



HAL
open science

Influence of stereochemistry in a local approach for calculating protein conformations

Wagner da Rocha, Leo Liberti, Antonio Mucherino, Thérèse Malliavin

► **To cite this version:**

Wagner da Rocha, Leo Liberti, Antonio Mucherino, Thérèse Malliavin. Influence of stereochemistry in a local approach for calculating protein conformations. *Journal of Chemical Information and Modeling*, In press, 10.1021/acs.jcim.4c01232 . hal-04770588

HAL Id: hal-04770588

<https://hal.science/hal-04770588v1>

Submitted on 19 Nov 2024

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.

Influence of stereochemistry in a local approach for calculating protein conformations

Wagner Da Rocha,[†] Leo Liberti,[†] Antonio Mucherino,[‡] and Thérèse E.
Malliavin^{*,¶}

¹ [†]*Laboratoire d'informatique de l'École Polytechnique, CNRS UMR 7161*

[‡]*Institut de Recherche en Informatique et Systèmes Aléatoires, CNRS UMR 6074,
University of Rennes, France*

[¶]*Laboratoire de Physique et Chimie Théoriques (LPCT), University of Lorraine,
Vandoeuvre-lès-Nancy, France*

E-mail: therese.malliavin@univ-lorraine.fr

² Wagner Da Rocha 0000-0002-3894-4002

³ Leo Liberti 0000-0003-3139-6821

⁴ Thérèse E Malliavin 0000-0002-3276-3366

⁵ Antonio Mucherino 0000-0003-1824-3724

⁶ Abstract

⁷ Protein structure prediction is usually based on the use of local conformational in-
⁸ formation coupled with long-range distance restraints. Such restraints can be derived
⁹ from the knowledge of a template structure or the analysis of protein sequence align-
¹⁰ ment in the framework of models arising from the physics of disordered systems. The
¹¹ accuracy of approaches based on sequence alignment, however, is limited in the case
¹² where the number of aligned sequences is small. Here we derive protein conformations

13 using only local conformations knowledge by means of the interval Branch-and-Prune
14 algorithm. The computation efficiency is directly related to the knowledge of stere-
15 ochemistry (bond angle and ω values) along the protein sequence, and in particular
16 to the variations of the torsion angle ω . The impact of stereochemistry variations is
17 particularly strong in the case of protein topologies defined from numerous long-range
18 restraints, as in the case of protein of β secondary structures. The systematic enumer-
19 ation of the conformations improves the efficiency of the calculations. The analysis of
20 DNA codons permits to connect the variations of torsion angle ω to the positions of
21 rare DNA codons.

22 November 3, 2024

23 Introduction

24 The approaches for predicting protein structures from the knowledge of their primary se-
25 quence have undergone enormous developments during the last decades.¹⁻³ One of the most
26 recent steps of this progress is the use of deep learning approaches.⁴⁻⁷ These *in silico* pre-
27 dictions pave the way towards protein function prediction and drug design and can be thus
28 considered as founding steps towards a reasoned interference with physiological processes,
29 health problems, or plant engineering.

30 In the domain of protein structure prediction, template-free *in silico* approaches uses
31 local structural information coupled to long-range proximities.⁸ The relative importance of
32 these two pieces of information is essential for a successful prediction, as pointed out by
33 Skolnick et al⁹ already long ago. As the development of covariance approaches for multiple
34 sequence alignments¹⁰⁻¹² permits the prediction of long-range restraints, a consensus was
35 found on the fact that prediction methods must be based on the local and long-range pieces
36 of information.¹³

37 The recently flourishing deep learning approaches⁴⁻⁷ have followed the same path, cap-
38 italizing on the availability of huge databases of protein structures and sequences.^{14,15} The

39 success of all prediction methods is thus quite dependent on the availability of long-range
40 restraints and consequently on the availability of multiple sequence alignment. Prediction
41 methods for the torsion angles ϕ and ψ , however, may rely on a unique protein sequence.¹⁶⁻²⁰
42 Consequently, local structure prediction can be inferred independently of alignment infor-
43 mation.

44 In several cases, long-range proximity information cannot be obtained because the size of
45 the corresponding sequence alignments is insufficient. An obvious case arises in the presence
46 of disordered regions involving many conformations, which prevents the determination of
47 precise proximities. Besides, for some protein families, the number of aligned sequences is
48 too small for statistically determining the long-range restraints.²¹ Proteins for which expres-
49 sion frameshift conducts to the expression of various polypeptides are also cases where the
50 multiple sequence alignment does not provide reliable information.²²

51 We investigate here whether local structure information is sufficient to determine the
52 protein fold. Of course, local and global structural pieces of information are closely linked:
53 we are aware of the artificial nature of their separation. The present work should be con-
54 sidered a geometric investigation of the relative importance of local and global information
55 for calculating protein conformations. In a previous analysis,²³ it was shown that the use
56 of distances restraints based on local geometry permitted to calculate protein conformations
57 closer to the target protein structures. In addition, some initial investigations in line with
58 the present work have been conducted.²⁴

59 For our purpose, we employ a purely geometric approach, the interval Branch-and-Prune
60 (iBP) algorithm, proposed some years ago to solve the problem of distance geometry in the
61 frame of protein structure.²⁵⁻²⁸ The adaptation of iBP to intrinsically disordered proteins
62 and regions is known as Threading-Augmented interval Branch-and-Prune (TAiBP).^{28,29} It
63 systematically enumerates protein conformations while heuristically overcoming the intrinsic
64 combinatorial barrier. Since then, TAiBP has been shown to allow the analysis of the
65 conformational space of various flexible or disordered proteins.³⁰⁻³²

66 In the present work, we test several variants of the iBP algorithm with different levels of
67 knowledge of the local geometry information on a database of 308 protein structures smaller
68 than 100 residues (Table S1 and Figure 1B). These high-resolution X-ray crystallographic
69 structures were selected in particular because they contain at least two secondary structural
70 elements, α helices or β strands. In the following, the torsion angles ϕ and ψ will be assumed
71 to be known within 5° intervals (or within 40° for loops in enumerating iBP runs), and the
72 focus will be put on the variations of the torsion angle ω and of the bond angles. Larger
73 variations of ϕ and ψ can be in principle taken into account using TAIiBP.²⁹

74 The present study shows that the efficiency of reconstructing the protein fold is very
75 sensitive to the knowledge of stereochemistry variations, namely the variations of bond an-
76 gle values between the heavy backbone atoms and of the torsion angles ω . We will show
77 that these stereochemistry variations depend more on the position of the residues in the
78 Ramachandran diagram than on the type of individual amino acid residues. From these
79 statistics, different types of stereochemistry were investigated, in particular, the case where
80 the stereochemistry parameters are averaged on the regions of the Ramachandran diagram
81 defined by Hollingsworth et al³³ and that we will denote in the following by *Hollingsworth*
82 *stereochemistry*. Two other stereochemistry types we analyze are the uniform one, in which
83 the parameters are taken from Engh and Huber,³⁴ and the pdb one in which the stereo-
84 chemistry parameters are extracted from each studied PDB entry. Using the Hollingsworth
85 stereochemistry, the exact knowledge of the ω backbone angles allowed us to recover most
86 of the protein folds. Even a discretized knowledge of ω allowed us to achieve decent recon-
87 struction levels. The enumeration of conformations using the iBP approach improves the
88 fold reconstruction whenever uniform stereochemistry is used. The calculations performed
89 here have been summarized in the flow chart described in Figure 1. Looking at the origin
90 of ω variability, some connection with the position of rare DNA codons was emphasized, in
91 agreement with recent literature results.³⁵⁻³⁷

92 **Materials and Methods**

93 **Preparation of the protein database**

94 The list of PDB entries of X-ray crystallographic structures with identity between sequences
95 smaller than 20%, resolution better than 1.6 Å and R factor better than 0.25, has been
96 downloaded from the server `dunbrack.fccc.edu/pisces`³⁸ providing 3757 protein chains.
97 From this list, 308 protein chains were selected, smaller than 100 residues, not containing
98 *cis* peptide bonds, and for which more than two secondary structure elements (α -helix or
99 β -strand) are present (Table S1 and Figure 1A).

100 The proteins forming the database display a size mostly in the range of 60-90 residues,
101 with a smaller number of proteins containing 20 to 60 residues (Figure S1). The percentage
102 of α helices is uniformly distributed among the proteins, whereas the β -strand and the loops
103 display more concentrated distributions in the 0-20% for β strands and 20-40% for loops.

104 **interval Branch-and-Prune approach**

105 The interval Branch-and-Prune approach (iBP) algorithm was initially proposed by Mucherino
106 and coworkers^{25,26,28,39-42} to enumerate the conformations of proteins verifying sets of dis-
107 tance constraints. The space of all possible protein structures is described as a tree and the
108 available geometric information permits tree branching and pruning. Each time a branch of
109 the tree is pruned, the iBP calculation is stopped and resumed at the previous positioned
110 atom. This branch-and-prune description of the problem makes possible a discrete enu-
111 meration of solutions, and consequently strongly contrasts with most of the optimization
112 approaches usually employed for the determination of biomolecular structure.

113 If not otherwise stated, the conformations of the proteins have been recalculated using
114 one-shot iBP runs, in which the run was stopped after producing the first solution. The
115 branching part was performed on ϕ and ψ torsion angles using intervals of 5° centered
116 around the true ϕ and ψ values. The torsion angles are converted into distance intervals,

117 which are discretized with a maximum of four branches separated by at least 0.1 Å, which
118 defines the discretization factor ϵ .

119 The ω values of the torsion angle of peptidic planes were used as pruning restraints as
120 well as the χ_1 torsion angle defining L amino acid residues. A last pruning restraint is related
121 to all interatomic distances which should be larger than the sum of van der Waals radii of the
122 atoms, using a scale factor of $\rho = 0.8$ on the radii if not otherwise stated. This approach is
123 reminiscent of the reduction of the van der Waals interactions during the simulated annealing
124 procedure in NMR structure calculation.⁴³

125 **Variations of stereochemistry during iBP calculations**

126 Several definitions of protein stereochemistry focusing on the backbone bond angles and ω
127 torsion angles were used as inputs for the calculations. Uniform stereochemistry was de-
128 fined using the values from the force field PARALLHDG (version 5.3)³⁴ (Table S2). Two
129 variations of the stereochemistry are explored: (i) pdb stereochemistry in which the bond
130 angles and ω torsion angle were extracted from the PDB conformation of the considered pro-
131 tein, (ii) Hollingsworth stereochemistry in which the bond angles and the ω torsion angle are
132 taken as the average stereochemistry values calculated from the regions of the Ramachandran
133 diagram defined by Hollingsworth et al from the analysis of high-resolution X-ray crystal-
134 lographic structures.³³ The correspondence between the regions displayed in Figure S2 and
135 the definition of Ref⁴⁴ is given in Table S3.

136 For pdb stereochemistry, each protein residue is defined by a 3-letter name, the alphabetic
137 order of the names coding for the positions of the residues in the primary sequence, the first
138 residue being AAA, the second one AAB, and so on. The topology files in CNS format⁴⁵
139 were modified by using this residue code to define the amino acid residues along the primary
140 sequence as well as the different atom types for each residue. Using these atom types, and
141 the stereochemistry values in the PDB structure, values of bond lengths and angles are then
142 generated for each residue along the sequence and stored in the CNS parameter file. This

143 allows us to take into account any possible variations of protein stereochemistry (pdb or
144 Hollingsworth) along the protein sequence.

145 **Analysis of obtained conformations**

146 The analysis of protein conformations obtained with iBP has been performed using the
147 MDAnalysis package⁴⁶ and STRIDE.⁴⁷ Sidechains were added to the protein backbone using
148 the Relax procedure⁴⁸ of Rosetta⁴⁹ for the refinement of a one-shot iBP run, in the case of
149 uniform and Hollingsworth stereochemistry. During the Relax procedure, 10 conformations
150 were generated and the procedure was repeated 5 times.

151 **Results**

152 **Analysis of protein stereochemistry**

153 The stereochemistry of the 3757 protein chains downloaded from the server `dunbrack.fccc.`
154 `edu/pisces`³⁸ has been analyzed (Figure 1) by calculating the average values of the backbone
155 bond angles $N-C_\alpha-C$, $C_\alpha-C-N$, $C-N-C_\alpha$, $C_\alpha-C-O$, and $O-C-N$ (Figure 2). The
156 negative torsion angles ω were shifted by 360° in order to obtain ω value variations around
157 180° . The averaged and standard deviation values of these bond and torsion angles are
158 plotted according to the type of amino acid (Figure 3, left column) and to the region of the
159 Ramachandran diagram defined by the backbone torsion angles ϕ and ψ (Figure 3, right
160 column). The Ramachandran regions were taken from the definition given in the work of
161 Hollingsworth et al³³ (Figure S2).

162 Almost all average angle and standard deviation values display flat profiles along the type
163 of amino acid (Figure 3, left column). The standard deviations for ω angles display slight
164 variations among the amino acids, especially for Glycine, Tryptophan, and Tyrosine (Table
165 S4). Unsurprisingly, the dashed line indicating the Engh and Huber³⁴ values is close to the
166 average values of angles. The bond angles $C_\alpha-C-N$ and $C-N-C_\alpha$ display the smallest

167 standard deviations, whereas the bond angles $\text{N}-\text{C}_\alpha-\text{C}$ and $\text{C}_\alpha-\text{C}-\text{O}$ display the largest
168 ones. The averaged values of the angle $\text{C}-\text{N}-\text{C}_\alpha$ display one outlier for Proline residues,
169 with a shift of around 2° . The averaged values of the angle $\text{N}-\text{C}_\alpha-\text{C}$ display four outliers,
170 all shifted by around 2° : two are shifted towards larger values for amino acids Glycine and
171 Proline, and two are shifted towards smaller values for amino acids Isoleucine and Valine. The
172 outliers positions of Isoleucine and Valine have been recently observed⁵⁰ for the propensity
173 scales of the Ramachandran regions. In addition, Proline and Glycines have been known for
174 decades to influence local geometry.^{51,52}

175 Interestingly, the profiles along the Hollingsworth regions (Figure 3, right column) are
176 much more variable for the average as well as for the standard deviation values (Table S4).
177 Among the bond angles, the angle $\text{N}-\text{C}_\alpha-\text{C}$ and $\text{C}_\alpha-\text{C}-\text{O}$ display the most variable profiles.
178 The large variability of these angles may arise from the involvement of the atoms N and O
179 into hydrogen bond network stabilizing the protein secondary structures.

180 The angle variability depends on the Ramachandran regions. The region A (Figure S2,
181 red), corresponding to the regular α helix, produces angles close to the Engh and Huber val-
182 ues, with the smallest standard deviations. The region D (Figure S2, green), corresponding
183 to the 3-10 helix, displays standard deviations similar to the region A, but average values
184 shifted to upper values for bond angles $\text{C}_\alpha-\text{C}-\text{N}$ and $\text{N}-\text{C}_\alpha-\text{C}$ and to lower values for bond
185 angle $\text{C}_\alpha-\text{C}-\text{O}$. The region B, corresponding to the regular β strand (Figure S2, blue), and
186 P, corresponding to the polyproline region (Figure S2, brown), displays also average values
187 mostly close to the Engh and Huber values except for the angle $\text{N}-\text{C}_\alpha-\text{C}$, but the standard
188 deviations are larger, especially for the angle ω . The regions g, Z, located between the α and
189 β regions of the Ramachandran diagram, the regions G, d, p, located in the loop region of
190 positive ϕ value, and the region E, all display large standard deviations and shifted average
191 values.

192 The right column of Figure 3 permits the definition of protein stereochemistry depending
193 on the Ramachandran region by averaging over each Hollingsworth region (Figure S2), the

194 values of bond angles and ω angles. Due to the profile variations of Figure 3, one may
195 expect that this Hollingsworth stereochemistry will be more variable than a stereochemistry
196 based on the amino-acid type. This is not surprising as the amino acid type is defined
197 by the sidechains which are more far apart from the backbone than the ϕ and ψ torsion
198 angles. In the following, the protein stereochemistry will be modeled as uniform ie. uniquely
199 defined from the atom type, following the measurements of Engh and Huber³⁴ (Table S2),
200 as Hollingsworth with averaged values determined from the ϕ , ψ torsion angles (Figure 3
201 and Table S4), and as a pdb, using angles measured on the PDB structure of the considered
202 protein.

203 **Effect of stereochemistry on the protein conformations generated** 204 **with iBP**

205 Several experiments were performed on the database of proteins to reconstruct the conformations using iBP with the previously chosen types of stereochemistry. First, one-shot runs were realized with calculations stopping after obtaining the first conformation (Figure 1C) and intervals of 5° for ϕ and ψ angles. Then, the protein targets were submitted to a full exploration of the tree, using narrow intervals (5°) of ϕ and ψ in the secondary structure elements, and larger intervals (40°) in the connecting loops (Figure 1D).

211 Figure 4 displays the distributions of the root-mean-square deviation (RMSD, \AA) between
212 the atomic coordinates of the iBP solution and of the initial PDB conformation for the one-shot runs using the definitions of stereochemistry described in Table 1. In the present work,
213 the RMSD values were calculated on the heavy atoms of the protein backbone. In the case
214 where the stereochemistry is defined from the initial PDB conformation (pdb stereochemistry
215 in Table 1), coordinate RMSD around 0.5 \AA are observed (Figure 4a). This provides a floor
216 value for the maximum possible precision which can be obtained using the discretization of
217 the ϕ and ψ intervals in iBP. It is interesting to note that if the bond lengths are also taken
218 variable from the PDB conformation, the same distribution of RMSD values is obtained (data
219

220 not shown). The bond length variations have thus much less influence on the variations of
221 conformations obtained by iBP than the bond angle variations.

222 As soon as the ω angle is set to 178° while getting the other parameter values from
223 pdb stereochemistry (Figure 4b), the RMSD distributions are switched towards much larger
224 values, up to 10-12 Å. Such behavior is also observed for Hollingsworth (Figure 4c,d) or for
225 uniform (Figure 4e) stereochemistry. Interestingly, quite different RMSD distributions are
226 observed according to the type of secondary structures. The shift is smaller for proteins folded
227 mostly as α helices (blue curves) or mostly as loops (red curves), producing an RMSD value
228 smaller than 3 Å for at least half of the structures. By contrast, the structures containing
229 mostly β strands (green curves) display distributions centered at RMSD values between 5
230 and 6 Å.

231 The coordinate RMSD values are known to display some limitations for precisely mea-
232 suring the accuracy of a protein structure prediction.⁵³ Thus, the TM score distribution⁵⁴
233 have been calculated (Figure S3) using the software code downloaded from [zhanggroup.](http://zhanggroup.org/TM-score)
234 [org/TM-score](http://zhanggroup.org/TM-score). For pdb stereochemistry (Figure S3a), the TM scores are close to the opti-
235 mal value of 1. Similarly to Figure 4, the TM-score values are getting worse if ω values of
236 178° are used (Figure S3b,d) or in the case of uniform stereochemistry (Figure S3e). Most of
237 the calculated conformations display TM-scores larger than 0.5, if the bond angles are taken
238 from the PDB entry and the ω values are set equal to 178° (Figure S3b).

239 An interesting difference between the TM score and RMSD is observed for Hollingsworth
240 stereochemistry. Indeed, the TM scores are worse for Hollingsworth stereochemistry (Fig-
241 ure S3c,d) than for any other calculation, whereas the RMSD values are similar between
242 Hollingsworth (Figure 4c,d) and uniform (Figure 4e) stereochemistry. This is probably due
243 to a distortion in interatomic distance distribution, produced by the use of bond and ω
244 angles averaged on Hollingsworth regions. Indeed, this distortion deteriorates the TM score
245 value as the distance distribution is the main ingredient of the score. Additional distance
246 distortions may arise from the use of the van der Waals scaling of $\rho = 0.8$ used during the

247 iBP calculation to avoid pruning of conformations.

248 To investigate more precisely the relationship between stereochemistry variations and the
249 efficiency in conformer generation, the global variation of bond angles along a structure has
250 been calculated as:

$$\Delta\theta = \sum_{i=1}^{N-1} |\theta_{i+1} - \theta_i| \quad (1)$$

251 where $|\cdot|$ stands for the absolute value, and N is the number of residues with residue number
252 indexed from 1 to N . A similar global variation for the torsion angle ω was defined as:

$$\Delta\omega = \sum_{i=1}^{N-1} |\delta\omega_{i+1} - \delta\omega_i| \quad (2)$$

253 where:

$$\delta\omega = \begin{cases} 180^\circ - \omega & \text{if } \text{sgn}(\omega) > 0, \\ -180^\circ - \omega & \text{if } \text{sgn}(\omega) < 0, \end{cases} \quad (3)$$

254 with $\text{sgn}(\cdot)$ being the sign function, i.e., $\text{sgn}(x) = 1$ if $x > 0$ or $\text{sgn}(x) = -1$ if $x < 0$.

255 The $\Delta\theta$ values calculated on bond angles C-N-C $_{\alpha}$, N-C $_{\alpha}$ -C and C $_{\alpha}$ -C-N and the $\Delta\omega$
256 values calculated on torsion angle ω were compared to the coordinate RMSD values between
257 the iBP and initial conformations (Figure S4). The global variations and the coordinate
258 RMSD display an obvious correlation which is also driven by the length of the protein
259 chains. In agreement with Figure 3, the largest global variations are obtained for $\Delta\omega$ (blue
260 points) and $\Delta\theta$ of the N-C $_{\alpha}$ -C bond angle (green points).

261 For the calculations performed using: (i) Hollingsworth stereochemistry (Figure 4c) and
262 (ii) uniform stereochemistry (Figure 4e), the protocol Relax⁴⁸ of Rosetta⁴⁹ was applied on
263 the iBP outputs, to add the residue sidechains. The minimal RMSD value with respect
264 to the initial PDB conformations (Figure S5a,b) shifts towards smaller values which is the
265 sign of a conformation drift towards the correct solution. Indeed, the comparison of RMSD

266 distribution with Hollingsworth (Figure S5a versus Figure 4c) and uniform (Figure S5b
267 versus Figure 4e) stereochemistry reveals a shift of 1-2 Å and even of 4 Å for the mostly β
268 folded proteins (green curve). The Rosetta scores have been also plotted along the coordinate
269 RMSD and display a similar variation towards more negative values for smaller RMSD values
270 (Figure S5c,d).

271 The iBP procedure presented here for reconstructing a protein complete fold could also
272 have an application for the reconstruction of missing parts of a given protein structure. To
273 evaluate this approach, the sub-chains for which coordinate RMSD to initial protein structure
274 was smaller than 2.5 Å were extracted and their lengths are plotted as the percentage of the
275 length of the full chain (Figure S6a,b,c) as well as numbers of residues (Figure S6d,e,f). The
276 distribution of the percentages (Figure S6a,b,c) agrees with the distribution of RMSD values
277 (Figure 4), with percentages close to 100% when RMSD values close to 0.5 Å are observed. As
278 soon as the stereochemistry becomes less variable, the distributions of percentages become
279 wider, but display very similar shapes in all runs, with two maxima located around 50%
280 and 90% (Figure S6b,c). The distribution of the numbers of residues are all larger than 30
281 residues and are mostly distributed in the range of 20-60 amino acids. These values compare
282 well with the results of the literature.⁵⁵ In addition, similar distributions are observed for the
283 different types of secondary structures in protein folds. These results are quite encouraging
284 in the perspective of reconstructing non visible regions in protein structures.

285 In the presence of Hollingsworth stereochemistry, the effect of different input ω values
286 on the reconstruction of protein folds was analyzed (Figure 5). If exact values are known
287 for the ω torsion angles (Figure 5a), the majority of structures containing mostly α helices
288 or loops display RMSD values smaller than 3 Å, corresponding to a good reconstruction of
289 the protein fold. On the other hand, the structures containing mostly β strands display a
290 shift in RMSD values, but their RMSD is still mostly smaller than 3 Å. Thus, knowing the
291 exact values for torsion angles ω is essential for building the protein structure from local
292 information.

293 Then, the effect of several discretizations of ω was tested on the reconstruction of protein
294 structures. In that case, the ω continuous values are replaced by ω_k values corresponding
295 to different discretization classes k . In the first discretization, the absolute value of the
296 parameter $\delta\omega$ previously introduced in Eq 3 was used to define four classes of ω values:

$$\omega_k = \begin{cases} 173^\circ \text{sgn}(\omega) & \text{if } 5^\circ < |\delta\omega| < 10^\circ, \\ 177^\circ \text{sgn}(\omega) & \text{if } |\delta\omega| < 5^\circ \end{cases} \quad (4)$$

297 This discretization induces a shift in RMSD values (Figure 5b). The mostly α and loop
298 structures are still correctly reconstructed, and half of the mostly β structures display RMSD
299 values larger than 3 Å.

300 A more crude discretization is used where ω_k is set equal to $178^\circ \text{sgn}(\omega)$. This two-
301 class discretization (Figure 5c) shifts the RMSD distribution to values larger than 3 Å for β
302 structures, but about the two third of α and one-half of loop structures display RMSD values
303 smaller than 3 Å. But, even this crude discretization allows us to obtain better RMSD values
304 than those observed for uniform stereochemistry (Figure 4e). The effect of ω discretization
305 on the fold reconstruction proves that classification approaches⁵⁶ could be interesting for
306 predicting protein conformations.

307 **Effect of the enumeration by iBP to the reconstruction of protein** 308 **fold**

309 The iBP approach has the advantage of allowing a systematic enumeration of all possible
310 solutions. This enumerating scheme was thus used here to improve the coordinate RMSD
311 of solutions with respect to the initial PDB structure. The inputs of the iBP runs were
312 intervals around the ϕ and ψ angles with intervals widths of 5° in α helices and β strands,
313 and of 40° in other protein regions. The number of branches is 4.

314 A disadvantage of the iBP approach is that execution can take a very long time and
315 ultimately prune all solutions. In order to quickly determine input values avoiding the full

316 pruning of solutions, short iBP runs were launched with an upper limit of 2 minutes, varying
317 systematically the values of the discretization factors ϵ and of the van der Waals scaling
318 ρ . Two stereochemistry inputs were used: uniform and Hollingsworth stereochemistry. A
319 conformation was stored only if the coordinate RMSD between the newly generated and the
320 previous solution was smaller than 3.5 Å.

321 The number of accepted solutions is mostly around 10^4 and increases up to 10^5 (Fig-
322 ure S7a). Around 20% of the calculations display no solutions. The number of solutions
323 rejected because of the RMSD criterion (Figure S7b) is in the range of 10^6 - 10^7 , much larger
324 than the range of accepted solutions. An RMSD of 3.5 Å is thus quite discriminating for
325 selecting solutions. The tree size is mostly in the range of 10^5 to 10^{10} (Figure S7c). Inter-
326 estingly, previous experiments realized with TAI BP showed²⁹ that a tree size of about 10^9
327 permits systematic enumeration of the tree solutions for protein fragments. The size of the
328 trees, as well as the numbers of accepted and rejected solutions, display the same distribution
329 for the Hollingsworth or the uniform stereochemistry. Similarly, the discretization factors ϵ
330 vary uniformly in the range 0.15-0.17 (Figure S7d) for Hollingsworth (black curve) as well as
331 for uniform (red curve) stereochemistry. By contrast, the van der Waals scaling ρ varies in
332 the 0.2-0.6 range for Hollingsworth stereochemistry and in the 0.3-0.6 range for the uniform
333 stereochemistry (Figure S7e). This shows that smaller ρ values were sometimes used to avoid
334 pruning in the case of Hollingsworth stereochemistry and agrees with the worse TM score
335 observed for the one-shot run with Hollingsworth stereochemistry (Figure S2).

336 Based on the fast exploration described above, input values for enumerating runs were
337 selected using the following rules: (i) the largest possible van der Waals scaling ρ for max-
338 imizing the pruning by steric hindrance, (ii) the largest possible discretization factor ϵ for
339 obtaining the smallest possible tree size to facilitate its full exploration. The corresponding
340 trees were then completely parsed using iBP. During the enumeration, the number of asked
341 conformations was set to 10^9 . All calculations produced a smaller number of conformations,
342 which proves that the corresponding trees were fully explored. Tree sizes centered around

343 10^4 , discretization factors ϵ around 0.17, and van der Waals scaling factors ρ around 0.5 were
344 used for these full runs (Figure S8). The discretization factor displays similar distributions
345 for Hollingsworth and uniform stereochemistry. In contrast, the tree size and the van der
346 Waals scaling factor ρ are slightly shifted towards higher values for Hollingsworth stereo-
347 chemistry. Indeed, the larger tree observed for this stereochemistry requires greater van der
348 Waals scaling to reduce the number of solutions by pruning.

349 The effect of the enumerating scheme for calculating structures was evaluated using
350 the distribution of coordinate RMSD between iBP and PDB target conformations (Figure
351 6). For each processed protein, the smallest RMSD value between the iBP solution and the
352 initial structure was selected and the corresponding RMSD distribution was compared to the
353 corresponding RMSD distributions for the one-shot runs (Figure 4c,e). For both uniform
354 (Figure 6a) and Hollingsworth (Figure 6b) stereochemistry, the use of enumeration induces a
355 shift of the RMSD values towards smaller values. Interestingly, this shift is more pronounced
356 in the case of uniform stereochemistry, as shown by the comparison of Figures 6b and 4e.
357 Thus, using the enumeration of conformations potentially improves the efficiency of the fold
358 reconstruction.

359 **A possible origin of the variability of stereochemistry**

360 During the previous sections, the effect of variability of stereochemistry on the calculation of
361 protein conformation based on local conformational restraints has been examined in various
362 situations. In this section, we intend to investigate the relationship between the distribution
363 of synonymous DNA codons and the variability of stereochemistry.

364 We first focused on the variability of bond angle values. The standard genetic code⁵⁷
365 was used to determine the number of synonymous codons for each amino acid residue. The
366 number of possible synonymous codons for each residue was summed along each of the 308
367 protein primary sequences to produce the cumulative number of synonym codons. Plotting
368 the global variations $\Delta\theta$ of the bond angles C–N–C $_{\alpha}$, N–C $_{\alpha}$ –C and C $_{\alpha}$ –C–N compared

369 to this cumulative number (Figure S9) reveals a correlation between the stereochemistry
370 variation and the number of synonymous codons similar to those previously observed in
371 Figure S4. As in Figure S4, the correlation is driven by the protein size. The 13 proteins
372 from *E coli* and expressed in *E coli* for structure determination are marked with green crosses
373 and display the same tendency as the whole set of proteins.

374 These 13 *E coli* proteins are drawn in cartoon and the residues displaying global variations
375 $\Delta\theta$ of bond angles larger than 6° are drawn in licorice and colored in green (Figure S10).
376 Most of these protein structures display a topology inducing interactions between secondary
377 structure elements located apart in the protein sequence. Also, the residues with the largest
378 local variation of bond angles are mostly located in loops or at the extremity of secondary
379 structure elements. In several structures (1C4Q, 1GYX, 1Q5Y, 3CCD, 4MAK, 4Q2L), most
380 variable residues are close to each other in the 3D structure, displaying even long-range
381 physico-chemical interactions. The variations of bond angle stereochemistry can be thus
382 related to the long-range interactions participating to the fold definition. The positions of
383 variable residues in the loops might be related to the importance of loop conformations for
384 orienting the protein backbone with the folded topology. In addition, the long-range inter-
385 actions of some variable residues suggest a cooperative effect between bond angle variations
386 arising during the protein folding.

387 In the second step, we focused on the relationship between the ω torsion angle variability
388 and the individual corresponding DNA sequences. Among the 308 protein structures, the
389 proteins issued from the organisms *Homo sapiens*, *Escherichia coli* and *Saccharomyces cere-*
390 *visiae* were selected using the descriptor SOURCE: ORGANISM_SCIENTIFIC. The relative
391 codon usage observed in three organisms: *Escherichia coli*, *Saccharomyces cerevisiae* and
392 *Homo sapiens* (Table 1 of Ref⁵⁸) was used to extract the different numbers of synonymous
393 codons for each amino-acid residue of these proteins. The PDB entries were then entered as a
394 query to the European Nucleotide Archive (ENA) www.ebi.ac.uk/ena. The corresponding
395 DNA sequences were programmatically downloaded and filtered to keep those corresponding

396 to the considered protein chain in the PDB entry.

397 The DNA sequence codons were then analyzed using the statistics on codons from Ref⁵⁸
398 calculated on the organisms *Homo sapiens*, *Escherichia coli* and *Saccharomyces cerevisiae*.
399 From each amino acid, the codons displaying statistics of presence smaller than the average
400 presence of all codons coding for the amino acid were considered rare codons. Then, the ω
401 angle values of all protein residues were analyzed (Figure 7) by calculating their average μ
402 and standard deviation σ^2 values on each considered protein sequence. The ω angle values
403 were centered and normalized using μ and σ^2 , producing a global averaged ω value on each
404 protein equal to zero. The ω values averaged on protein residues corresponding to rare
405 codons, as well as to protein residues corresponding to neighbors or second-neighbors of
406 rare codons were centered and normalized using the μ and σ^2 values obtained for the full
407 sequence of the corresponding PDB entry. The distributions of these centered and normalized
408 ω values display slight shifts towards positive or negative values (Figure 7a). Looking at these
409 distributions, the ω values are more apart from zero for residues corresponding to rare codons
410 than for neighbor and second-neighbor residues (Figure 7a). All standard deviation values
411 (Figure 7b) are centered around 1° , similarly to the standard deviation values normalized
412 on the whole primary sequence.

413 The rare codons have been pointed out to be related to the kinetics of protein folding
414 during the protein synthesis in the ribosome.^{57,59,60} In addition, recent bioinformatics analysis
415 has established a relation between the genetic code and the protein structure.^{35,36} In that
416 frame, the relationship put in evidence here between the variability of ω values, the rare
417 codons, and the reconstruction of the protein structure connects the protein folding and the
418 kinetics of protein synthesis in the ribosome.

419 In that respect, it is interesting to observe that mostly β folded proteins are specifically
420 sensitive to the variability of ω values. Indeed, their folding requires intricate cooperation
421 between the establishment of long-range interactions forming the β sheets. This may be
422 related to the analysis of Figure S10 performed above.

423 The analyses performed here point out the importance of mRNA in the variability of
424 stereochemistry in proteins. They complement the relationships put in evidence in the
425 literature³⁵ between the mRNA sequence and populations of α and β regions, as we have
426 also shown here that the variations of stereochemistry are related to the Hollingsworth regions
427 of the Ramachandran diagram (Figure 3).

428 Discussion

429 The present work has been investigating the exclusive use of local conformational informa-
430 tion, namely the values of the torsion angles ϕ and ψ for calculating protein conformations.
431 The results obtained here were made possible in an essential way by the development of
432 the interval Branch-and-Prune approach (iBP),⁴⁰ providing a framework for the systematic
433 enumeration of conformations. The analyses performed here have put in evidence the essen-
434 tial impact of the variability in stereochemistry and represent, to the best of our knowledge,
435 the first attempt to relate these stereochemistry aspects to the calculation and prediction of
436 protein conformations.

437 The variations of stereochemistry are certainly influenced by the refinement protocols
438 used for determining X-ray crystallographic structures, in which the application of long-
439 range restraints can produce effects in variations of local stereochemistry, in a way that is
440 not mastered in the details. During the last decades, the stereochemistry aspects have not
441 been taken into account during the protein structure prediction thanks to the use of long-
442 range distance/angle restraints.⁶¹ On the other side, the use of long-range restraints might
443 influence the appearance of stereochemistry outliers. The relative weights of the different
444 types of information in the protein structure calculation should be further investigated for
445 example using a Bayesian approach.^{62,63}

446 To alleviate the impact of variability, a Ramachandran-based definition of the bond
447 angle stereochemistry, the Hollingsworth definition, has been proposed. The efficiency of

448 this definition is improved with the use of enumeration during the iBP approach or by the
449 knowledge of ω values. The combination of these aspects provides thus a way to overcome
450 the variability problem for most of the protein structures examined here, especially in the
451 case of α proteins.

452 The calculations performed here have been scored with respect to reference protein con-
453 formation, using coordinate RMSD and TM-score. In most of the calculations, TM-scores
454 display better values than RMSD, in agreement with the general knowledge on this score.⁵⁴
455 But, if Hollingsworth stereochemistry is used, better RMSD values are obtained than the
456 TM-score values, probably because the deformation of local stereochemistry impacts the
457 distribution of inter-atomic distances used in the TM-score calculations. Indeed, the TM-
458 score was derived to correct the bias of coordinate RMSD on structures determined in the
459 framework of uniform stereochemistry and should be adapted to the case of Hollingsworth
460 stereochemistry.

461 Two approaches have been used to reduce the conformational drift produced by the
462 lack of precision in the modeling of stereochemistry: the enumeration of conformation in
463 the framework of iBP, and the Relax procedure⁴⁸ of Rosetta.⁴⁹ Both approaches permit to
464 improve the results.

465 The analyses carried out here make it possible to propose that the origin of stereochemical
466 variations could be linked to the information contained in the mRNA sequence. The finer
467 investigation of this aspect is out of the scope of the present work but could provide a more
468 integrated modeling of protein structure and folding.

469 **Acknowledgements**

470 CNRS, Lorraine University, IRISA, and ANR PRCI multiBioStruct (ANR-19-CE45-0019)
471 are acknowledged for funding. High Performance Computing resources were provided by the
472 EXPLOR Centre at Lorraine University (2022CPMXX2687).

473 Data and Software Availability

474 The version of iBP²⁸ modified to handle variable stereochemistry is available at: `github.`
475 `com/tmalliavin/ibp-ng-fullchain`. For the other software, not developed by the authors,
476 the literature references are given.

477 References

- 478 (1) Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment
479 of methods of protein structure prediction (CASP)-Round XV. *Proteins* **2023**, *91*,
480 1539–1549.
- 481 (2) Lupas, A.; Pereira, J.; Alva, V.; Merino, F.; Coles, M.; Hartmann, M. The breakthrough
482 in protein structure prediction. *Biochem J* **2021**, *478*, 1885–1890.
- 483 (3) Jisna, V.; Jayaraj, P. Protein Structure Prediction: Conventional and Deep Learning
484 Perspectives. *Protein J* **2021**, *40*, 522–544.
- 485 (4) Senior, A. W. et al. Improved protein structure prediction using potentials from deep
486 learning. *Nature* **2020**, *577*, 706–710.
- 487 (5) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*
488 **2021**, *596*, 583–589.
- 489 (6) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-
490 track neural network. *Science* **2021**, *373*, 871–876.
- 491 (7) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Ka-
492 bel, O.; Shmueli, Y.; Dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.;
493 Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a lan-
494 guage model. *Science* **2023**, *379*, 1123–1130.

- 495 (8) Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat Rev*
496 *Mol Cell Biol* **2019**, *20*, 681–697.
- 497 (9) Skolnick, J.; Kolinski, A.; Ortiz, A. R. MONSSTER: a method for folding globular
498 proteins with a small number of distance restraints. *J Mol Biol* **1997**, *265*, 217–241.
- 499 (10) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct
500 residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci*
501 *U S A* **2009**, *106*, 67–72.
- 502 (11) Szurmant, H.; Weigt, M. Inter-residue, inter-protein and inter-family coevolution:
503 bridging the scales. *Curr Opin Struct Biol* **2018**, *50*, 26–32.
- 504 (12) Russ, W. P.; Figliuzzi, M.; Stocker, C.; Barrat-Charlaix, P.; Socolich, M.; Kast, P.;
505 Hilvert, D.; Monasson, R.; Cocco, S.; Weigt, M.; Ranganathan, R. An evolution-based
506 model for designing chorismate mutase enzymes. *Science* **2020**, *369*, 440–445.
- 507 (13) Mortuza, S. M.; Zheng, W.; Zhang, C.; Li, Y.; Pearce, R.; Zhang, Y. Improving
508 fragment-based ab initio protein structure assembly using low-accuracy contact-map
509 predictions. *Nat Commun* **2021**, *12*, 5011.
- 510 (14) Sayers, E. et al. Database resources of the national center for biotechnology information.
511 *Nucleic Acids Research* **2022**, *50*, D20–D26.
- 512 (15) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.;
513 Bourne, P. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.
- 514 (16) Kotowski, K.; Smolarczyk, T.; Roterman-Konieczna, I.; Stapor, K. ProteinUnet-An effi-
515 cient alternative to SPIDER3-single for sequence-based prediction of protein secondary
516 structures. *J Comput Chem* **2021**, *42*, 50–59.
- 517 (17) Moffat, L.; Jones, D. T. Increasing the accuracy of single sequence prediction methods
518 using a deep semi-supervised learning framework. *Bioinformatics* **2021**, *37*, 3744–3751.

- 519 (18) Singh, J.; Litfin, T.; Paliwal, K.; Singh, J.; Hanumanthappa, A. K.; Zhou, Y. SPOT-1D-
520 Single: improving the single-sequence-based prediction of protein secondary structure,
521 backbone angles, solvent accessibility and half-sphere exposures using a large training
522 set and ensembled deep learning. *Bioinformatics* **2021**, *37*, 3464–3472.
- 523 (19) Singh, J.; Paliwal, K.; Litfin, T.; Singh, J.; Zhou, Y. Reaching alignment-profile-based
524 accuracy in predicting protein secondary and tertiary structural properties without
525 alignment. *Sci Rep* **2022**, *12*, 7607.
- 526 (20) Peracha, O. PS4: a next-generation dataset for protein single-sequence secondary struc-
527 ture prediction. *Biotechniques* **2024**, *76*, 63–70.
- 528 (21) Warnow, T. Revisiting Evaluation of Multiple Sequence Alignment Methods. *Methods*
529 *Mol Biol* **2021**, *2231*, 299–317.
- 530 (22) Radjasandirane, R.; de Brevern, A. G. Structural and Dynamic Differences between
531 Calreticulin Mutants Associated with Essential Thrombocythemia. *Biomolecules* **2023**,
532 *13*.
- 533 (23) Hengeveld, S. B.; Malliavin, T.; Lin, J.; Liberti, L.; Mucherino, A. A Study on the
534 Impact of the Distance Types Involved in Protein Structure Determination by NMR.
535 *Computational Structural Bioinformatics Workshop (CSBW21), IEEE International*
536 *Conference on Bioinformatics and Biomedicine (BIBM21)* **2021**, 2502–2510.
- 537 (24) Hengeveld, S. B.; Merabti, M.; Pascale, F.; Malliavin, T. E. A Study on the Covalent
538 Geometry of Proteins and Its Impact on Distance Geometry. 6th International Con-
539 ference on Geometric Science of Information (GSI'23). Saint Malo, France, 2023; pp
540 520–530.
- 541 (25) Lavor, C.; Liberti, L.; Maculan, N.; Mucherino, A. The Discretizable Molecular Distance
542 Geometry Problem. *Computational Optimization and Applications* **2012**, *52*, 115–146.

- 543 (26) Cassioli, A.; Bardiaux, B.; Bouvier, G.; Mucherino, A.; Alves, R.; Liberti, L.; Nilges, M.;
544 Lavor, C.; Malliavin, T. An algorithm to enumerate all possible protein conformations
545 verifying a set of distance constraints. *BMC Bioinformatics* **2015**, *16*, 23–37.
- 546 (27) D’Ambrosio, C.; Vu, K.; Lavor, C.; Liberti, L.; Maculan, N. New Error Measures and
547 Methods for Realizing Protein Graphs from Distance Data. *Discrete & Computational*
548 *Geometry* **2017**, *57*, 371–418.
- 549 (28) Worley, B.; Delhommel, F.; Cordier, F.; Malliavin, T.; Bardiaux, B.; Wolff, N.;
550 Nilges, M.; Lavor, C.; Liberti, L. Tuning interval Branch-and-Prune for protein struc-
551 ture determination. *Journal of Global Optimization* **2018**, *72*, 109–127.
- 552 (29) Malliavin, T. E.; Mucherino, A.; Lavor, C.; Liberti, L. Systematic Exploration of Pro-
553 tein Conformational Space Using a Distance Geometry Approach. *J Chem Inf Model*
554 **2019**, *59*, 4486–4503.
- 555 (30) Malliavin, T. E. Tandem domain structure determination based on a systematic enu-
556 meration of conformations. *Sci Rep* **2021**, *11*, 16925.
- 557 (31) Förster, D.; Idier, J.; Liberti, L.; Mucherino, A.; Lin, J. H.; Malliavin, T. E. Low-
558 resolution description of the conformational space for intrinsically disordered proteins.
559 *Sci Rep* **2022**, *12*, 19057.
- 560 (32) Huang, S.-Y.; Chang, C.-F.; Lin, J.-H.; Malliavin, T. In *Geometric Science of Informa-*
561 *tion : 6th International Conference, GSI 2023, St. Malo, France, August 30 – Septem-*
562 *ber 1, 2023, Proceedings, Part II*; Nielsen, F., Barbaresco, F., Eds.; Lecture Notes in
563 Computer Science; Springer Nature Switzerland, 2023; Vol. 14072; pp 531–540.
- 564 (33) Hollingsworth, S. A.; Lewis, M. C.; Berkholz, D. S.; Wong, W. K.; Karplus, P. A.
565 (ϕ, ψ)₂ motifs: a purely conformation-based fine-grained enumeration of protein
566 parts at the two-residue level. *J. Mol. Biol.* **2012**, *416*, 78–93.

- 567 (34) Engh, R.; Huber, R. Accurate bond and angle parameters for X-ray protein structure
568 refinement. *Acta Crystallogr A* **1991**, *47*, 392–400.
- 569 (35) Rosenberg, A. A.; Marx, A.; Bronstein, A. M. Codon-specific Ramachandran plots show
570 amino acid backbone conformation depends on identity of the translated codon. *Nat*
571 *Commun* **2022**, *13*, 2815.
- 572 (36) Ackerman-Schraier, L.; Rosenberg, A. A.; Marx, A.; Bronstein, A. M. Machine learning
573 approaches demonstrate that protein structures carry information about their genetic
574 coding. *Sci Rep* **2022**, *12*, 21968.
- 575 (37) Rosenberg, A. A.; Yehishalom, N.; Marx, A.; Bronstein, A. M. An amino-domino model
576 described by a cross-peptide-bond Ramachandran plot defines amino acid pairs as local
577 structural units. *Proc Natl Acad Sci U S A* **2023**, *120*, e2301064120.
- 578 (38) Wang, G.; Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics*
579 **2003**, *19*, 1589–1591.
- 580 (39) Lavor, C.; Liberti, L.; Mucherino, A. The interval Branch-and-Prune algorithm for
581 the discretizable molecular distance geometry problem with inexact distances. *J Glob*
582 *Optim* **2013**, *56*, 855–871.
- 583 (40) Liberti, L.; Lavor, C.; Mucherino, A. The discretizable molecular distance geometry
584 problem seems easier on proteins. *Distance Geometry: Theory, Methods and Applica-*
585 *tions. Mucherino, Lavor, Liberti, Maculan (eds.)* **2014**, 47–60.
- 586 (41) Liberti, L.; Lavor, C.; Maculan, N.; Mucherino, A. Euclidean Distance Geometry and
587 Applications. *SIAM Rev* **2014**, *56*, 3–69.
- 588 (42) Lavor, C.; Alves, R.; Figueiredo, W.; Petraglia, A.; Maculan, N. Clifford Algebra and
589 the Discretizable Molecular Distance Geometry Problem. *Adv. Appl. Clifford Algebras*
590 **2015**, *25*, 925–942.

- 591 (43) Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. ARIA: automated NOE assignment
592 and NMR structure calculation. *Bioinformatics* **2003**, *19*, 315–316.
- 593 (44) Hollingsworth, S. A.; Karplus, P. A. A fresh look at the Ramachandran plot and the
594 occurrence of standard structures in proteins. *Biomol Concepts* **2010**, *1*, 271–283.
- 595 (45) Brunger, A.; Adams, P.; Clore, M.; Delano, W.; Gros, P.; Grosse-Kunstleve, W.;
596 Jiang, J.-S.; Nilges, M.; Pannu, N.; Read, R.; Rice, L.; Simonson, T.; Warren, G.
597 Crystallography & NMR System: A New Software Suite for Macromolecular Structure.
598 *Acta Cryst* **1998**, *D54*, 905–921.
- 599 (46) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Dotson, D.;
600 Domanski, J.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: A Python package
601 for the rapid analysis of molecular dynamics simulations. *Proceedings of the 15th Python*
602 *in Science Conference, Austin, TX, 2016* **2016**, *32*, 102–109.
- 603 (47) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Pro-*
604 *teins* **1995**, *23*, 566–579.
- 605 (48) Nivón, L. G.; Moretti, R.; Baker, D. A Pareto-optimal refinement method for protein
606 design scaffolds. *PLoS One* **2013**, *8*, e59004.
- 607 (49) Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Mod-
608 eling and Design. *J Chem Theory Comput* **2017**, *13*, 3031–3048.
- 609 (50) Balasco, N.; Esposito, L.; De Simone, A.; Vitagliano, L. Local Backbone Geome-
610 try Plays a Critical Role in Determining Conformational Preferences of Amino Acid
611 Residues in Proteins. *Biomolecules* **2022**, *12*, 1184.
- 612 (51) Woolfson, D. N.; Williams, D. H. The influence of proline residues on alpha-helical
613 structure. *FEBS Lett* **1990**, *277*, 185–188.

- 614 (52) Krieger, F.; Moglich, A.; Kiefhaber, T. Effect of proline and glycine residues on dynam-
615 ics and barriers of loop formation in polypeptide chains. *J Am Chem Soc* **2005**, *127*,
616 3346–3352.
- 617 (53) Betancourt, M. R.; Skolnick, J. Universal similarity measure for comparing protein
618 structures. *Biopolymers* **2001**, *59*, 305–309.
- 619 (54) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure
620 template quality. *Proteins* **2004**, *57*, 702–710.
- 621 (55) Barozet, A.; Chacón, P.; Cortés, J. Current approaches to flexible loop modeling. *Curr*
622 *Res Struct Biol* **2021**, *3*, 187–191.
- 623 (56) Guermeur, Y. A generic model of multi-class support vector machine. *International*
624 *Journal of Intelligent Information and Database Systems (IJIIDS)* **2012**, *6*, 555–577.
- 625 (57) Chaney, J.; Clark, P. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu*
626 *Rev Biophys* **2015**, *44*, 143–166.
- 627 (58) Komar, A. A. The Yin and Yang of codon usage. *Hum Mol Genet* **2016**, *25*, R77–R85.
- 628 (59) Liu, Y.; Yang, Q.; Zhao, F. Synonymous but Not Silent: The Codon Usage Code for
629 Gene Expression and Protein Folding. *Annu Rev Biochem* **2021**, *90*, 375–401.
- 630 (60) Buhr, F.; Jha, S.; Thommen, M.; Mittelstaet, J.; Kutz, F.; Schwalbe, H.; Rod-
631 nina, M. V.; Komar, A. A. Synonymous Codons Direct Cotranslational Folding toward
632 Different Protein Conformations. *Mol Cell* **2016**, *61*, 341–351.
- 633 (61) Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved
634 protein structure prediction using predicted interresidue orientations. *Proc Natl Acad*
635 *Sci U S A* **2020**, *117*, 1496–1503.
- 636 (62) Habeck, M.; Rieping, W.; Nilges, M. Weighting of experimental evidence in macro-
637 molecular structure determination. *Proc Natl Acad Sci U S A* **2006**, *103*, 1756–1761.

638 (63) Bernard, A.; Vranken, W. F.; Bardiaux, B.; Nilges, M.; Malliavin, T. E. Bayesian
639 estimation of NMR restraint potential and weight: a validation on a representative set
640 of protein structures. *Proteins* **2011**, *79*, 1525–1537.

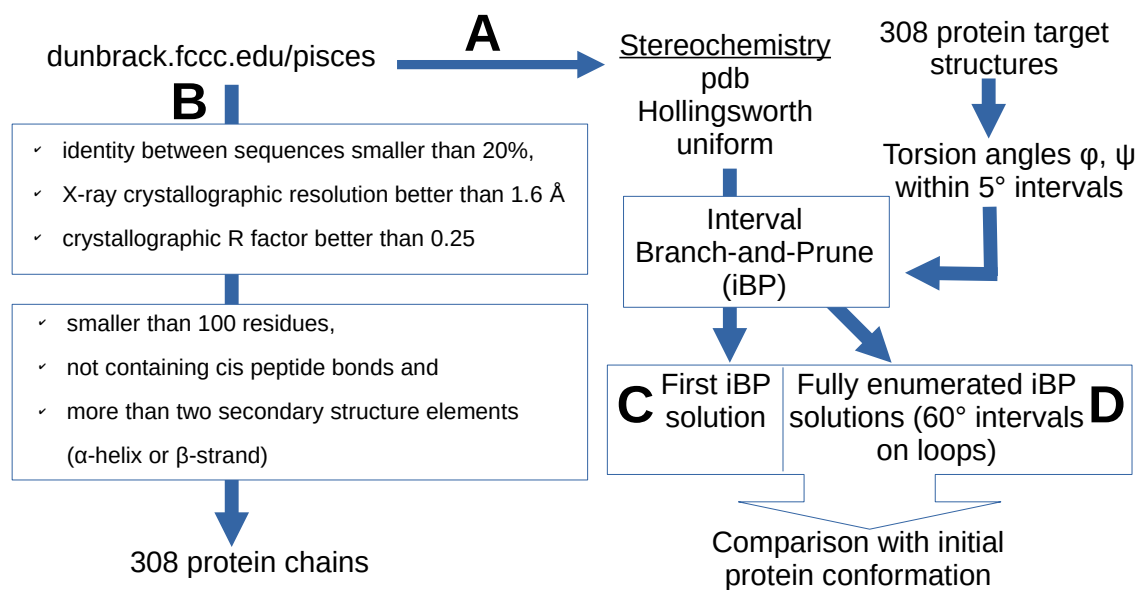


Figure 1: Flow-chart of the calculations. A. Obtaining the statistics of stereochemistry from a protein database. B. Generating a subset of 308 protein chains which will be the targets for iBP calculations. Using the torsion angle values measured on the protein target conformations along with different hypotheses on stereochemistry (Table 1), protein conformations were recalculated using iBP, selecting the first generated conformation (one-shot iBP run) (C) or enumerating all possible conformations (D).

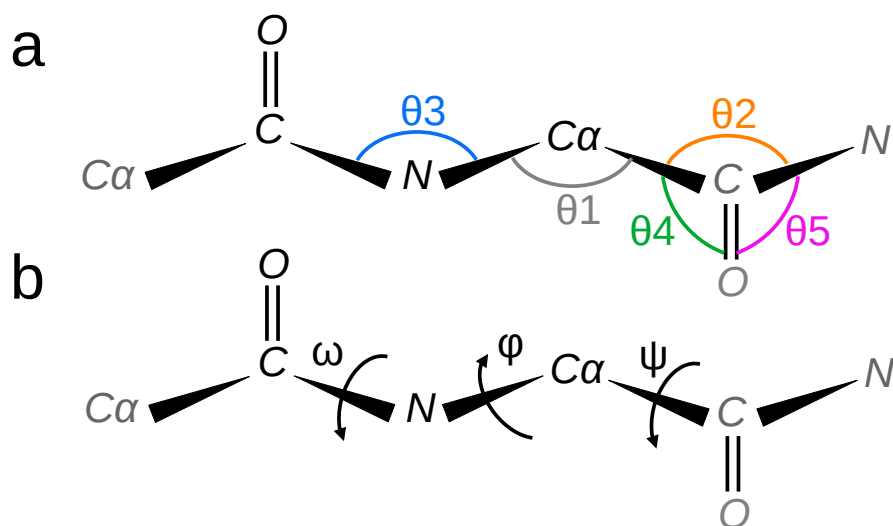


Figure 2: Scheme of the succession of protein backbone heavy atoms N, C_α , C, and O along with definitions of angle parameters. a. Bond angles are: $N-C_\alpha-C$ (θ_1 , grey), $C_\alpha-C-N$ (θ_2 , orange) and $C-N-C_\alpha$ (θ_3 , blue), $C_\alpha-C-O$ (θ_4 , green) and $O-C-N$ (θ_5 , magenta). b. The backbone torsion angles ϕ , ψ , and ω are indicated by circular arrows.

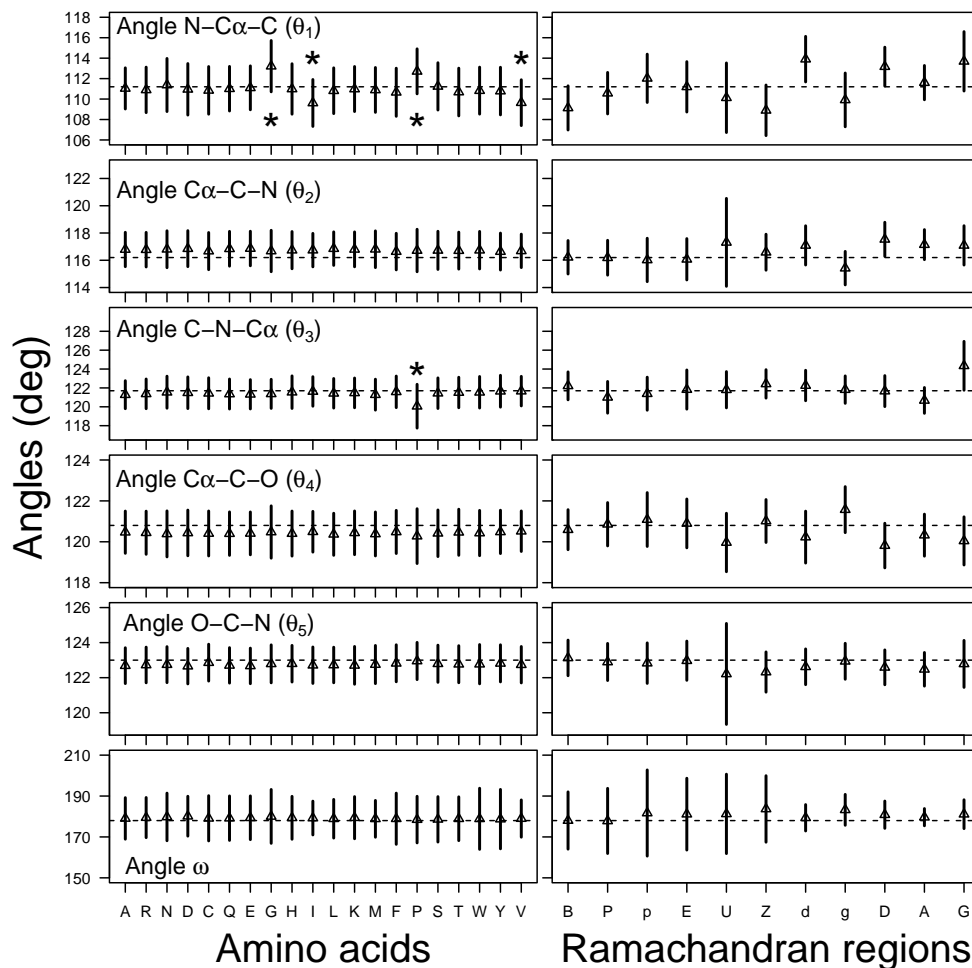


Figure 3: Average and standard deviation values calculated on the bond angles and ω dihedral angle, defining the stereochemistry of protein backbone. The bond angles labels are the same than those displayed on Figure 2. The regions of the Ramachandran diagram were taken from Ref³³ and are displayed in Figure S2. The dashed lines correspond to the angle values in the parameter set of Engh and Huber³⁴ (Table S2). Asterisk indicate the most variable bond angles along the amino acid type.

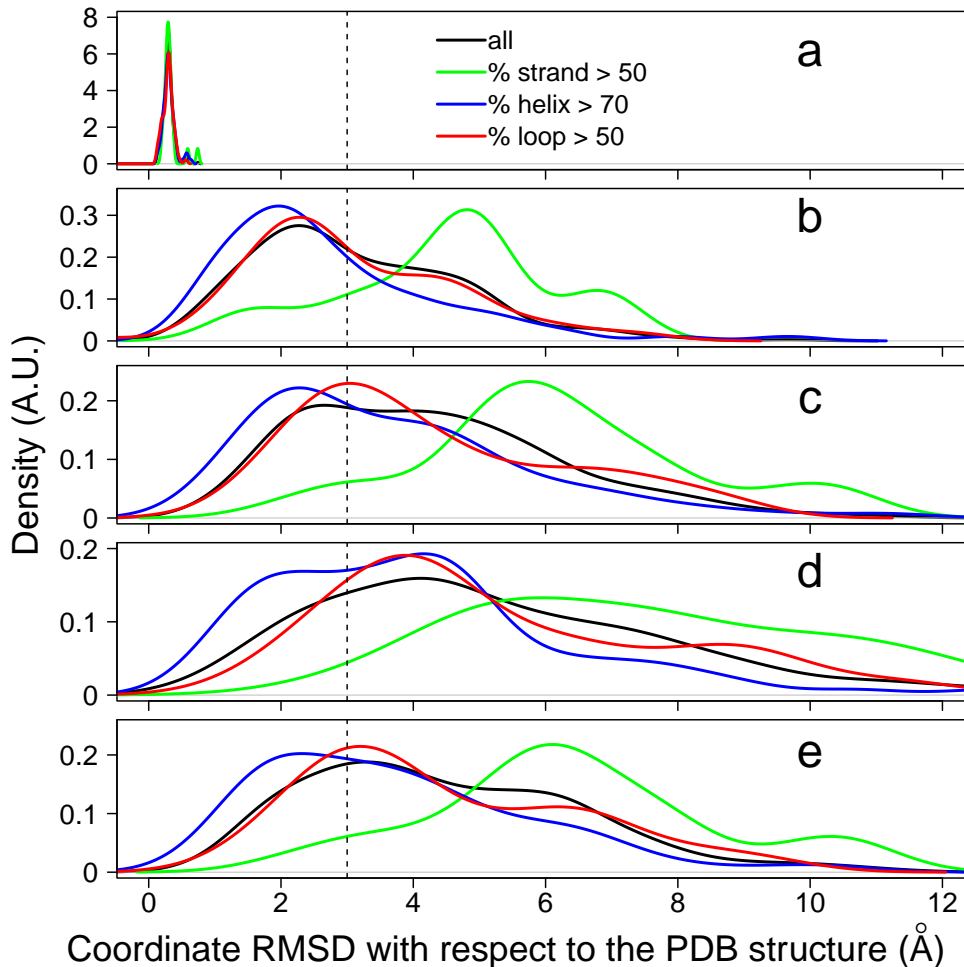


Figure 4: Distribution of the root-mean-square deviation (RMSD, Å) of atomic coordinates between the initial PDB conformation and the conformation reconstructed using iBP. The following stereochemistry inputs (Table 1) were used: (a) pdb stereochemistry taken from the PDB input, (b) pdb stereochemistry with ω values of 178 deg, (c) Hollingsworth stereochemistry with bond and ω angles averaged along the Hollingsworth regions (Figure S2), (d) Hollingsworth stereochemistry with ω values of 178 deg, (e) uniform stereochemistry³⁴ (Table S2). The vertical dashed line indicated the RMSD value of 3 Å. The curves are colored depending on the percentage of residues belonging to α -helices, to β -strands, or to loops as described in the legend. The secondary structures were determined using STRIDE.⁴⁷

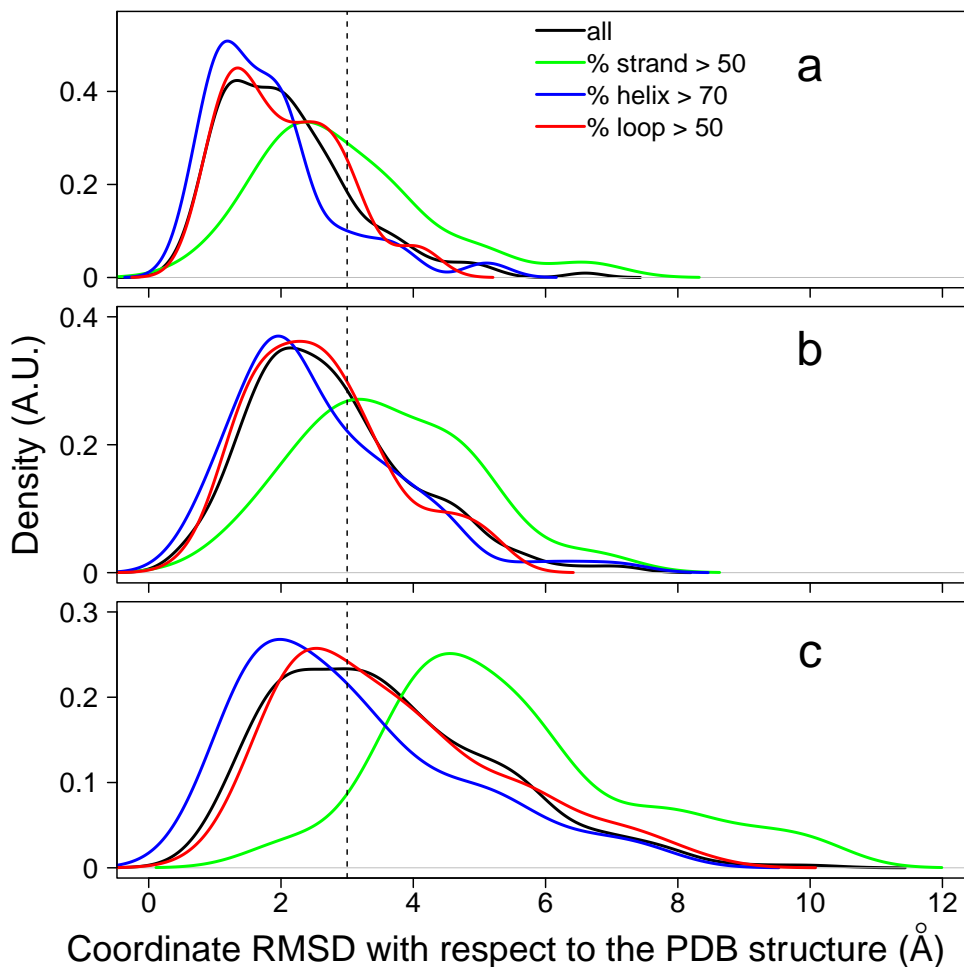


Figure 5: Distribution of the root-mean-square deviation (RMSD, Å) of atomic coordinates between the initial PDB conformation and the conformation reconstructed using iBP with the Hollingsworth stereochemistry for bond angles along with various definitions of the ω angles: (a) ω values taken from the PDB initial conformation, (b) discretization of ω values among four classes (Eq 4), (c) discretization of ω to $178^\circ \text{sgn}(\omega)$, where $\text{sgn}(\omega)$ is the sign of ω in the initial PDB structure.

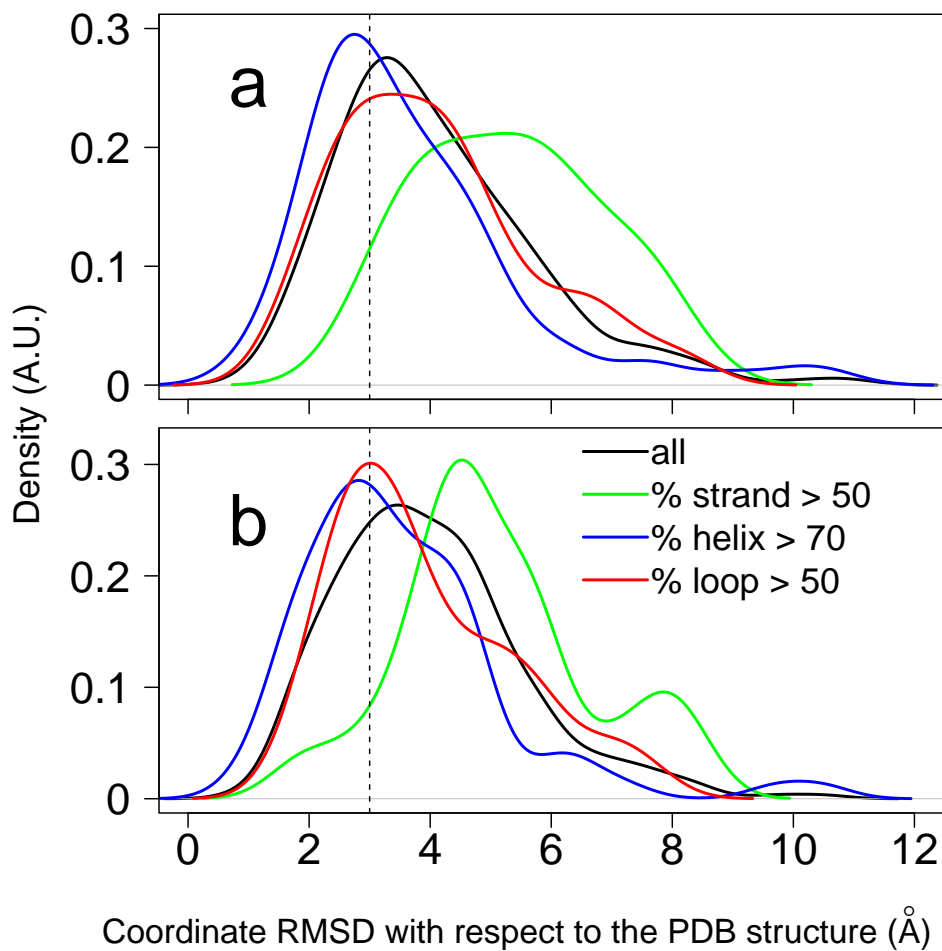


Figure 6: Distribution of the root-mean-square deviation (RMSD, Å) of atomic coordinates between the initial PDB conformation and the conformation reconstructed using enumerating iBP runs with Hollingsworth (a) or uniform (b) stereochemistry. The coordinate RMSD was taken as the smallest RMSD value obtained among all iBP solutions.

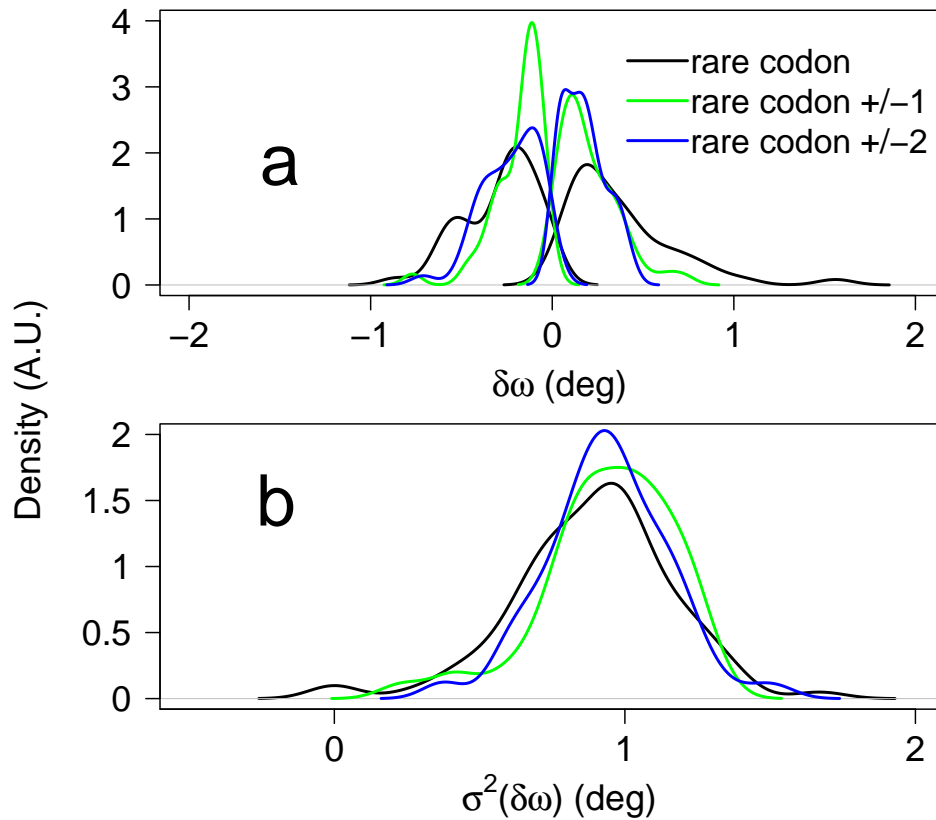


Figure 7: Distribution of the variations of $\delta\omega$ (Eq 3) (a) and of the standard deviations $\sigma^2(\delta\omega)$ (b) for various positions in the protein sequences: at the residues for which the rare codons are observed (black curve), at the residues neighboring the rare codon (green curve) and at the residues second neighbor of the rare codon (blue codon).

Table 1: Definitions of stereochemistry inputs.

name	origin of stereochemistry
pdb	initial conformation from the Protein Data Bank
Hollingsworth	averaged angle values from the Ramachandran regions ³³
uniform	stereochemistry parameters from Engh and Huber ³⁴ (Table S2)