pSic1² using a threshold of 0.011. The boxes and the corresponding sequence are colored according to the considered residue. The pT and pS residues are marked as X and Z in the sequences.

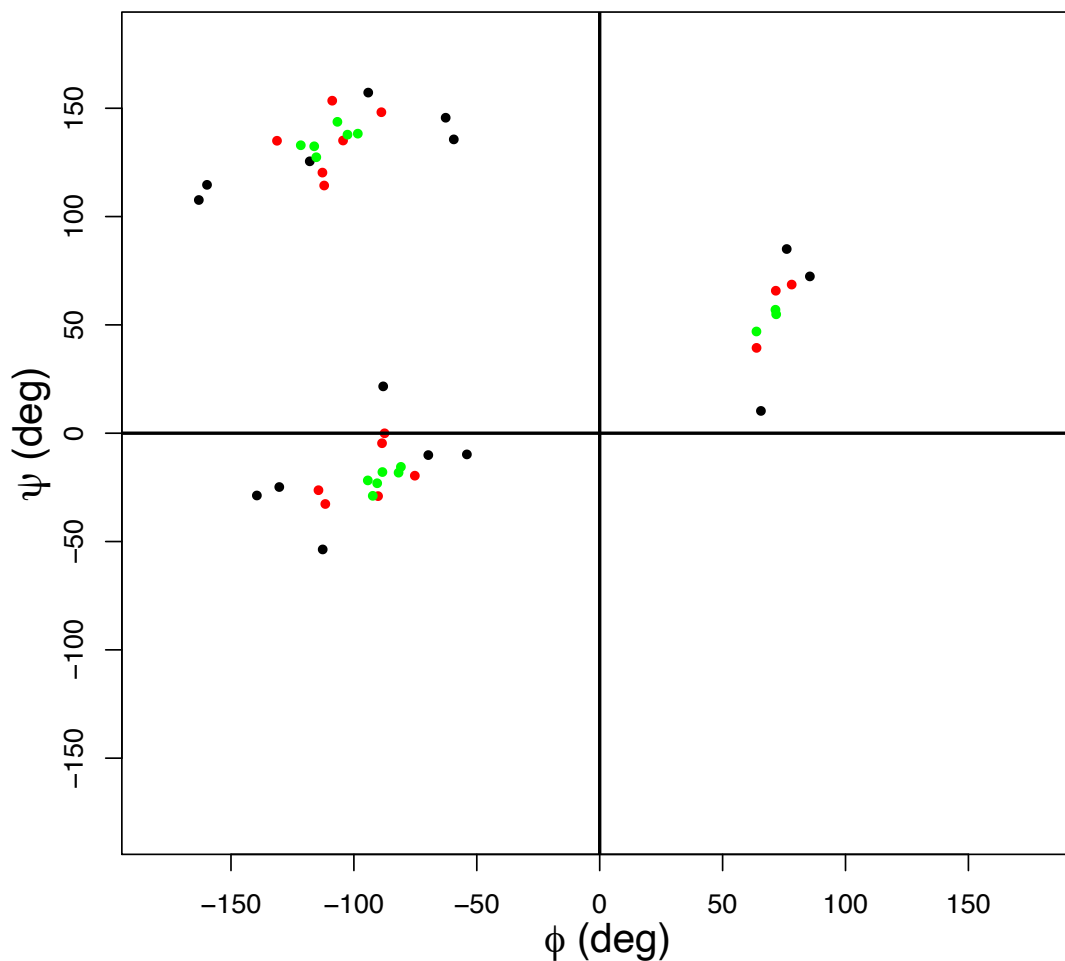


Figure S5: Synthetic Ramachandran plots used for validation of RamaMix. The colors black, red and green correspond to most (large), averaged (medium) and least (narrow) scattered 15 ϕ and ψ values. These synthetic data correspond to five hypothetical residues located in three conformations, the relative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

Figure S6

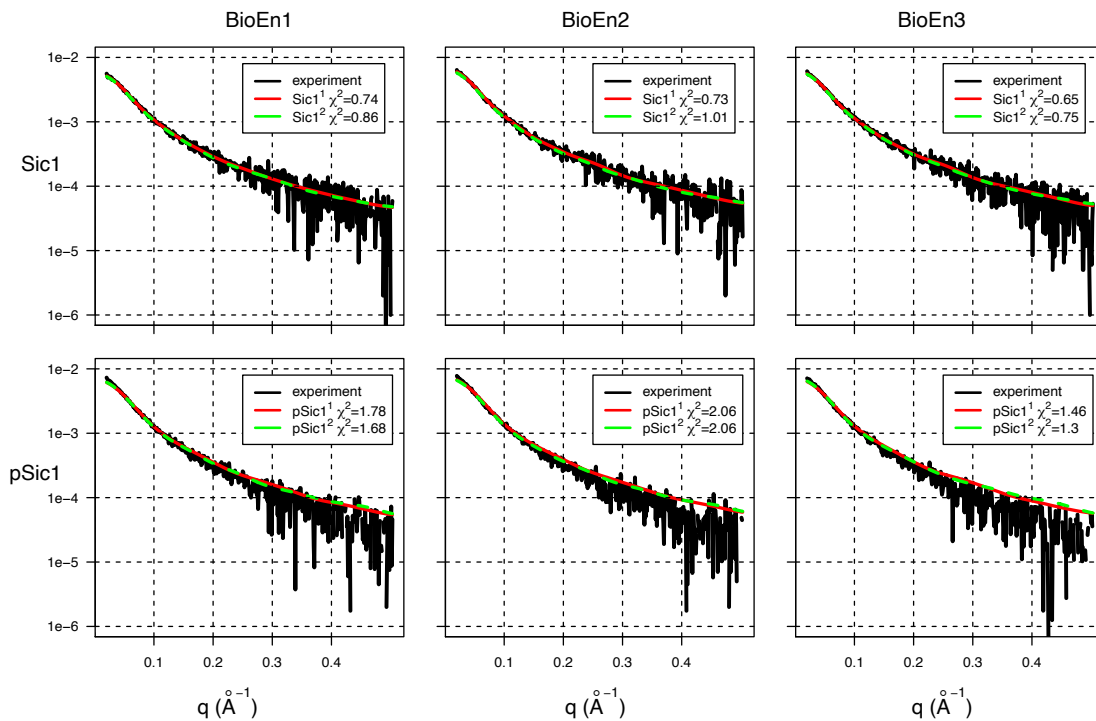


Figure S6: Superimposition of experimental SAXS curves with the reconstructed SAXS curves from the conformations of Sic1 and pSic1 selected by BioEn. The reconstructions of the SAXS curves from the selected conformations are plotted using red and green solid lines, depending on the TAiBP first (Sic1¹, pSic1¹) or second (Sic1², pSic1²) run.

Figure S7

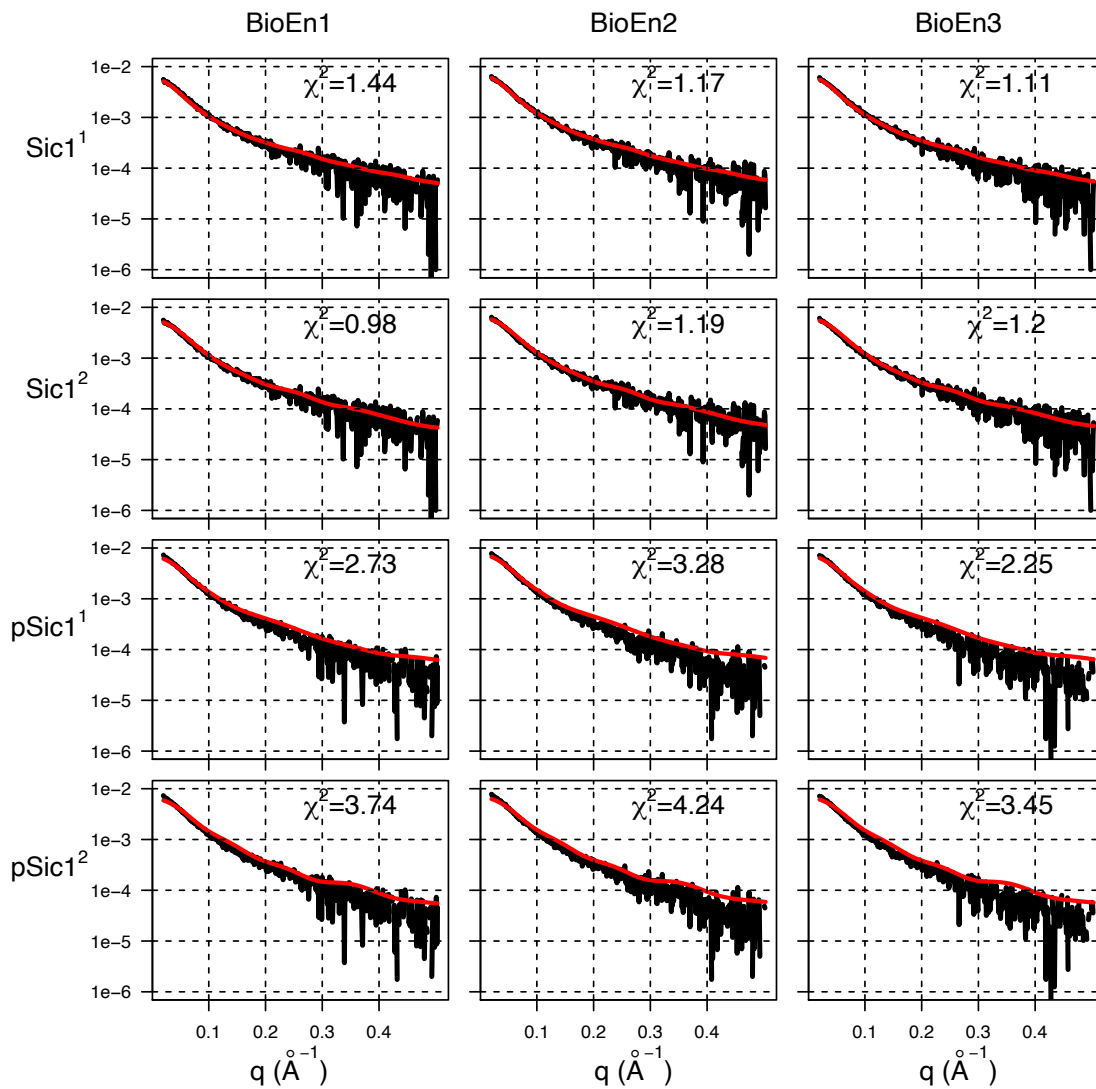


Figure S7: Superimposition of experimental SAXS curves (black) with the reconstructed SAXS curves (red) from the conformations of Sic1 and pSic1 selected by RamaMix.

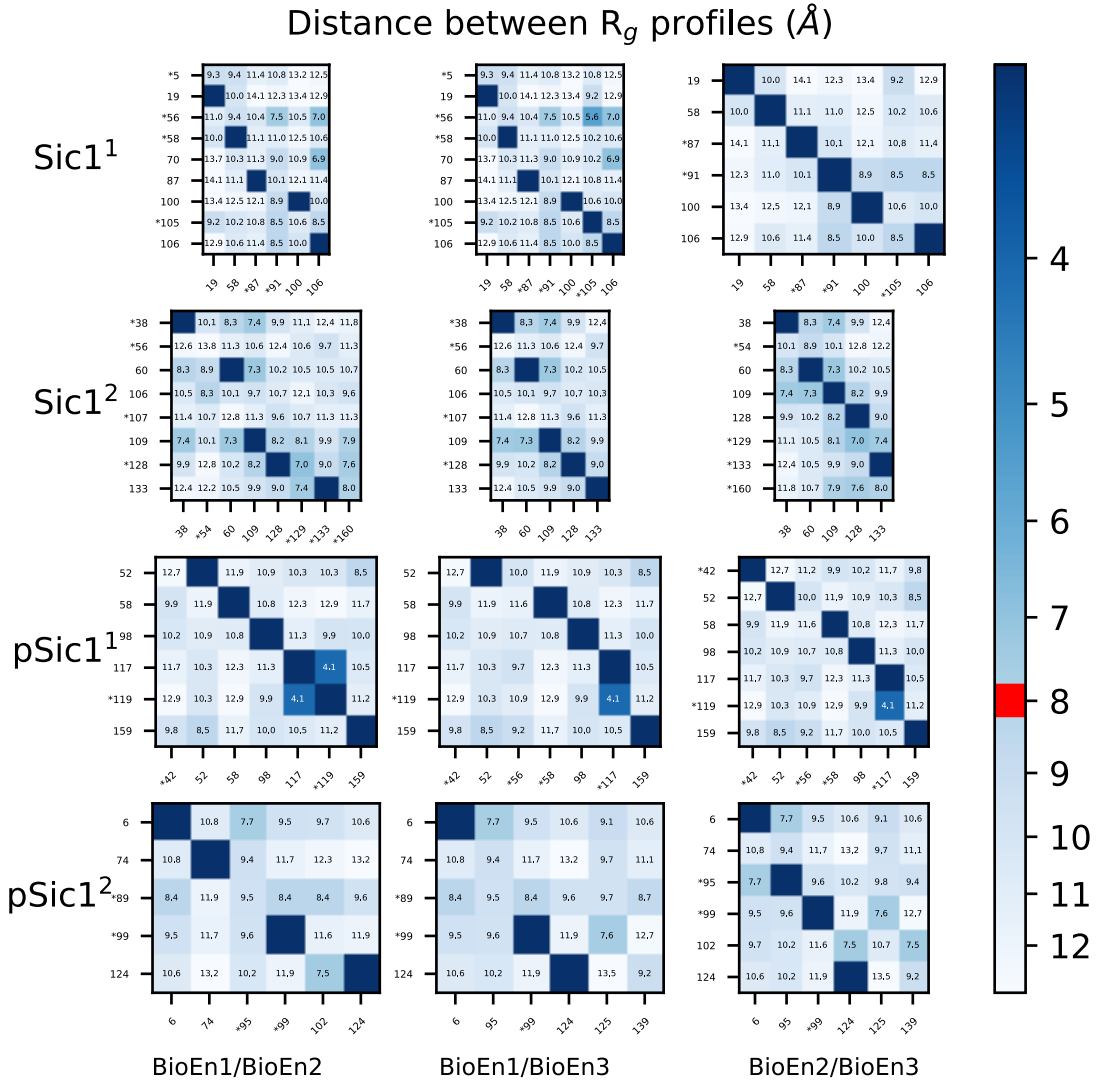


Figure S8: Distances between the profiles P_q (Eq. 11 of the main text) for local gyration radii between the conformations selected from different fittings of SAXS curves (BioEn1, BioEn2, BioEn3). The limit of 8 \AA used for the superposed plots of profiles (Figure 5 of the main text) is drawn in red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

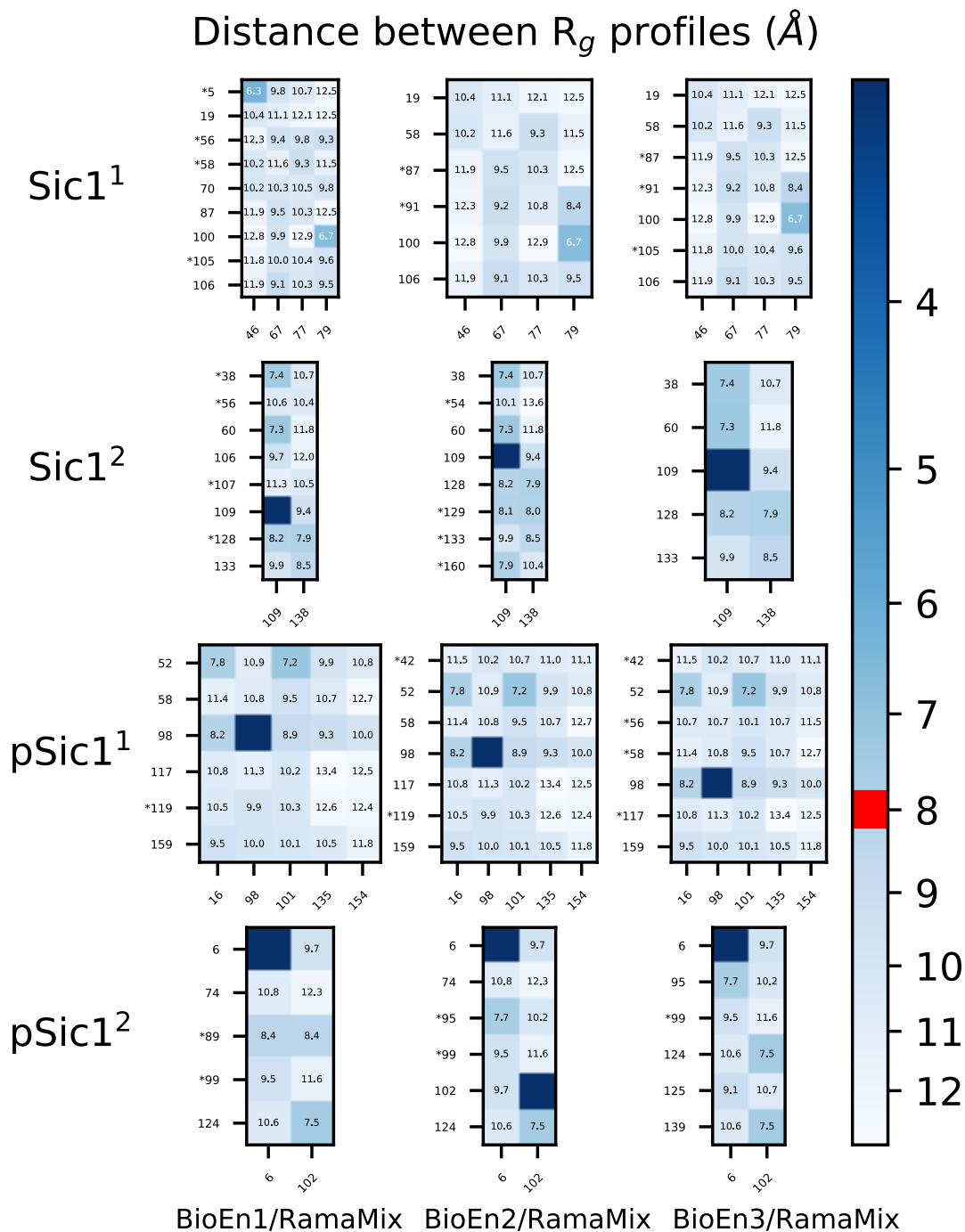
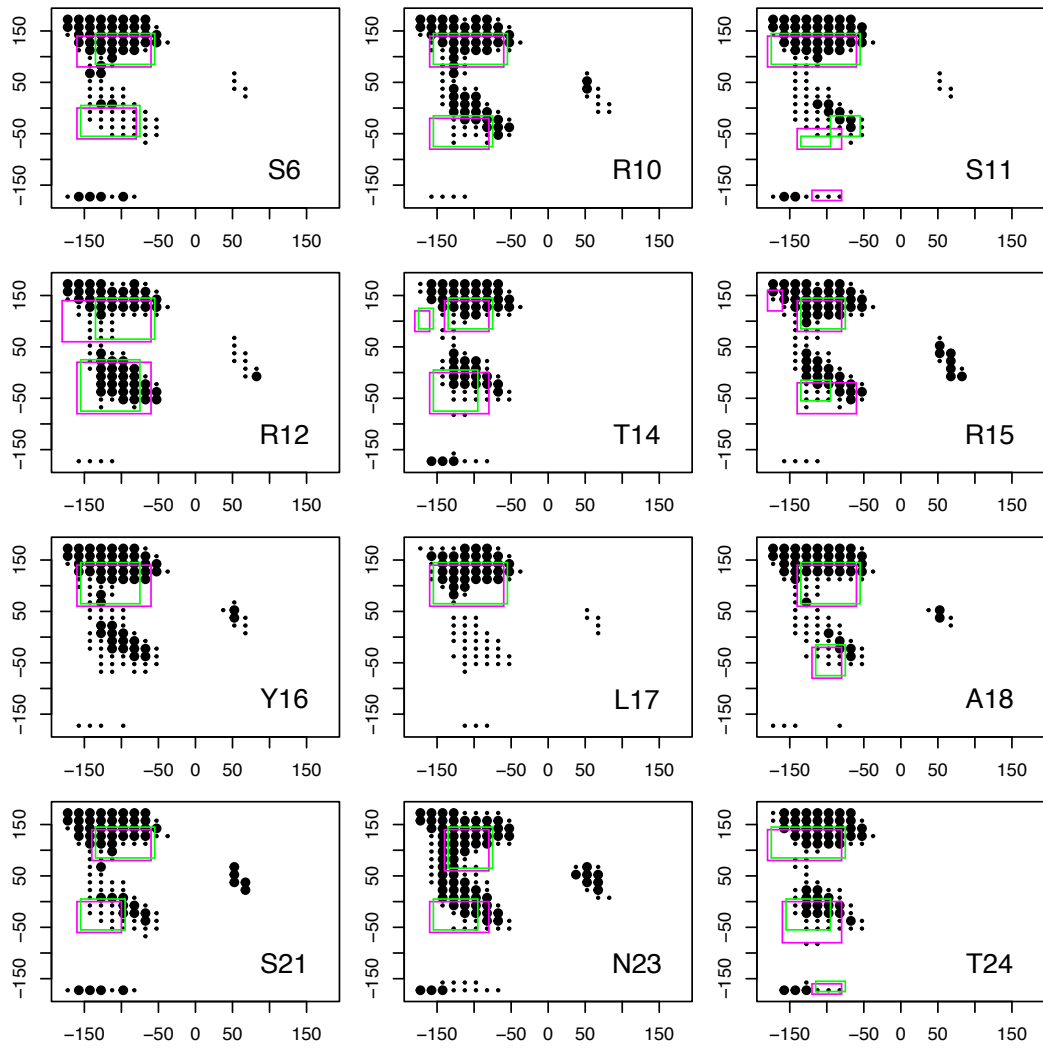
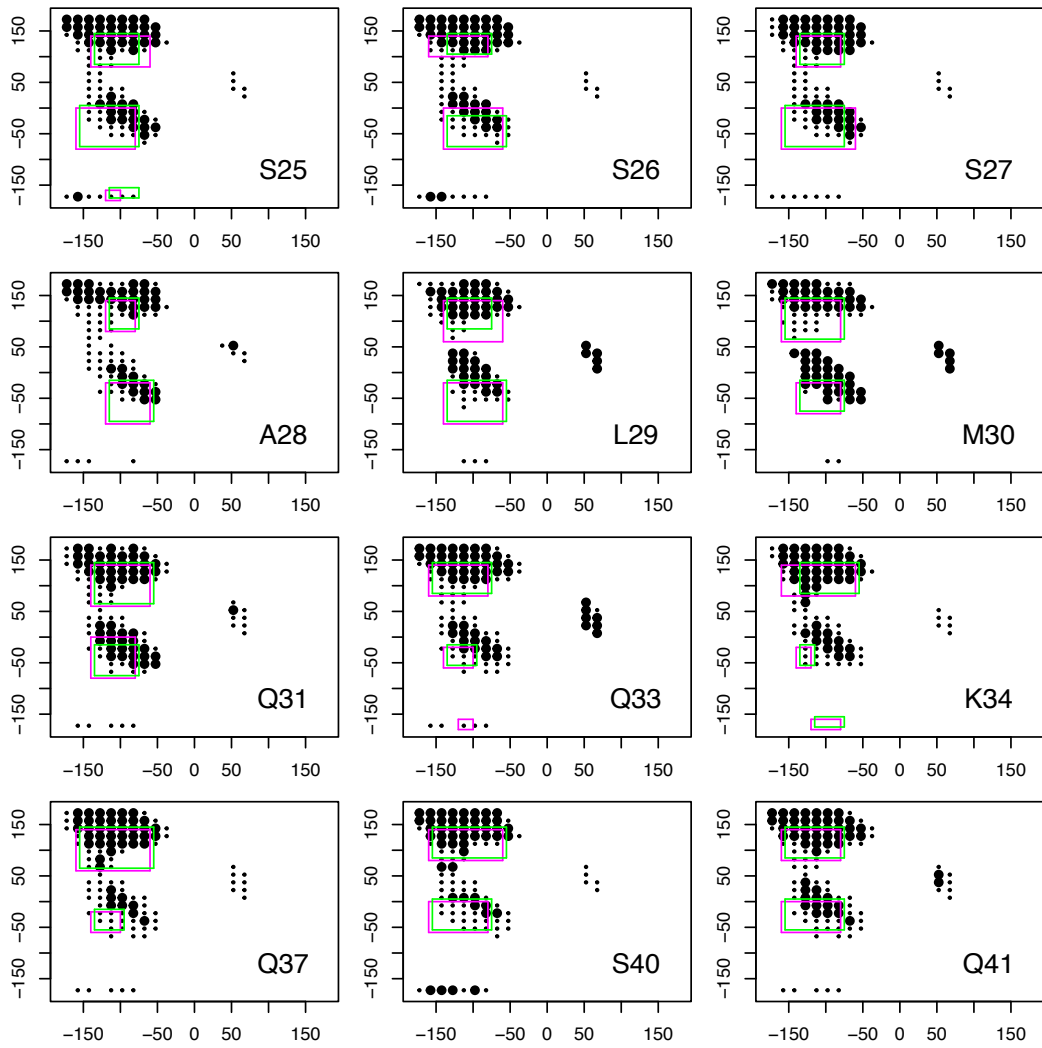
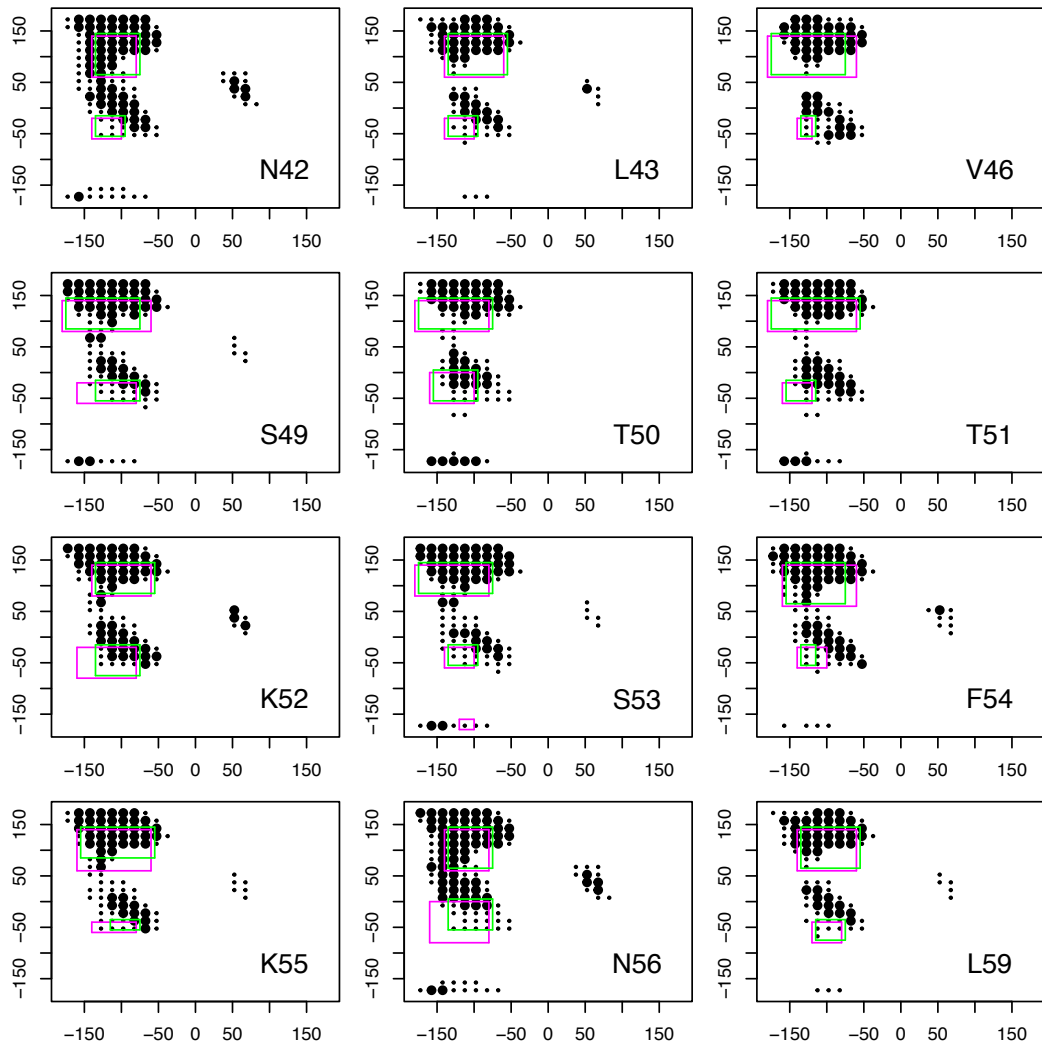


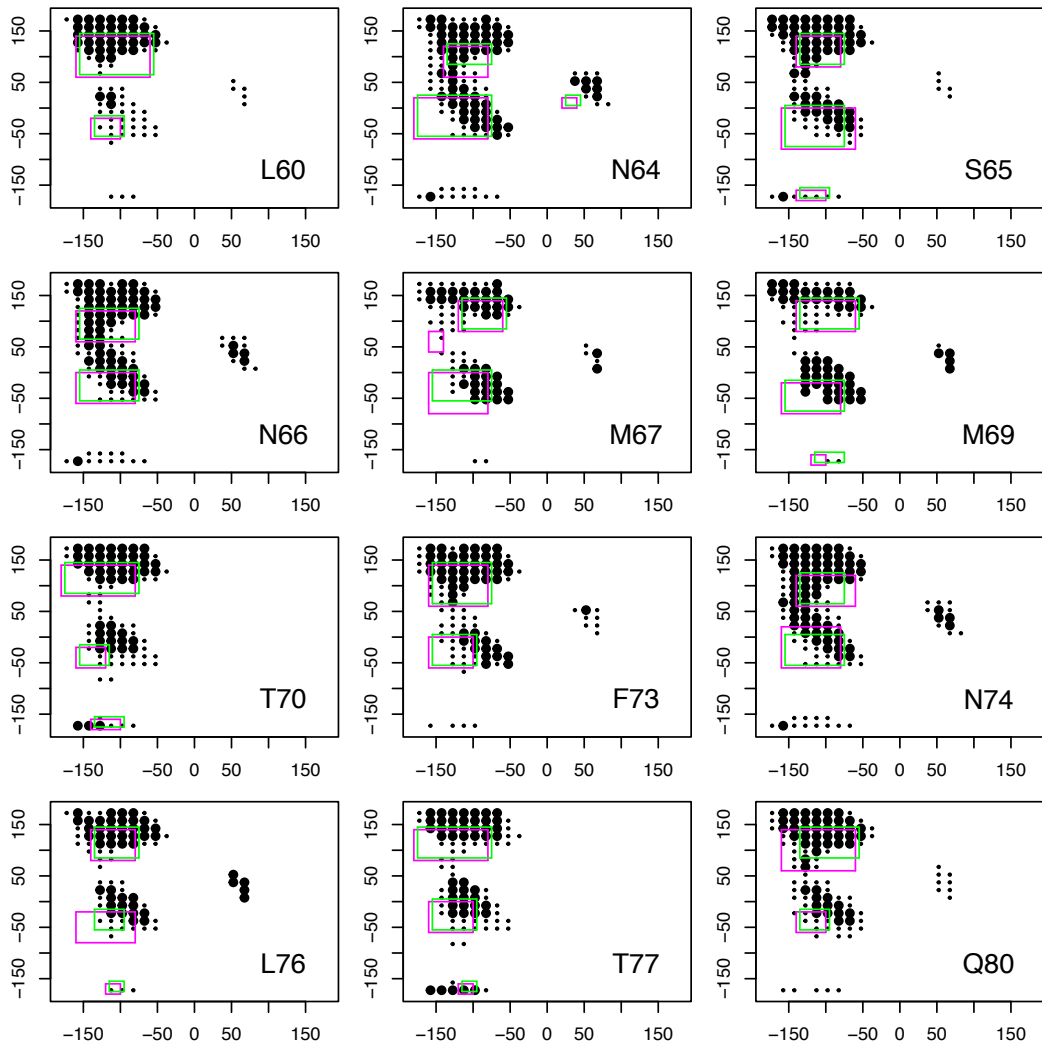
Figure S9: Distances between the profiles P_q (Eq. 11 of the main text) for local gyration radii between the conformations selected from fitting of SAXS curves (BioEn1, BioEn2, BioEn3) and of Ramachandran maps (RamaMix). The limit of 8 \AA used for the superposed plots of profiles (Figure 5 of the main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

Figure S10: Superimposition of the MERA ϕ , ψ distributions obtained on residues of Sic1 with the (ϕ, ψ) input boxes for TAIiBP. 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 TAIiBP input boxes are colored in magenta and green for the duplicated TAIiBP runs: Sic1¹ and Sic1².









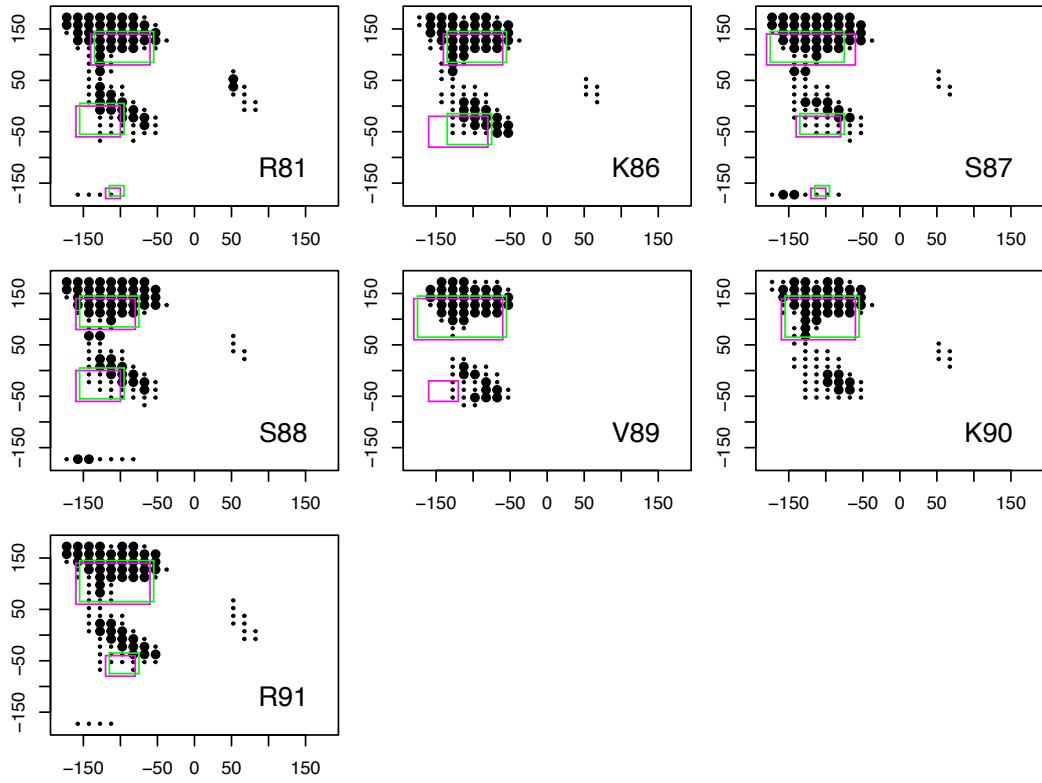
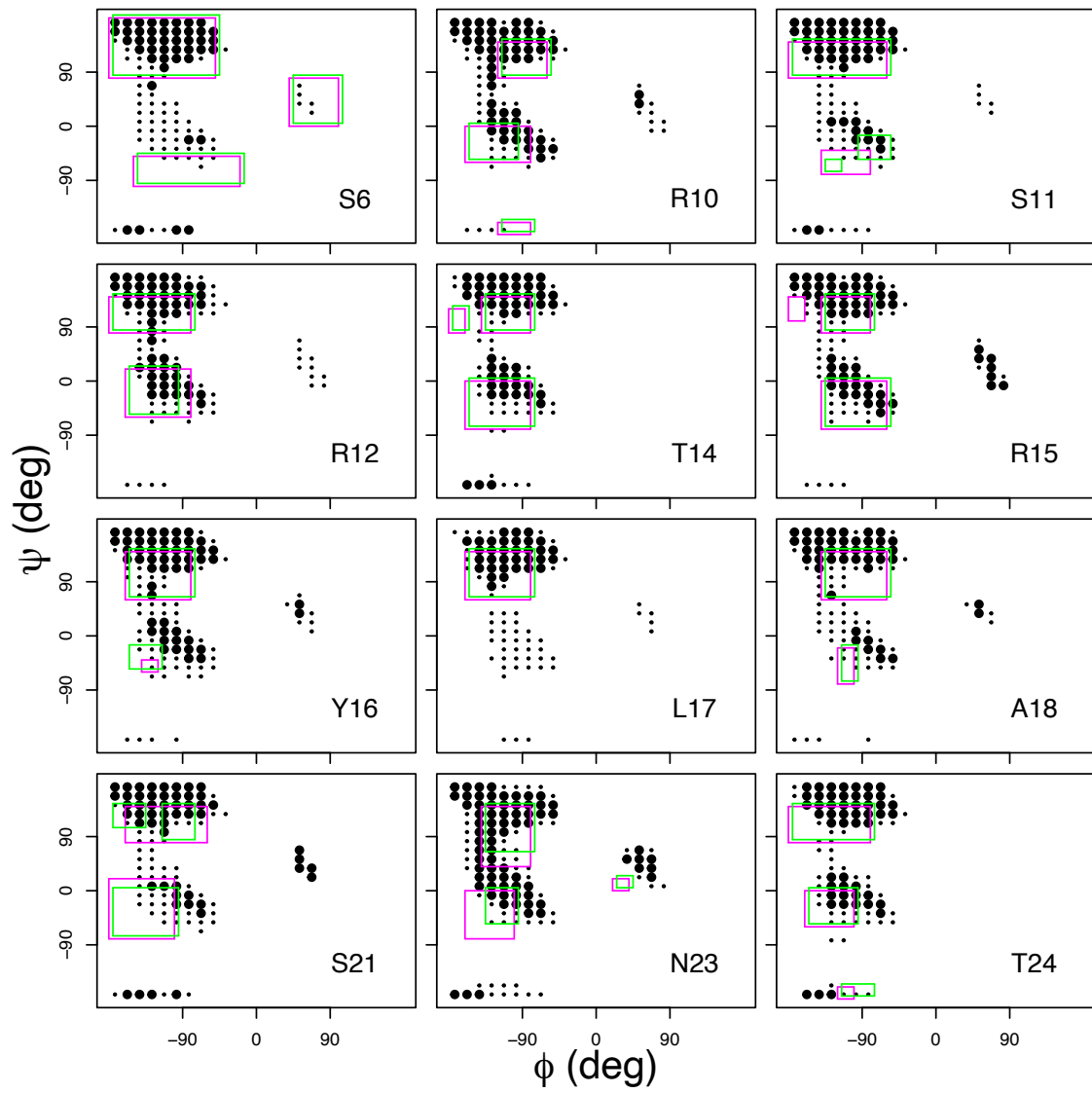
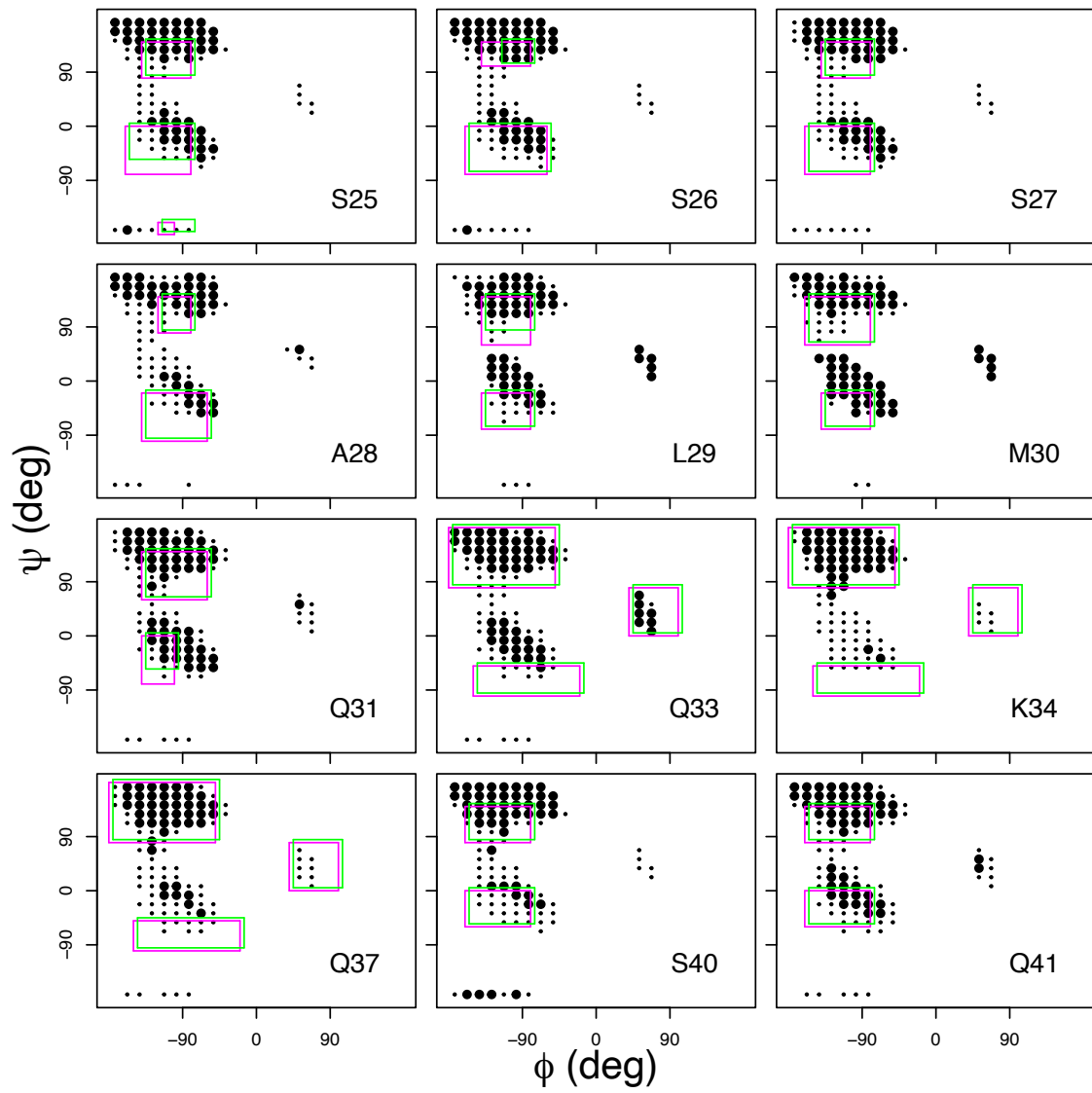
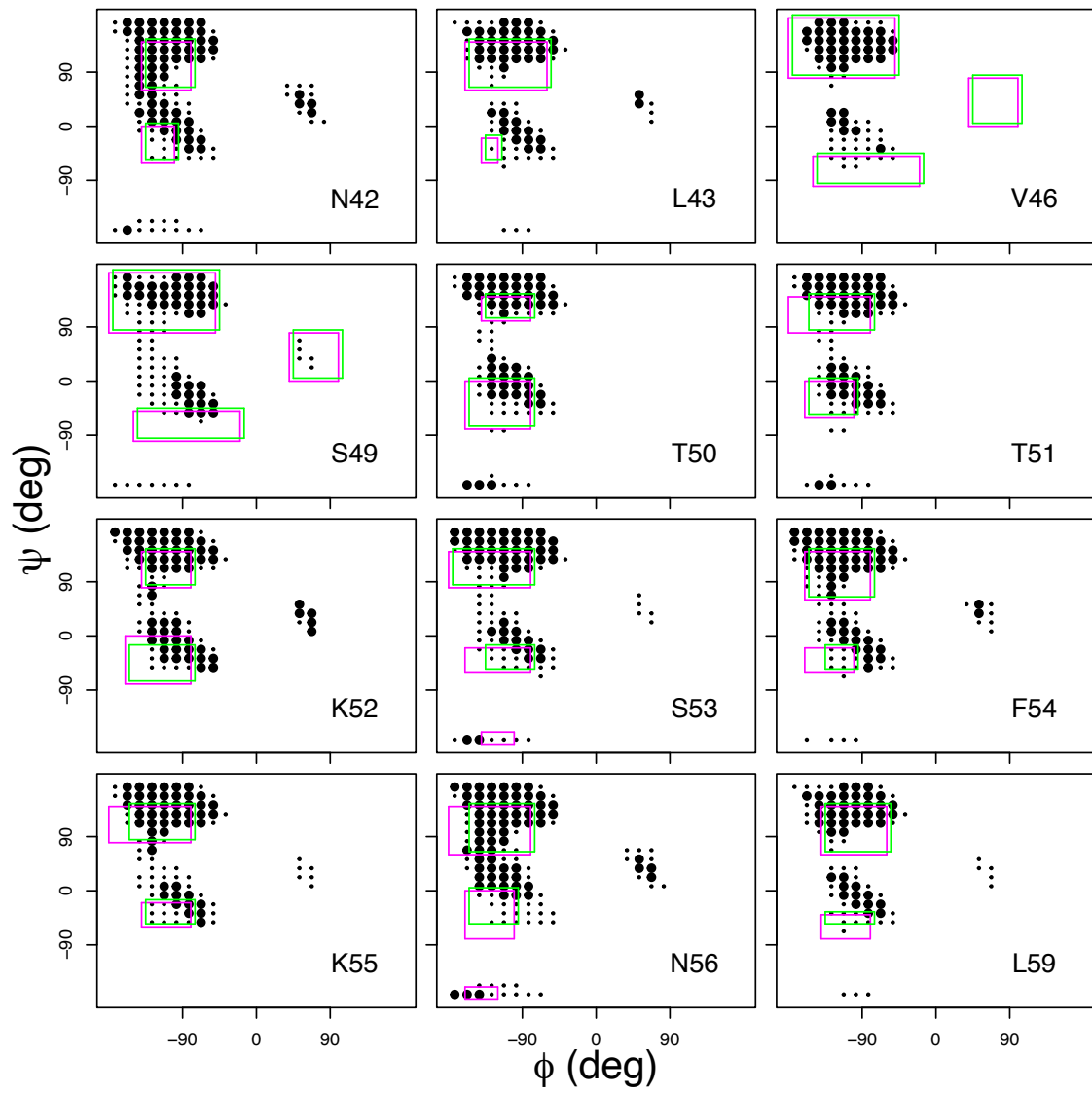
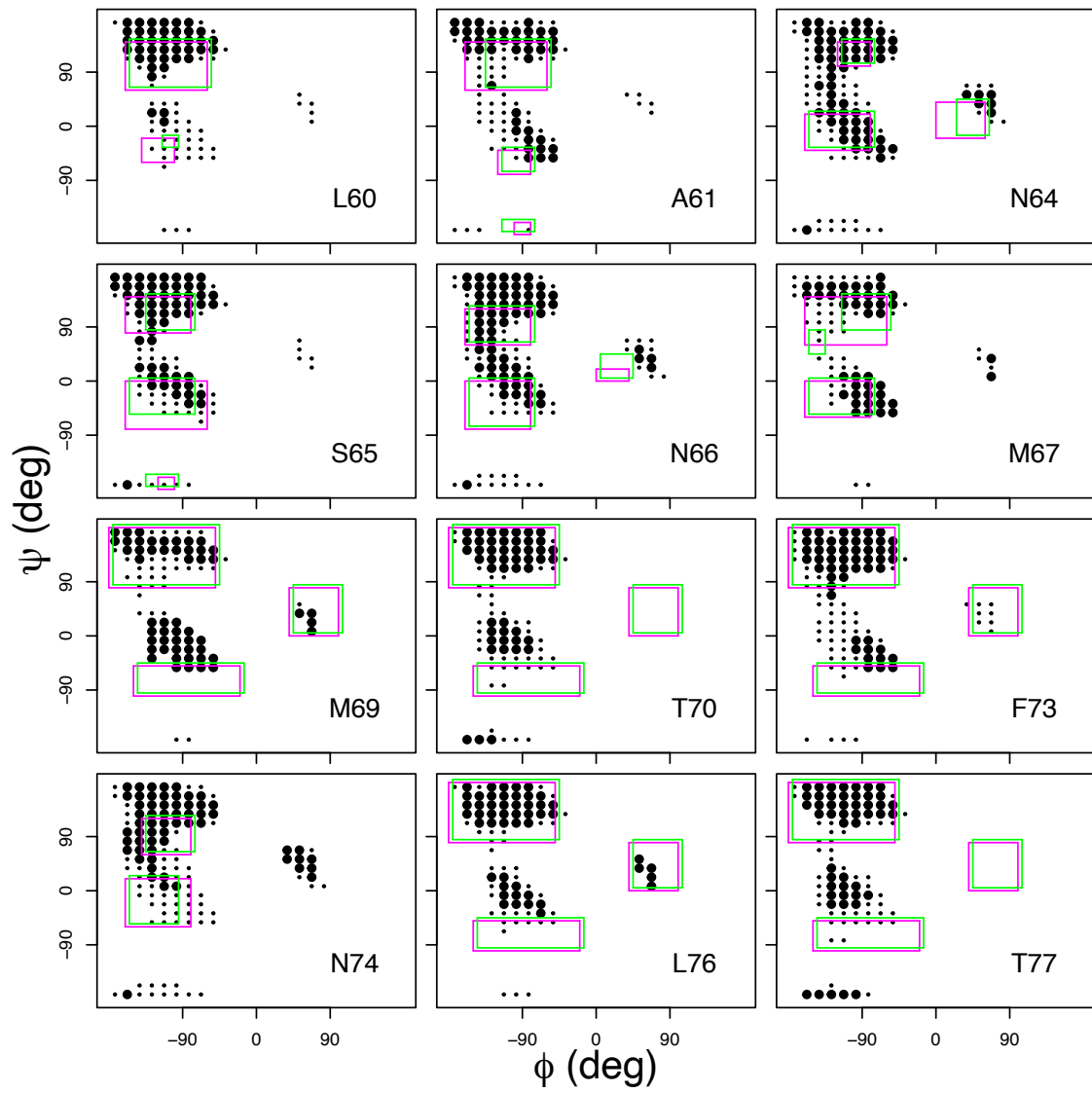


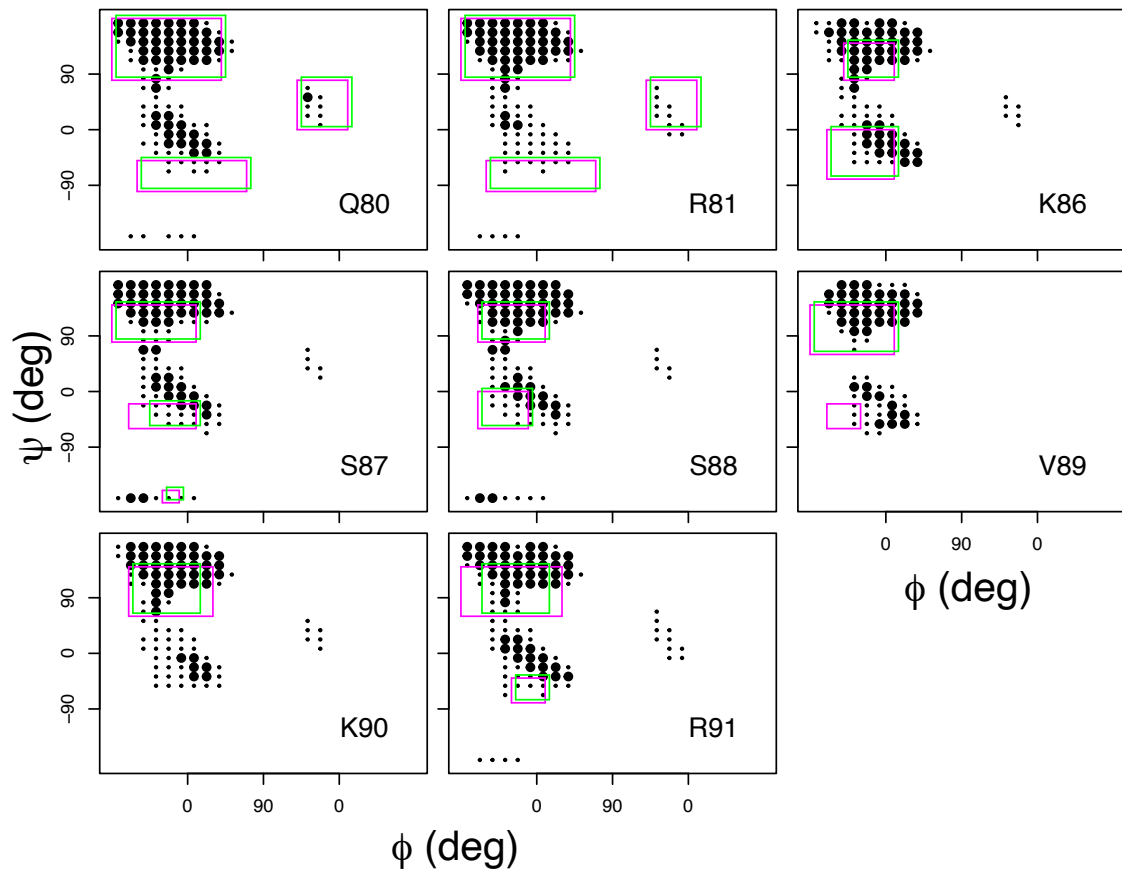
Figure S11: Superimposition of the MERA ϕ , ψ distributions obtained on residues of pSic1 with the (ϕ, ψ) input boxes for TAI_{BP}. The size of the points on MERA distribution is large for predicted probability values larger than 0.005 and small for the other probability values. The TAI_{BP} input boxes are colored in magenta and green for the duplicated TAI_{BP} runs: pSic1¹ and pSic1².











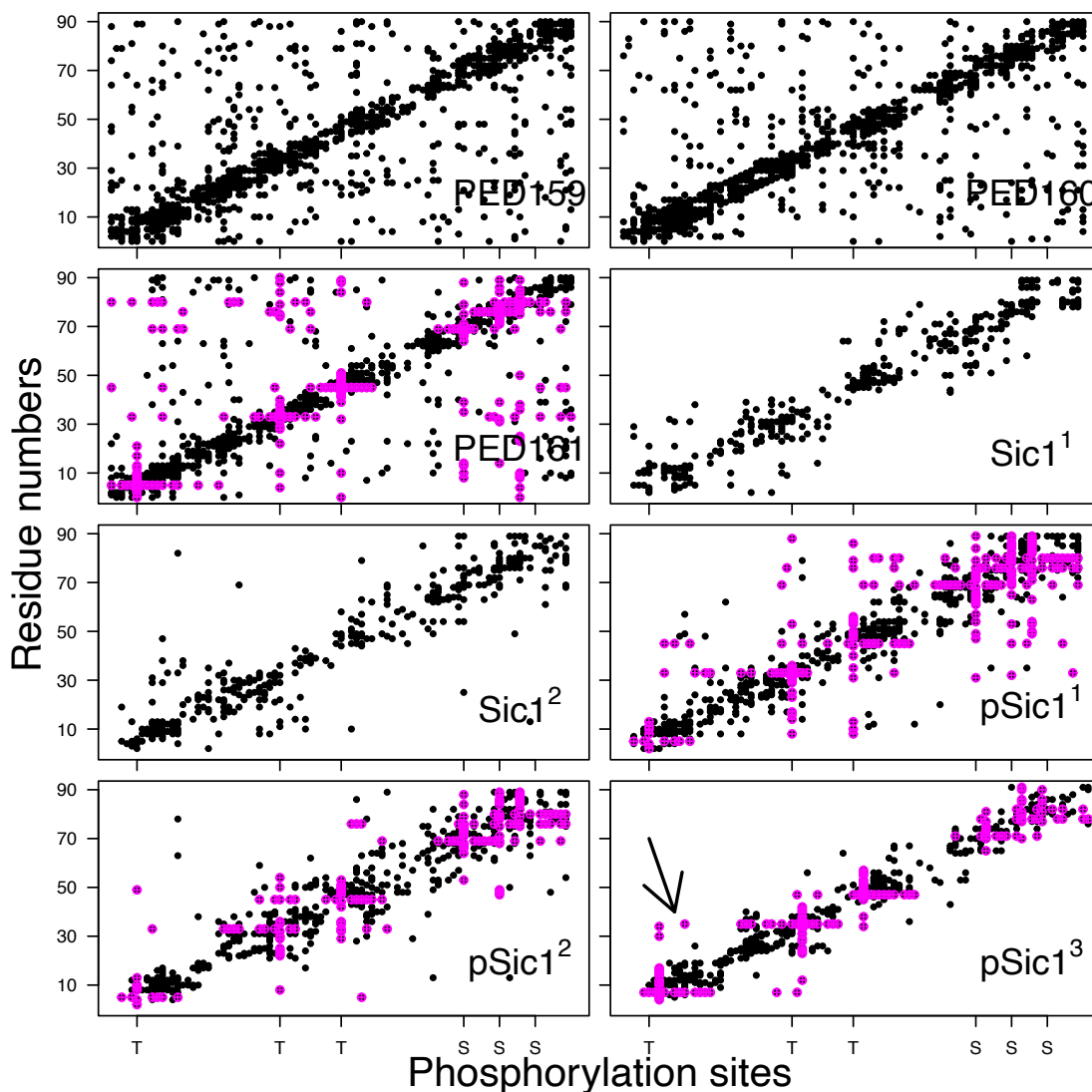


Figure S12: Contact maps displaying cumulative hydrogen bonds observed for the conformation sets PED159(Sic1), PED160(Sic1), PED161(pSic1) refined using MD simulations, as well as for the TAIiBP conformational sets Sic1¹, Sic1², pSic1¹, pSic1² and pSic1³. The hydrogen bonds involving sidechains of phosphorylated residues are plotted in magenta, whereas the other hydrogen bonds are plotted in black. An arrow on the contact map of pSic1³ 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 TAI_{BP} 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
Sic1 ¹ / <i>BioEn1</i>	0.74	3.74	1.45
Sic1 ¹ / <i>BioEn2</i>	3.64	0.73	1.5
Sic1 ¹ / <i>BioEn3</i>	1.27	1.57	0.65
Sic1 ² / <i>BioEn1</i>	0.86	4.71	1.98
Sic1 ² / <i>BioEn2</i>	1.97	1.01	0.76
Sic1 ² / <i>BioEn3</i>	0.96	2.13	0.75
pSic1 ¹ / <i>BioEn1</i>	1.78	4.21	1.59
pSic1 ¹ / <i>BioEn2</i>	3.65	2.06	2.53
pSic1 ¹ / <i>BioEn3</i>	1.8	3.41	1.46
pSic1 ² / <i>BioEn1</i>	1.68	4.42	1.51
pSic1 ² / <i>BioEn2</i>	2.24	2.06	1.59
pSic1 ² / <i>BioEn3</i>	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 χ^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 χ^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 χ^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 pSic1³ and then pooled with pSic1¹ and pSic1² to produce pSic1¹³ pSic1²³. 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 pSic1¹³ and pSic1²³ were assigned to the conformations from pSic1³.**

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 pSic1¹³ and for pSic1²³. 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 pSic1¹³ and pSic1²³ were assigned to the conformations from pSic1³.

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.

- 270 [9] T. Kohonen. Self-organizing maps. *Springer Series in Information Sciences, Heidelberg,*
271 *Germany.*, 2001.
- 272 [10] L. Miri, G. Bouvier, A. Kettani, A. Mikou, L. Wakrim, M. Nilges, and T. E. Malliavin.
273 Stabilization of the integrase-DNA complex by Mg^{2+} ions and prediction of key residues
274 for binding HIV-1 integrase inhibitors. *Proteins*, 82(3):466–478, Mar 2014.
- 275 [11] G Bouvier, N Duclert-Savatier, N Desdouits, D Meziane-Cherif, A Blondel, P Courvalin,
276 M Nilges, and TE Malliavin. Functional motions modulating VanA ligand binding
277 unraveled by self-organizing maps. *J Chem Inf Model*, 54:289–301, 2014.
- 278 [12] YG Spill, G Bouvier, and M Nilges. A convective replica-exchange method for sampling
279 new energy basins. *J Comput Chem*, 34:132–140, 2013.
- 280 [13] G. W. Gomes, M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon,
281 J. D. Forman-Kay, and C. C. Gradinaru. Conformational Ensembles of an Intrinsically
282 Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am*
283 *Chem Soc*, 142:15697–15710, 2020.
- 284 [14] JC Phillips, R Braun, W Wang, J Gumbart, E Tajkhorshid, E Villa, C Chipot, RD Skeel,
285 L Kale, and K Schulten. Scalable molecular dynamics with NAMD. *J Comput Chem*,
286 26:1781–1802, 2005.
- 287 [15] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Mackerell.
288 Optimization of the additive CHARMM all-atom protein force field targeting improved
289 sampling of the backbone ϕ and ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J*
290 *Chem Theory Comput*, 8:3257–3273, 2012.

- 291 [16] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller,
292 and A. D. MacKerell. CHARMM36m: an improved force field for folded and intrinsically
293 disordered proteins. *Nat Methods*, 14:71–73, 2017.
- 294 [17] D. E. Tanner, K. Y. Chan, J. C. Phillips, and K. Schulten. Parallel Generalized Born
295 Implicit Solvent Calculations with NAMD. *J Chem Theory Comput*, 7(11):3635–3642,
296 Nov 2011.
- 297 [18] J.P. Ryckaert, G. Ciccotti, and HJC Berendsen. Numerical integration of the cartesian
298 equations of motion of a system with constraints and Molecular dynamics of n-alkanes.
299 *J. Comput. Phys.*, 23:327–341, 1977.
- 300 [19] HC Andersen. Rattle: a "Velocity" Version of the Shake Algorithm for Molecular
301 Dynamics Calculations. *J Comp Phys*, 52:24–34, 1983.
- 302 [20] D Frenkel and B Smit. *Understanding molecular simulation: from algorithms to appli-*
303 *cations*. Academic press, San Diego, California, 2002.
- 304 [21] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Texts in Statistics.
305 Springer-Verlag, New York, NY, 2nd edition, 1998.
- 306 [22] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Exten-*
307 *sions*. Wiley series in probability and statistics. John Wiley and Sons, Inc., 1997.
- 308 [23] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Sci-*
309 *ence and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

- 310 [24] Kanti V. Mardia, Gareth Hughes, Charles C. Taylor, and Harshinder Singh. A multi-
311 variate von Mises distribution with applications to bioinformatics. *Canadian Journal*
312 *of Statistics*, 36(1):99–109, 2008.
- 313 [25] Harshinder Singh, Vladimir Hnizdo, and Eugene Demchuk. Probabilistic model for two
314 dependent circular variables. *Biometrika*, 89(3):719–723, 2002.
- 315 [26] Kanti V. Mardia, Charles C. Taylor, and Ganesh K. Subramaniam. Protein Bioinfor-
316 matics and Mixtures of Bivariate von Mises Distributions for Angular Data. *Biometrics*,
317 63(2):505–512, 2007.
- 318 [27] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamel-
319 ryck. A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci*
320 *U S A*, 105:8932–8937, 2008.
- 321 [28] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound con-
322 strained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, Septem-
323 ber 1995.
- 324 [29] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*.
325 Springer Verlag, New York, NY, 2005.
- 326 [30] D. E. Amos. Computation of modified Bessel functions and their ratios. *Mathematics*
327 *of Computation*, 28(125):239–251, 1974.
- 328 [31] J. Köfinger, L. S. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer. Efficient
329 Ensemble Refinement by Reweighting. *J Chem Theory Comput*, 15(5):3390–3401, May
330 2019.