



HAL
open science

A refined pH-dependent coarse-grained model for peptide structure prediction in aqueous solution

Pierre Tufféry, Philippe Derreumaux

► To cite this version:

Pierre Tufféry, Philippe Derreumaux. A refined pH-dependent coarse-grained model for peptide structure prediction in aqueous solution. *Frontiers in Bioinformatics*, 2023, 3, 10.3389/fbinf.2023.1113928 . hal-04030567

HAL Id: hal-04030567

<https://hal.sorbonne-universite.fr/hal-04030567v1>

Submitted on 15 Mar 2023

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.

A refined pH-dependent coarse-grained model for peptide structure prediction in aqueous solution

Pierre Tufféry^{1,*} and Philippe Derreumaux^{2,3}

¹ *Université Paris Cité, CNRS UMR 8251, INSERM U1133, Paris, France.*

² *CNRS, Université Paris Cité, UPR 9080, Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, Fondation Edmond de Rothschild, 13 rue Pierre et Marie Curie, 75005 Paris, France.* ³ *Institut Universitaire de France (IUF), 75005, Paris, France.*

Correspondence*:

Pierre Tufféry, Université Paris Cité, 35 rue H. Brion case 7113, 75205 Paris cedex 13, France
pierre.tuffery@u-paris.fr

2 ABSTRACT

3 we selected the 5

4 **Keywords:** peptide, structure, pH dependence, coarse grained models, prediction

1 INTRODUCTION

5 Peptides of less than 40 amino acids have diverse biological functions, acting as signaling entities in all
6 domains of life, and targeting receptors or interfering with molecular interactions. Hormones and their
7 bacterial mimetics (?), neuropeptides and their roles in neurodegenerative diseases (?), antimicrobial
8 peptides contribution to host defence (?), and immunomodulatory peptides in the perspective of vaccine
9 design (?) are some current directions motivating their study at a fundamental level. Due to their specific
10 features, peptides have also gained interest as therapeutical agents ?, particularly to target protein-protein in-
11 teractions (?). They are also considered as having interest in the development of new functional biomimetic
12 materials (?). Peptides have limitations though, as they can be highly flexible (?), which motivate efforts to
13 understand and predict their conformational energy landscapes.

14 Structure prediction of polypeptides with amino acid lengths up to 40 amino acids in aqueous solution
15 can be performed by a series of methods including machine-learning approaches such as AlphaFold2 (?),
16 TrRosetta (?), and APPTTEST (?). Looking at AlphaFold2, which revolutionized structure prediction of
17 single folded domain to a root-mean-square deviation (RMSD) accuracy of 0.2 nm, its capability lies
18 on machine learning based on protein data bank (PDB) (?) templates, multiple sequence alignments,
19 co-evolution rules and sophisticated algorithms to predict local backbone and side conformations, and
20 side chain - side chain contact probability within distances bins. AlphaFold2 builds the protein by energy
21 minimization using a protein-specific energy potential.

22 TrRosetta is basically similar to AlphaFold2. It builds the protein structure based on direct energy
23 minimizations with a restrained Rosetta. The restraints include inter-residue distance and orientation

24 distributions predicted by a deep neural network. Homologous templates are included in the network
25 prediction to improve the accuracy further.

26 APPTTEST also uses machine learning on the PDB structures with a chain length **varying** between 5 and
27 40 amino acids. APPTTEST derives $C\alpha$ - $C\alpha$ and $C\beta$ - $C\beta$ distance restraints, and backbone dihedral restraints
28 that are input of simulated annealing and energy minimization.

29 Other methods which are accessible by WEB-servers or can be downloaded include Rosetta (?), I-
30 TASSER (?), PepStrMod (?) and PEPFOLD (??). Rosetta is a fragment-assembly approach based on
31 Monte Carlo simulation, a library of predicted 9 and then 3 residues, and a coarse-grained model, followed
32 by all-atom refinement. I-TASSER is a hierarchical approach that identifies structural templates from
33 the PDB by multiple threading approaches, with full-length atomic models constructed by iterative
34 template-based fragment assembly simulations.

35 The PEPstrMOD server predicts the tertiary structure of small peptides with sequence length varying
36 between 7 to 25 residues. The prediction strategy is based on the realization that β -turn is an important
37 feature of small peptides. Thus, the method uses both the regular secondary structure information predicted
38 from PSIPRED and the β -turns information predicted from BetaTurns. The structure is further refined with
39 energy minimization and molecular dynamic simulations.

40 PEP-FOLD2 is a fast accurate structure peptide approach based on the prediction of a profile of the
41 structural alphabet of **4 amino acid lengths** along the sequence, and a chain growth method based on the
42 coarse-grained sOPEP2 model followed by Monte Carlo steps. It should be noted that PEP-FOLD2 is not
43 free of learning as it uses an Support Vector Machine predictor relying on multiple sequence alignment. Of
44 practical interest, during the time of this study, we could not access the APPTTEST and PepStrMod servers.
45 Also, TrRosetta cannot be applied to sequences with < 10 amino acids.

46 Overall, all these methods generate good models for well-structured peptides at pH 7 in aqueous
47 solution because most structures deposited in the PDB from nuclear magnetic resonance (NMR) and X-ray
48 diffraction **experiments** were determined at neutral pH, and the PDB contains close to 200,000 structures
49 as of October 30th, 2022.

50 These methods face, however, two current limitations: correct conformational ensemble sampling of
51 intrinsically disordered peptides or proteins (IDPs) which lack stable secondary and tertiary structures,
52 and accurate conformational ensemble prediction of peptides as a function of pH and salt conditions. The
53 first issue has motivated the development of new force fields, such as CHARMM36m-TIP3P modified (?),
54 AMBER99-DISP (?) and many others (?). The current approach to address the impact of pH variations is
55 to perform your own extensive molecular dynamics and replica exchange molecular dynamics simulations
56 at your desired pH. Alternatively one can use pH-replica exchange molecular dynamics using a discrete
57 protonation method (?) or all-atom and coarse-grained continuous constant pH molecular dynamics
58 (CpHMD) methods (???). **Accurate and fast peptide structure predictions at different pH and salt conditions**
59 **are the objectives of the present study.**

60 The organization of this paper is as follows. In section 2, we present an extension of the coarse-
61 grained, sOPEP2, force field to integrate Debye-Hückel charge interactions as a function of pH and salt
62 concentrations. Next, we present the **TrRosetta**, AlphaFold2 and PEP-FOLD with and without Debye-
63 Hückel protocols and the analysis methods. In section 3, we present the results of structure predictions
64 of six poly-charged **peptides** as a function of pH and compare them to experimental **circular dichroism**
65 **(CD)** data, and the predicted models obtained by TrRosetta and AlphaFold2. The charged polypeptides are
66 particularly interesting to assemble the sOPEP2 interactions and the Debye-Hückel charge interactions.

67 This is followed by the prediction of 15 ordered peptides, which have NMR structures **resolved at a pH**
 68 **varying from 2 to 8**. We finish this section on the prediction of four peptides for which low-resolution
 69 experimental data and topological descriptions are available. Finally section 4 summarizes our findings.

2 METHODS

70 2.1 the sOPEP2 force field

71 The sOPEP2 potential, to be used in a discrete space, originates from the OPEP potential which uses **an**
 72 **explicit representation** of the backbone (N, H, C α , N and H atoms) and one bead for each side chain, whose
 73 position to C α and **Van der Walls radius depend on the amino acid type(??)**. The sOPEP2 is expressed as a
 74 sum of local, nonbonded and hydrogen-bond (H-bond) terms, with all parameters described in (?).

$$E = E_{local} + E_{nonbonded} + E_{H-bond} \quad (1)$$

75 Since the geometry in PEP-FOLD is mainly imposed by the superimposition of the discrete **structural**
 76 **alphabet (SA) letters**, the local contributions are restricted to a simple flat-bottomed quadratic potential to
 77 described the energy associated with dihedral angles ϕ and ψ , described by:

$$E(\phi_i) = \epsilon_{\phi} (\phi_i - \phi_{0_sc_i})^2 \quad (2)$$

78 where $\phi_{0_sc_i} = \phi$ within the interval $[\phi_{low_sc_i}, \phi_{high_sc_i}]$ and $\phi_{0_sc_i} = \min(\phi - \phi_{low_sc_i}, \phi - \phi_{high_sc_i})$
 79 outside of the interval $\phi_{low_sc_i}$ and $\phi_{high_sc_i}$ are specific to each amino acid type (Binette et al. 2022).

80

81 Nonbonded interactions corresponding to repulsion/attraction effects are described using the Mie
 82 potential (?) given by :

$$E_{vdw_ij} = \epsilon_{ij} \times \left[\frac{m}{n-m} \left(\frac{r_{ij}^0}{r_{ij}} \right)^n - \frac{n}{n-m} \left(\frac{r_{ij}^0}{r_{ij}} \right)^m \right] \quad (3)$$

83 where ϵ_{ij} is the potential depth and r_{ij}^0 is the position of the potential minimum function of atomic types
 84 for i and j . The combination of exponents, n and m , gives the relationship between the position of the
 85 potential minimum (r^0) and the position where it is zero ($gR0$):

$$gR0 = \left(\frac{m}{n} \right)^{\frac{1}{n-m}} r_0 \quad (4)$$

86 Hydrogen bonds are considered explicitly, using a combination of two types of contributions:

$$E_{H-bond} = E_{HBpairwise} + E_{HBcoop} \quad (5)$$

87 where $E_{HBpairwise}$ corresponds to the two-body contributions of all hydrogen bonds between residue (i)
 88 and residue (j), characterized by the hydrogen/acceptor distance r_{ij} and the donor/hydrogen/acceptor angle
 89 α_{ij} :

$$E_{HBpairwise}(r_{ij}, \alpha_{ij}) = \epsilon_{\alpha}^{HB} \sum_{ij, j=i+4} \mu(r_{ij}) \cdot \nu(\alpha_{ij}) + \epsilon_{\beta}^{HB} \sum_{ij, j>4} \mu(r_{ij}) \cdot \nu(\alpha_{ij}) \quad (6)$$

$$\mu(r_{ij}) = \epsilon_{ij} \cdot \left[5 \left(\frac{\sigma}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma}{r_{ij}} \right)^{10} \right] \quad (7)$$

$$\nu(\alpha_{ij}) = \begin{cases} \cos(\alpha_{ij}) & \text{if } \alpha_{ij} > 90^{\circ} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

90 where $\sigma = 0.18$ nm is the position of the potential minimum and ϵ is the potential depth. We distinguish
 91 between α -helix-like hydrogen bonds defined by O(i)-H(i+4) and other hydrogen bonds. Hydrogen bonds
 92 between a pair of residues separated by less than four amino acids are not considered.

93 E_{HBcoop} involves four-body interactions involving pairs of hydrogen bonds (between residues (i) and (j)
 94 and residues (k) and (l)), so as to stabilize secondary structure motifs. The cooperation energy is given by
 95 the following equations:

$$E_{HBcoop}(r_{ij}, r_{kl}) = \epsilon_{\alpha}^{coop} \sum C(r_{ij}, r_{kl}) \times \Delta(ijkl) + \epsilon_{\beta}^{coop} \sum C(r_{ij}, r_{kl}) \times \Delta'(ijkl) \quad (9)$$

$$C(r_{ij}, r_{kl}) = \exp(-0.5(r_{ij} - \sigma)^2) \cdot \exp(-0.5(r_{kl} - \sigma)^2) \quad (10)$$

$$\Delta(ijkl) = \begin{cases} 1 & \text{if } (k, l) = (i + 1, j + 1) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$\Delta'(ijkl) = \begin{cases} 1 & \text{if } (k, l) = (i + 2, j - 2) \\ & \text{or } (k, l) = (i + 2, j + 2) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$(13)$$

96 2.2 Debye-Hückel charge interactions

97 The new sOPEP version introduces the possibility to consider pH-dependent charge interactions, using
 98 the Debye-Hückel formulation (?).

$$E_{DH_{ij}} = (q_i * q_j * e^{-r_{ij}/l_{DH}}) / (\epsilon(r_{ij}) * r_{ij}) \quad (14)$$

99 where q_i and q_j correspond to the charge of particles i and j , $j > i + 1$, respectively, r_{ij} is the distance
 100 between the particles, l_{DH} is the Debye length that depends of the ionic strength of the solvent, and $\epsilon(r_{ij})$
 101 is the dielectric constant that depends on the distance between the charges:

$$\epsilon(r) = D_w - (D_w - D_p)(s^2 r^2 + D_p s r + D_p) e^{-sr} / D_p \quad (15)$$

102 where D_w is the dielectric constant of water, D_p is the dielectric constant inside a protein, and s is the
 103 slope of the sigmoidal function. In practice, we used values of 78, 2 and 0.6 for D_w , D_p and s , respectively,
 104 as stated in (?).

105 Since the sOPEP representation does not include all-atom side chains, but charges associated with
106 particles of heterogeneous sizes, it is necessary to shift the energy curve to have energy values compatible
107 with those of the Mie formulation. For each pair of particle, we shifted the distance using:

108 $r_{ij}^{SH} = r_{ij} + shift_{ij}$ and we evaluated $E_{DH_{ij}}$ using r_{ij}^{SH} except for $\epsilon(r)$, where the unshifted distance is
109 used.

110 Shift values were adjusted for r such as $E_{vdw_r} = k$, $E_{DH_r} = E_{vdw_r}$, as illustrated Figure 1. In practice,
111 we found that values of k on the order of 4 kcal/mol are convenient, and the Debye-Hückel energy was
112 truncated to E_{DH_r} to avoid redundancy with the Mie potential. Also, as sOPEP2 side-chain side-chain
113 potential already includes some of **the interactions between the charged residues**, the Mie potential is
114 restricted to only the repulsive part for charged side chains.

115 Charges were assigned to particles depending on the pH using pKa values of 3.9, 4.2, 6.0, 10.5 and
116 12.5 for ASP, GLU, HIS, LYS and ARG side chains, respectively, and 9.0 and 2.0 for the N-terminal
117 α -ammonium and C-terminal α -carboxyl groups, respectively. Note that it is possible to consider blocking
118 the extremities using acetyl and N-methyl on the N-terminus and C-terminus groups, respectively, in which
119 case no charge is assigned to the extremity.

120 Finally, we have considered weighting differently the **electrostatic** contributions depending on the
121 separation of the amino acids in the sequence. In our experience, best results were obtained using a weight
122 of 10 for residue separation of less than 7 **amino acids** and a weight of 2, otherwise.

123 **2.3 PEP-FOLD, TrRosetta and AlphaFold2 protocols and analysis**

124 Our validation test set includes a total of 25 peptides as described in Section 3. For each peptide, we
125 performed 2 PEP-FOLD simulations, one TrRosetta simulation which uses PDB templates and homologous
126 sequences, and one AlphaFold2 simulation in its standard version using 3 recycles, template information,
127 and AMBER refinement. **Both TrRosetta and AlphaFold2 simulations return five models that we considered**
128 **equiprobable. The PEP-FOLD simulations without Debye-Hückel (referred to as PF-noDH), and with**
129 **Debye-Hückel (PF-DH) generated 200 models each. We selected the representative models of the five**
130 **best clusters identified among the 200 generated models based on their rankings using sOPEP2 energies -**
131 **i.e. the standard PEP-FOLD model selection - for PF-noDH and the sum of sOPEP2 and Debye-Hückel**
132 **energies for PF-DH.**

133 We have considered 15 peptides for which a PDB structure is available. These correspond to peptides
134 previously studied during PEP-FOLD development and new peptides **with their structures** released after
135 September 1st, 2019, and **determined in pure aqueous environment. The predicted models of the 15 peptides**
136 **were evaluated by computing the CAD-score (?). The reported CAD-score corresponds to the largest**
137 **value of the cross CAD-scores between the five predicted models and all NMR structures. Following our**
138 **previous work, if the CAD-score calculated on the backbone atoms is > 0.60 , the model is associated**
139 **with largely correct secondary structure prediction, otherwise if it is > 0.65 the model is correct in terms**
140 **of secondary and tertiary structures. For the poly-charged peptides, we also computed their secondary**
141 **structure contents using STRIDE program (?). For the four sequences free of any NMR structure, we**
142 **compared their predicted and experimental topologies.**

3 RESULTS AND DISCUSSION

143 3.1 Predicted Models of poly-charged peptides

144 For the simulations of the six **poly-charged peptides**, namely (EK)15, (EK)5, (H)30, (E)15, (K)15
145 and (R)25, we calculated the alpha-helix, coil and turn contents averaged over **the five models of each**
146 **method and compared with circular dichroism (CD)** experiments. It is to be noted that by default TrRosetta,
147 AlphaFold2 and PF-noDH only perform simulations at neutral pH. CD values are not available at all pH
148 varying from 3 to 13. We report, however, on the pH-dependent conformations using **PF-DH**. Results are
149 summarized in Table 1.

150 TrRosetta, AlphaFold2 and PF-noDH have a very high propensity to report alpha-helical conformations
151 for the six polypeptides at pH 7.4, the exception being (H)30, with alpha-content varying from 54% to
152 97%, while CD displays only coil or beta-turn signals. For instance, for (EK)5, TrRosetta reports 68%
153 helix and 26% coil, AlphaFold2 reports 54% helix and 10% coil and PF-noDH reports 90% helix and 10%
154 coil. Only PF-DH is able to predict the CD coil character of (EK)5 with 68% coil and 32% turn.

155 PF-DH is the single method to predict 85% coil and 15% turn consistent at pH 7.4 with the beta-turn CD
156 signal of (EK)15 (?), and PF-DH predicts 100% coil at pH 5 consistent with the coil CD signal of (H)30
157 (?). There is strong experimental evidence that (H)30 polymerizes at pH 7.4 forming beta-sheets. At this
158 pH, PF-noDH and TrRosetta predict strong helical conformations, while PF-DH and AlphaFold2 predict a
159 random coil, with contents of 95% and 81%, respectively.

160 The polypeptides (K)15 and (E)15 are particularly interesting because the alpha-helix content changes
161 inversely with the pH. As observed by CD, the helical content of (K)15 increases with pH, while the helical
162 content of (E)15 decreases with pH (?). (K)15 have 0% helix at pH 3.6 and 83.7% helix at pH 11-13 by
163 CD. PF-DH finds 0% helix at pH 3.6 and 93% helix at pH 11-13 (Figure 2). In contrast to (K)15, (E)15
164 (Figure 3) have 42% helix at pH 3.6 and 0% helix at pH 11-13 by CD. PF-DH finds 93% helix at pH 3.6
165 and 100% coil at pH 11-13.

166 The conformation ensemble of (R)25 is predicted to have 50% coil and 31% beta-sheet at pH 5.7 and
167 have 51% coil and 21% beta-sheet at pH 11.3 by CD (?). PF-DH predicts 100% coil, independently of the
168 pH values. Its performance is however much better than those of PF-noDH, AlphaFold2 and TrRosetta
169 which predict a high helical signal varying from 77% to 96%.

170 Overall, the structure predictions of the six polypeptides at neutral pH (7.4) give quite different contents of
171 the secondary structure using PF-noDH and PF-DH, with PF-noDH behaving and failing like AlphaFold2
172 and TrRosetta predictions. This result emphasizes the role of the Debye-Hückel charged-charged interactions
173 when treating poly-charged peptides. The results also demonstrate that the learning stage of the local
174 conformations in PEP-FOLD performed from structures at neutral pH can be counterbalanced by the force
175 field, making possible to explore new conformations depending on the pH. In contrast, AlphaFold2 and
176 TrRosetta rely on homologous structures and multiple sequence alignments. Since neither is available for
177 poly-charged peptides, it is normal for both predictors to fail. But surprisingly, the LDDT (local distance
178 difference test) metric predicted by both methods is, on average, very high (>80%) for all amino acids of
179 the six poly-charged peptides.

180 It is important to emphasize that in this study, we assume the standard pka values of charged amino acids
181 irrespective of the amino acid composition of the peptides and the conformations of the peptides. This is a
182 strong limitation of our current approach. Determining the pka values of charged amino acids in protein
183 structures has motivated the development of many theoretical methods (????). To illustrate the variation

184 of the pka values, we used the H++ server which is based on classical continuum electrostatics and basic
185 statistical mechanisms (?). Using (K)15, we found pka values ranging from 10.1 to 9.4 (versus 10.5 in our
186 model); using (R)25, we found pka ranging from 9.6 to 11.6 in one conformation, and from 10.9 to 11.7 in
187 another conformation (versus 12.5 in our model), and using (H)30, we found pka variations from 4.7 to
188 6.3 (versus 6.0 in our model). Clearly this change of pka of the amino acids will impact the equilibrium
189 conformations of PF-DH.

190 3.2 Predicted Models of polypeptides with NMR structures

191 The experimental information of each well-ordered peptide, given in Table 2, includes the amino acid
192 length varying from 8 to 35 amino acids, the number of NMR models, the WDC (well-defined rigid core)
193 according to the PDB, the topology, the pH varying from 4.3 to 7, the ionic strength varying from 0 to 150
194 mM NaCl, the blocking of the extremities and the amino acid sequence.

195 Table 3 reports on the CAD scores using the full sequences and the rigid cores of each of the 15 peptides
196 using the four methods. Note we give the results of PF-noDH, AlphaFold2 and TrRosetta, because these
197 methods which are pH independent are used irrespective of the experimental pH conditions.

198 The first striking result is that (the mean, standard deviation and median) values of the CAD-scores
199 averaged over the 15 peptides are nearly identical for the four methods using both the full sequences or
200 the rigid cores. They reach (0.73, 0.07, 0.75) for PF-noDH, (0.74, 0.10, 0.74) for AlphaFold2, (0.74, 0.06,
201 0.74) for PF-DH and (0.76, 0.08, 0.77) for TrRosetta using the full sequences. Similar trends are observed
202 considering the well-defined rigid cores, the average CAD-scores being of 0.75, 0.76, 0.76 and 0.78 for
203 PF-noDH, PF-DH, AlphaFold2 and TrRosetta, respectively.

204 The second result is that PF-noDH and PF-DH do not predict any low quality models (CAD-score < 0.6),
205 while AlphaFold2 produces CAD-scores of 0.59, 0.55 and 0.6 for the three targets 6nm3, legs and 7li2
206 (Figures 4, 5, 6, respectively). It has to be noted that the structures of these three peptides were solved at
207 pH 5.8, 6.5 and 7. This low score results in differences between the experimental and predicted topologies.
208 Experimentally, 6nm3 adopts a helical-like conformation, legs adopts a beta2-like conformation and 7li2
209 adopts a beta-2 like conformation. Of note, a beta-2 like conformation has the topology of a beta-hairpin
210 but lacks the H-bond network.

211 For these three systems, AlphaFold2 predicts an extended-unstructured conformation. The 7li2 target is
212 also problematic for TrRosetta, as it is the single system with a CAD-score <0.6, namely 0.58 leading to
213 an extended-unstructured conformation. Inversely, TrRosetta is the best method to predict the beta-hairpin
214 of 1pgbF (?) with a CAD-score of 0.91 versus 0.83 with AlphaFold2 and 0.79 with PF-DH.

215 The third result is related to the performances of PF-DH with respect to PF-noDH, which provides
216 evidence that the weights of the Debye-Huckel salt bridge interactions are consistent with the weights of
217 sOPEP2 interactions. It was far from being evident that the addition of charges at extremities and charged
218 amino acids in the core of the sequences would not change the quality of the models for pH varying
219 between 2.9 and 7. The number of titratable amino acids varies from 1-2 (1le1, legs - 6nm3, 6evq), 4 for
220 1j4m, 5 for 6j9p and 1pgbF, 6 for 6mi9, 7 for 1wbr and 7li2, 9 for 6r2x, 10 for 6svc and 7b2f, to 12 for
221 1fsd. The results also show that the pH-independent PEP-FOLD version and the pH-dependent PEP-FOLD
222 version perform similarly for peptides containing charged, hydrophilic and hydrophobic amino acids.

223 Finally, using NMR structures as a reference, a very recent study benchmarked the accuracy of AlphaFold2
224 in predicting 588 peptide structures between 10 and 40 amino acids, including soluble peptides, membrane-
225 associated peptides, and disulfide-rich peptides. Although the study ignores pH conditions and the presence

226 of the membrane, AlphaFold2 can be used for the modeling of peptide structures anticipated to have a
227 well-defined secondary structure. AlphaFold2 is particularly successful in the prediction of alpha-helical
228 membrane-associated peptides and disulfide-rich peptides, but also shows some shortcomings in predicting
229 phi and psi angles. It was found that AlphaFold2 performs at least as well if not better than TrRosetta and
230 PEP-FOLD using our 2016 set of parameters (Pierre: add Meiler J. Structure 2023)

231 3.3 Predicted Models of polypeptides without any NMR structures.

232 The last four peptides have been discussed in literature in terms of topological features without delivering
233 any NMR structure. Their sequences are given at the bottom of Table 2.

234 Two peptides are rather well described by all four methods. Pep17 has been shown as a stable monomeric
235 helix at pH 2 using CD and NMR experiments (?). PF-noDH, PH-DH at pH 2 and AlphaFold2 predict a
236 helical conformation with a frayed N-terminus, while TrRosetta predicts a full helical conformation (Figure
237 7). Pep38 determined experimentally as a helix-turn-helix at pH 3.6 (?) is also well reproduced by the four
238 methods.

239 There are two cases, where AlphaFold2 and TrRosetta fail to produce the experimental data. The first
240 peptide is pep10 which is described experimentally by an ensemble of distinct transient beta-hairpins at pH
241 4.3(?). It is described as an unstructured turn-like conformation by TrRosetta (8D), and an ensemble of
242 extended and beta2-like conformations by AlphaFold2 (8C). In contrast, PF-noDH and PF-DH predict well
243 a beta-hairpin conformation (8A and B).

244 The second peptide is the tau fragment encompassing residues 295-306 containing the aggregation-prone
245 PHF6 motif (306-311). Using cross-linking mass-spectrometry, ab initio Rosetta (?), and CS-Rosetta which
246 leveraged available chemical shifts (?) for the tau repeat spanning residues 243-365, the tau fragment
247 295-306 was predicted as a beta-hairpin at pH 7 (?). PF-noDH and PF-DH predict the same conformation
248 (Figure 9A and B). In contrast, AlphaFold2 predicts extended conformations (9C), and surprisingly
249 TrRosetta finds a random coil conformation (9D).

250 Overall, this small set of peptides provides evidence of some limitations of AlphaFold2 and TrRosetta
251 when the target does not have an homologous sequence in the PDB.

4 CONCLUSIONS

252 Integrating pH variation effects to a coarse-grained model, where the side chains are represented by one
253 single bead, is an important step toward accurate polypeptide structure prediction in aqueous solution,
254 as coarse-graining with various granularities (??), enhance sampling. This task has been performed by
255 combining a Debye-Hückel formalism for charged - charged side chain interactions and the sOPEP2
256 potential. By using a total of 25 peptides of amino acid lengths varying between 7 to 38 amino acids, this
257 study provides evidence that PF-noDH, PF-DH, AlphaFold2 and TrRosetta perform similarly on peptides
258 deposited in the Protein data Bank, but PF-DH outperforms the two recent machine-learning methods for
259 poly-charged peptides, and peptides for which homologous sequences are not deposited in the PDB. Of
260 note, our new formulation takes into account the impact of salt concentration variations, but we could not
261 identify from the literature any case reporting a conformation change upon ionic strength variation.

262 Overall this work is one step towards peptide structure prediction in mimicking in vivo conditions. We
263 are currently working on IDP's in aqueous solution and de novo structure prediction of peptides at the
264 surface of two-dimensional cell membranes.

	Exp. Block	Experiment		PF-noDH		PF-DH		AlphaFold2		TrRosetta					
		pH	CD	α	β -turn	α	β -turn	α	β -turn	α	β -turn				
(E)K)15	-	3		94	6	0	28	69	3	97	3	0			
		7.4	β -turn				0	85	15						
		11					24	68	8						
(E)K)5	-	3	coil	90	10	0	0	100	0	54	10	36	6		
		7.4					0	68	32						
		11					0	78	22						
(H)30	-	5	coil				0	100	0						
		7.4	aggregation	75	5	20	0	95	5	0	81	19	3	0	
(K)15	Ace, Nmt.	3.6	0% α				0	100	0						
		7.4	0% α	93	7	0	0	100	0	73	8	19	93	7	0
		11.2-13	83.7% α				93	7	0						
(E)15	Ace, Nmt.	3-3.6	ND-42% α				93	7	0						
		4.2	75% α				93	7	0						
		7.4	0% α	93	7	0	0	100	0	93	7	0	93	7	0
(R)25	-	5.7	50% coil				0	100	0						
		7.4	ND	96	4	0	0	100	0	77	4	19	96	4	0
		11.3	50% coil 15% α				0	100	0						
		13		96	0	0	96	0	0						

Table 1. Structural impact of pH variation on poly-charged peptides. Results are presented for PEP-FOLD without and with Debye-Hueckel (PF-noDH, PF-DH), AlphaFold2 and TrRosetta. Experimental information about the blocking of extremities (Exp. Block.) using acetyl (Ace) or N-methyl (Nmt) and Circular Dichroism (CD) data, is reported. ND stands for not determined.

Target	L	# mod.	WDC	topo.	pH	ion.Str	Exp. Blok.	Sequence
6mi9	19	10	3-19	α distort.	4.3	0	Nmet.	PMARNKILGKILRKIAAFK
6j9p	12	10	2-12	α 5	0	-	-	RRLRLRLRLR
1fsd	28	41	3-25	$\beta_2\alpha$	5	0	-	QQYTAKIKGRTRNEKELRDFIEKFKGR
1j4m	14	1	ND	β_2	5	0	-	RGKWTYNGITYEGR
1le1	12	20	2-11	β_2	5.5	0	Nmet.	SWTWENGKWTWK
6nm3	8	5	ND	α -like	5.8	0	Nmet.	RKIWWWWL
6svc	35	20	2-35	β_3	6	150	-	SKLPPGWEEKRMSRNSGRVYFNNHITNASQFERPSG
2evq	12	43	2-10	β_2	6	20	-	KTNWPAATGKWTE
1egs	9	20	1-9	β_2 -like	6.5	0	Ace, Nmet.	TKSAGGIVL
6r2x	25	15	9-20	C α	6.5	148	-	FETLRGDERILSILRHQNLKELQD
7b2f	31	20	6-21	α C α	6.5	100	-	MNNNELTSLPLAERKRLLLELAKAAKLSRQHY
6s0n	9	20	ND	α -like	6.8	1	-	QDVNTAVAW
7H2	22	20	1-22	β_2 -like	7	0	-	AGTMRVTYPDGGQKPGQSDVEKD
1wbr	17	32	1-16	α C'	7	0	Ace, Nmet.	QAERMSQIKRLLSEKKT
1pgbF	16	1	ND	β_2	7	0	-	GEWTYDDATKTFVTE
pep17	17	-	-	α	2	0	-	ETGTKAELLAKYEATHK
pep38	38	-	-	α T α	3.6	20	-	DWLKARVEQELQALEARGTDSNAELRAMEAKLKAEIQK
pep10	11	-	-	β_2	4.3	0	-	IYSNSDGGWTWT
tau fragment	17	-	-	β_2	7	0	-	DNIKHVPGGGSVQIVYK

Table 2. Peptide set.

For each peptide with an experimental structure available we specify its PDB identifier (PDB), size (L), the number of NMR models available (# mod.), its well defined core according to the PDB (WDC), its topology (topo.); and the experimental conditions including the pH, the ionic strength (ion. Str.) and the presence of extra groups to block the N terminus (acetyl) and C terminus (N methyl) (Exp. Blok.), and the amino acid sequence. Four additional peptides without deposited structures but for which information exists in the literature are reported at the bottom of the table.

Target	Exp.		PF-noDH		PF-DH		AlphaFold2		TrRosetta		
	topo	pH	topo	Full	WDC	topo	Full	WDC	topo	Full	WDC
6mi9	α	4.3	α	0.85	0.86	α, α -like	0.84	0.83	α	0.84	0.84
6j9p	α	5	α	0.75	0.77	C α	0.70	0.74	α	0.74	0.75
1fsd	β_2 <i>alpha</i>	5	$\alpha_2\alpha$	0.62	0.63	$\beta_2\alpha$	0.66	0.73	$\beta_2\alpha$	0.71	0.73
1j4m	β_2	5	β_2	0.67	0.67	β_2	0.76	0.74	β_2	0.73	0.73
1le1	β_2	5.5	β_2	0.76	0.8	β_2	0.75	0.83	β_2	0.8	0.79
6nm3	α -like	5.8	α	0.72	0.72	$\alpha, unstr$	0.73	0.59	-	-	-
6svc	β_3	6	β_3	0.66	0.66	β_3	0.70	0.79	β_3	0.78	0.79
2evq	β_2	6	β_2	0.84	0.86	β_2	0.82	0.89	β_2	0.8	0.79
1egs	β_2 -like	6.5	β_2, β_2 -like	0.71	0.71	β_2 -like	0.70	0.55	-	-	-
6r2x	C α	6.5	C α	0.75	0.9	C α	0.74	0.8	C α	0.8	0.9
7b2f	α C α	6.5	α C, <i>alpha</i>	0.77	0.82	α C α	0.79	0.79	α C α	0.77	0.83
6s0n	α -like	6.8	α	0.76	0.76	α	0.79	0.67	-	-	-
7H2	β_2 -like	7	β_2, β_2 -like	0.62	0.62	β_2, β_2 -like	0.64	0.6	coil	0.58	0.58
1wbr	α C	7	α	0.69	0.7	α	0.70	0.69	α	0.68	0.69
1pgbF	β_2	7	β_2	0.77	0.77	β_2	0.79	0.83	β_2	0.91	0.91
MEAN				0.73	0.75		0.74	0.74		0.76	0.78
STDEV				0.07	0.09		0.06	0.10		0.08	0.09
MEDIAN				0.75	0.76		0.74	0.74		0.77	0.78

Table 3. Performance prediction for structured peptides.

For each structure, we report a short description of the topology of the 5 best models, and the CADscore values (see methods) obtained for PEP-FOLD without and with Debye-Huckel (PF-noDH, PF-DH), AlphaFold2 and TrRosetta. Note that TrRosetta is not functional for amino acid lengths < 10.

CONFLICT OF INTEREST STATEMENT

266 The authors declare that the research was conducted in the absence of any commercial or financial
267 relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

268 PT, and PD contributed to conception and design of the study. PT performed the PEP-FOLD implementation
269 and ran the PEP-FOLD and Colabfold (AlphaFold2) simulations. PD ran the TrRosetta simulations. PT
270 and PD performed the analyses. PT, and PD wrote sections of the manuscript. All authors contributed to
271 manuscript revision, read, and approved the submitted version.

FUNDING

272 This work was supported by the “Initiative d’Excellence” program from the French State, Grant
273 “DYNAMO”, ANR-11-LABX-0011, and by INSERM U1133 recurrent funding.

ACKNOWLEDGMENTS

274 The authors thank Andrew Doig for helpful discussions about pKa.

FIGURE CAPTIONS

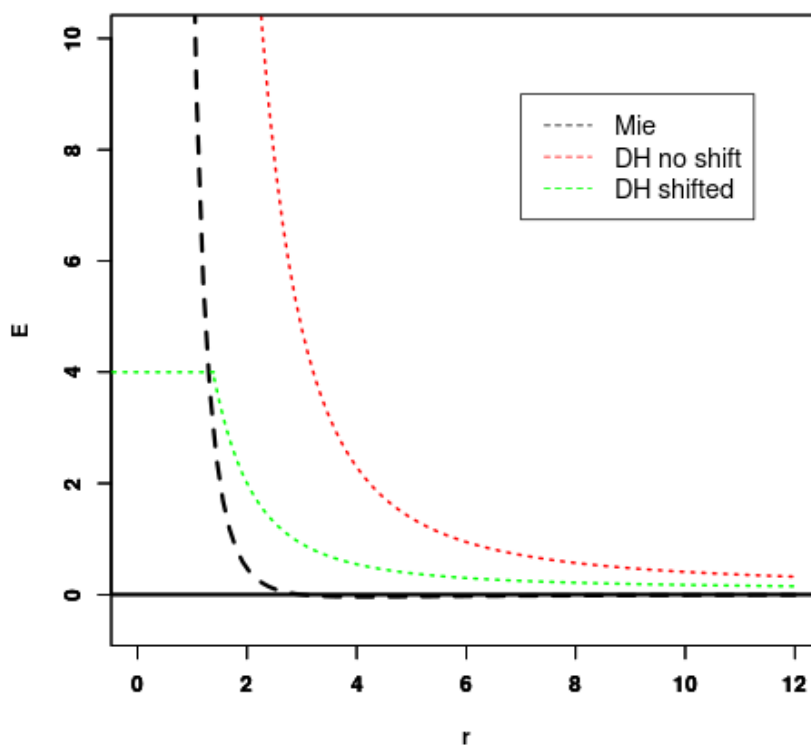


Figure 1. Fitting Debye-Hückel (DH) energy to Mie potential. The shift of the unshifted DH potential (red) is set so that the Mie (black) and shifted DH potential (green) cross for some energy threshold (4 kcal/mol in this case).

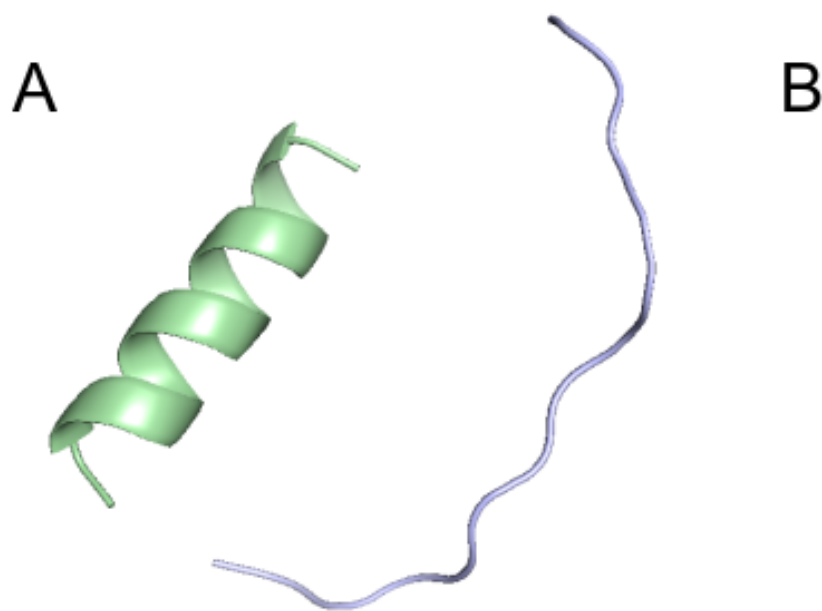


Figure 2. PF-DH Conformational ensemble of (K)15 as a function of pH. A: pH 7.4 B: pH 13. Only the lowest energy model (rank 1) is depicted.

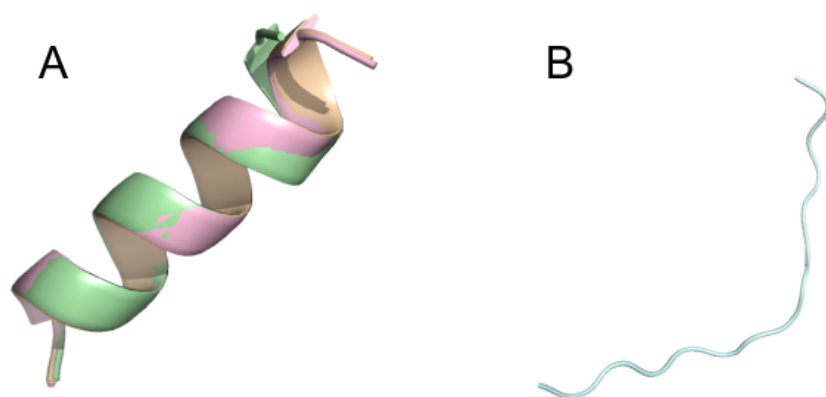


Figure 3. Conformational ensemble of (E)15 at pH7.4. A: PF-noDH, AlphaFold2 and TrRosetta B: PF-DH. Only the lowest energy (rank 1) model is depicted.

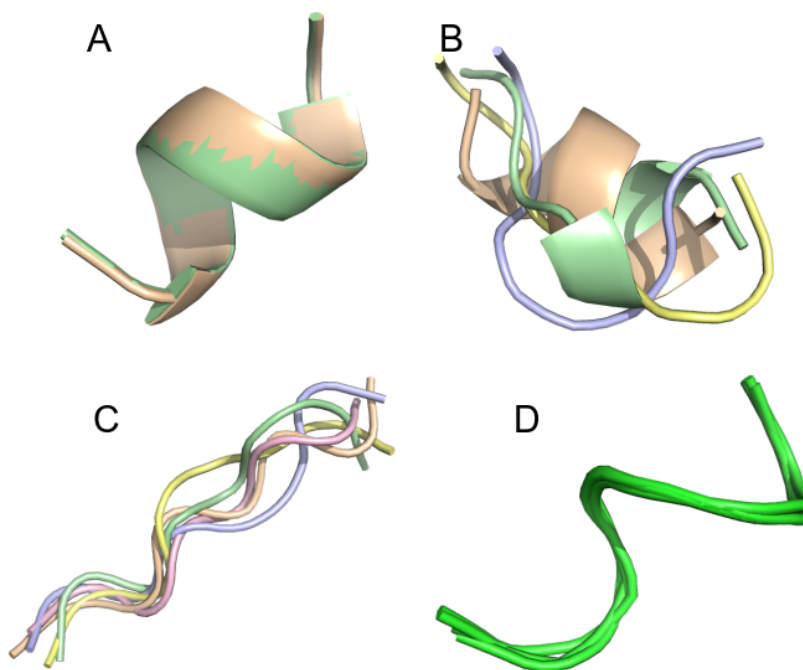


Figure 4. Conformational ensemble of 6nm3.

A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: NMR structure. For A, B and C, the 5 predicted models are depicted. For D, all models provided in the PDB are depicted.

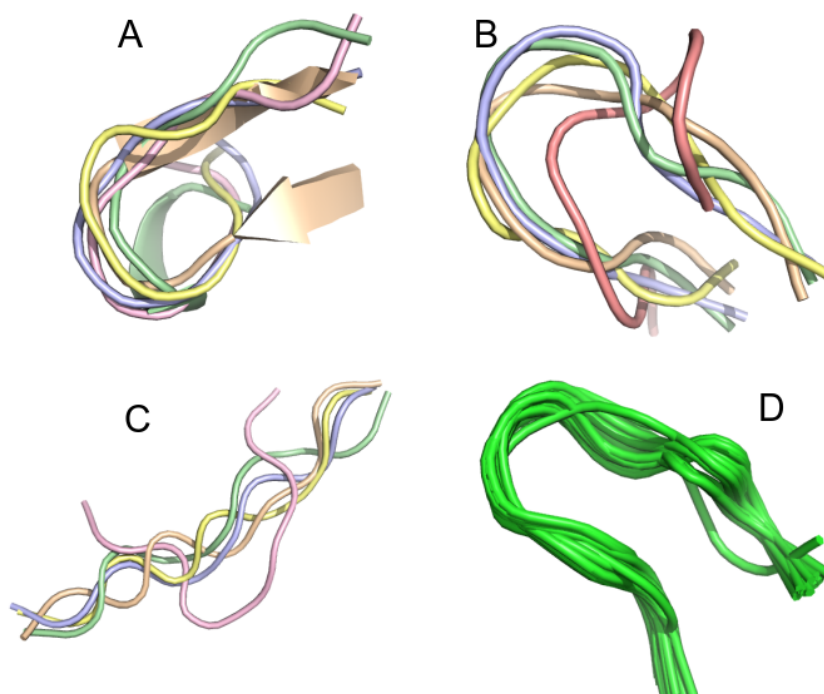


Figure 5. Conformational ensemble of 1egs.

A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: NMR structure. For A, B and C, the 5 predicted models are depicted. For D, all models provided in the PDB are depicted.

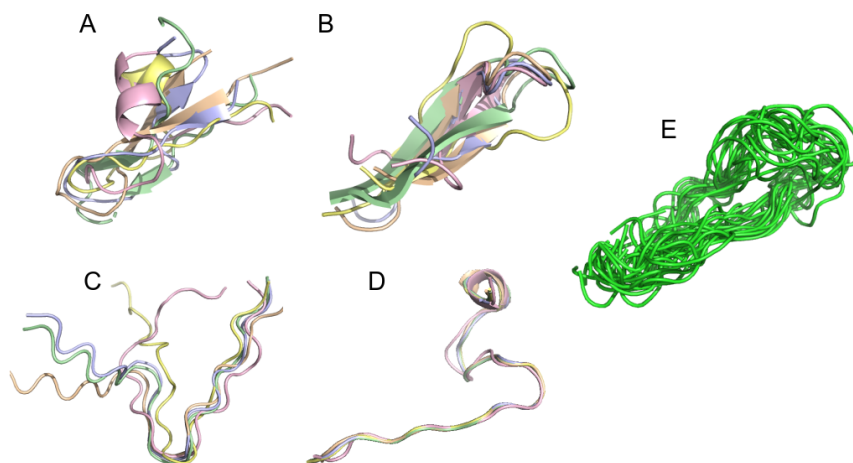


Figure 6. Conformational ensemble of 7li2.

A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: TrRosetta, E: NMR structure. For A, B, C and D the 5 predicted models are depicted. For E, all models provided in the PDB are depicted.

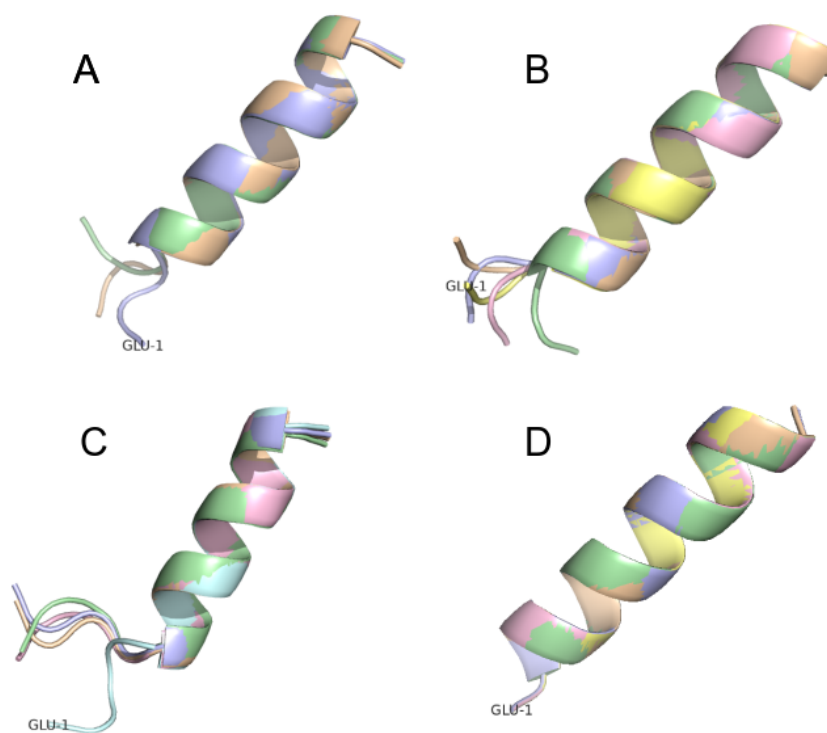


Figure 7. Conformational ensemble of pep17. A: PF-noDH, B: PF-DH - pH 2, C: AlphaFold2 and D: TrRosetta. For each method, the 5 predicted models are depicted.

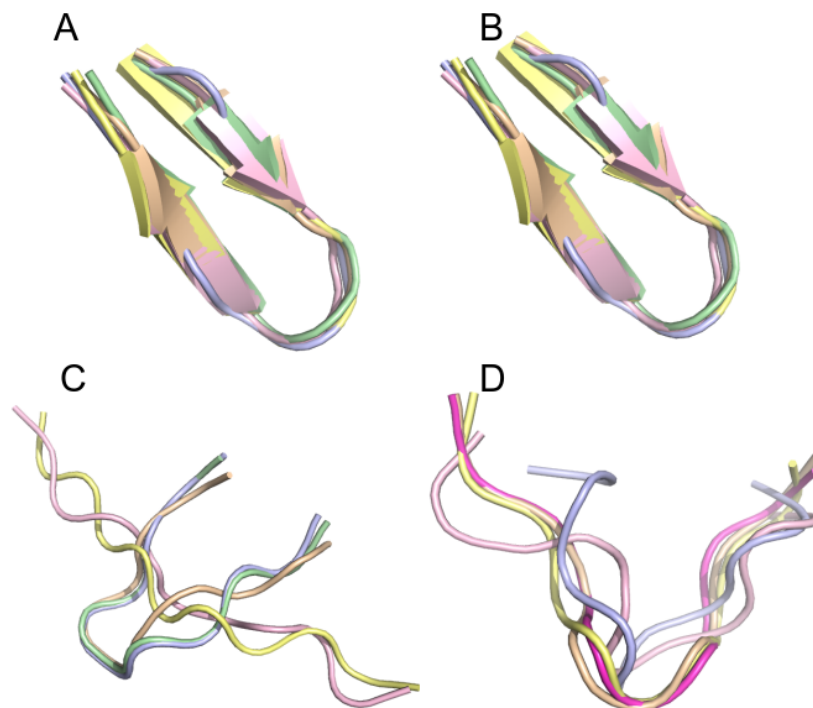


Figure 8. Conformational ensemble of pep10. A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: TrRosetta. For each method, the 5 predicted models are depicted.

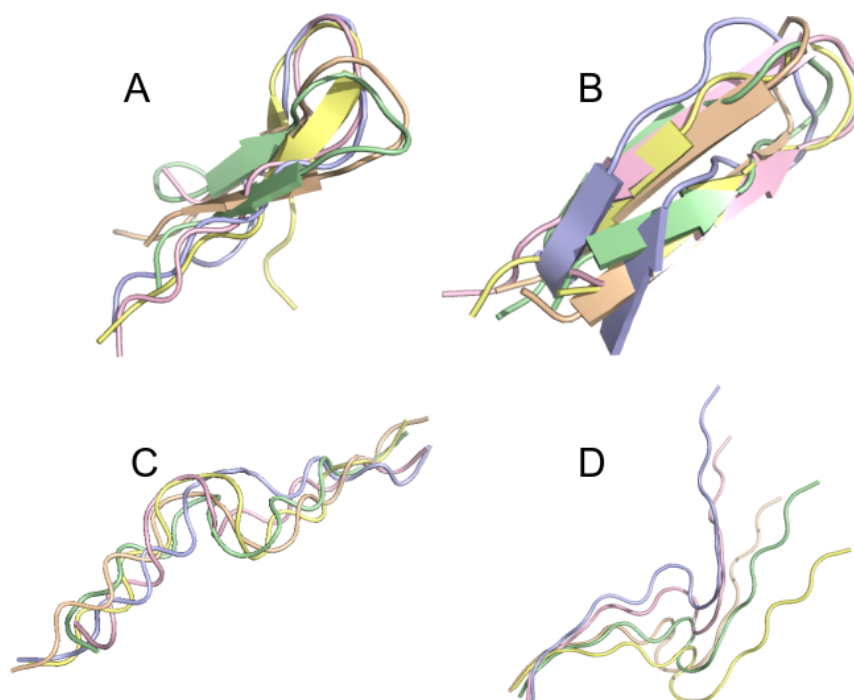


Figure 9. Conformation ensemble of tau-fragment at pH 7. A: PF-noDH, B: PF-DH, C: AlphaFold2 and D: TrRosetta. For each method, the 5 predicted models are depicted.