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

Pierre Tufféry, Philippe Derreumaux

frontiers

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

ABSTRACT

we selected the 5

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

1 INTRODUCTION

Peptides of less than 40 amino acids have diverse biological functions, acting as signaling entities in all domains of life, and targeting receptors or interfering with molecular interactions. Hormones and their bacterial mimetics (**?**), neuropeptides and their roles in neurodegenerative diseases (**?**), antimicrobial peptides contribution to host defence (**?**), and immunomodulatory peptides in the perspective of vaccine design (**?**) are some current directions motivating their study at a fundamental level. Due to their specific features, peptides have also gained interest as therapeutical agents **?**, particularly to target protein-protein interactions (**?**). They are also considered as having interest in the development of new functional biomimetic materials (**?**). Peptides have limitations though, as they can be highly flexible (**?**), which motivate efforts to understand and predict their conformational energy landscapes.
Structure prediction of polypeptides with amino acid lengths up to 40 amino acids in aqueous solution can be performed by a series of methods including machine-learning approaches such as AlphaFold2 (**?**), TrRosetta (**?**), and APPTEST (**?**). Looking at AlphaFold2, which revolutionized structure prediction of single folded domain to a root-mean-square deviation (RMSD) accuracy of 0.2 nm, its capability lies on machine learning based on protein data bank (PDB) (**?**) templates, multiple sequence alignments, co-evolution rules and sophisticated algorithms to predict local backbone and side conformations, and side chain - side chain contact probability within distances bins. AlphaFold2 builds the protein by energy minimization using a protein-specific energy potential.

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

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

APPTEST also uses machine learning on the PDB structures with a chain length varying between 5 and 40 amino acids. APPTEST derives Cα-Cα and Cβ-Cβ distance restraints, and backbone dihedral restraints that are input of simulated annealing and energy minimization.

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

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

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

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

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

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

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

2 METHODS

2.1 the sOPEP2 force field

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

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

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

$$E(\phi_i) = \epsilon_\phi \left(\phi_i - \phi_{0\_sc\_i}\right)^2 \quad (2)$$

where $\phi_{0\_sc\_i} = \phi$ within the interval $\left[\phi_{low\_sc\_i}, \phi_{high\_sc\_i}\right]$ and $\phi_{0\_sc\_i} = \min(\phi - \phi_{low\_sc\_i}, \phi - \phi_{high\_sc\_i})$ outside of the interval $\phi_{low\_sc\_i}$ and $\phi_{high\_sc\_i}$ are specific to each amino acid type (Binette et al. 2022).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$(13)$$

2.2 Debye-Hückel charge interactions

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

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

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

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

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

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

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

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

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

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

2.3 PEP-FOLD, TrRosetta and AlphaFold2 protocols and analysis

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

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

3 RESULTS AND DISCUSSION

3.1 Predicted Models of poly-charged peptides

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

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

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

The polypeptides (K)15 and (E)15 are particularly interesting because the alpha-helix content changes inversely with the pH. As observed by CD, the helical content of (K)15 increases with pH, while the helical 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 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 (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 and 100% coil at pH 11-13.

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

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

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

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

3.2 Predicted Models of polypeptides with NMR structures

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

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

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

The second result is that PF-noDH and PF-DH do not predict any low quality models (CAD-score < 0.6), while AlphaFold2 produces CAD-scores of 0.59, 0.55 and 0.6 for the three targets 6nm3,1egs and 7li2 (Figures 4, 5, 6, respectively). It has to be noted that the structures of these three peptides were solved at pH 5.8, 6.5 and 7. This low score results in differences between the experimental and predicted topologies. Experimentally, 6nm3 adopts a helical-like conformation, 1egs adopts a beta2-like conformation and 7li2 adopts a beta-2 like conformation. Of note, a beta-2 like conformation has the topology of a beta-hairpin but lacks the H-bond network.

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

The third result is related to the performances of PF-DH with respect to PF-noDH, which provides evidence that the weights of the Debye-Huckel salt bridge interactions are consistent with the weights of sOPEP2 interactions. It was far from being evident that the addition of charges at extremities and charged amino acids in the core of the sequences would not change the quality of the models for pH varying between 2.9 and 7. The number of titratable amino acids varies from 1-2 (1le1, 1egs - 6nm3, 6evq), 4 for 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 1fsd. The results also show that the pH-independent PEP-FOLD version and the pH-dependent PEP-FOLD version perform similarly for peptides containing charged, hydrophilic and hydrophobic amino acids.

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

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

3.3 Predicted Models of polypeptides without any NMR structures.

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

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

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

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

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

4 CONCLUSIONS

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

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

265 ## 4.1 Tables

	Experiment			PF-noDH			PF-DH			AlphaFold2			TrRosetta		
	Exp. Blck	pH	CD	α	coil	β-turn	α	coil	β-turn	α	coil	β-turn	α	coil	β-turn
		3					28	69	3						
(EK)15	-	7.4	β-turn	94	6	0	0	85	15	97	3	0	97	3	0
		11					24	68	8						
		3					0	100	0						
(EK)5	-	7.4	coil	90	10	0	0	68	32	54	10	36	68	26	6
		11					0	78	22						
(H)30	-	5	coil				0	100	0						
		7.4	aggregation	75	5	20		95	5	0	81	19	97	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						
		3-3.6	ND-42% α				93	7	0						
(E)15	Ace, Nmt.	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						

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. Blck.) using acetyl (Ace) or N-methyl (Nmt), pH and Circular Dichroism (CD) data, is reported. ND stands for not determined.

Target	L	# mod.	WDC	topo.	pH	ion.Str	Exp. Blck.	Sequence
6mi9	19	10	3-19	α distort.	4.3	0	Nmet.	PMARNKILGKILRKIAAFK
6j9p	12	10	2-12	α 5	0	-		RRLIRLILRLLR
1fsd	28	41	3-25	$\beta_2\alpha$	5	0	-	QQYTAKIKGRTFRNEKELRDFIEKFKGR
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	-	SKLPPGWEKRMSRNSGRVYYFNHITNASQFERPSG
2evq	12	43	2-10	β_2	6	20	-	KTWNPATGKWTE
1egs	9	20	1-9	β_2-like	6.5	0	Ace, Nmet.	TKSAGGIVL
6r2x	25	15	9-20	Cα	6.5	148	-	FETLRGDERILSILRHQNLLKELQD
7b2f	31	20	6-21	$\alpha C\alpha$	6.5	100	-	MNNNELTSLPLAERKRLLELAKAAKLSRQHY
6s0n	9	20	ND	α-like	6.8	1	-	QDVNTAVAW
7li2	22	20	1-22	β_2-like	7	0	-	AGTMRVTYPDGQKPGQSDVEKD
1wbr	17	32	1-16	αC	7	0	Ace, Nmet.	QAERMSQIKRLLSEKKT
1pgbF	16	1	ND	β_2	7	0		GEWTYDDATKTFTVTE
pep17	17	-		α	2	0	-	ETGTKAELLAKYEATHK
pep38	38	-		$\alpha T\alpha$	3.6	20	-	DWLKARVEQELQALEARGTDSNAELRAMEAKLKAEIQK
pep10	11	-		β_2	4.3	0	-	IYSNSDGWTWT
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 strenght (ion. Str.) and the presence of extra groups to block the N terminus (acetyl) and C terminus (N methyl) (Exp. Blck.), 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.

	Exp.		PF-noDH			PF-DH			AlphaFold2			TrRosetta		
Target	topo	pH	topo	Full	WDC	topo	Full	WDC	topo	Full	WDC	topo	Full	WDC
6mi9	α	4.3	α	0.85	0.86	α, α-like	0.84	0.86	α	0.83	0.86	α	0.84	0.84
6j9p	α	5	α	0.75	0.77	Cα	0.70	0.71	α	0.74	0.76	α	0.74	0.75
1fsd	$\beta_2 alpha$	5	$\alpha_2\alpha$	0.62	0.63	$\beta 2\alpha$	0.66	0.69	$\beta_2\alpha$	0.73	0.76	$\beta_2\alpha$	0.71	0.73
1j4m	β_2	5	β_2	0.67	0.67	β_2	0.76	0.76	β_2	0.74	0.74	β_2	0.73	0.73
1le1	β_2	5.5	β_2	0.76	0.8	β_2	0.75	0.75	β_2	0.83	0.87	β_2	0.8	0.79
6nm3	α-like	5.8	α	0.72	0.72	α,unstr	0.73	0.73	coil	0.59	0.59	-	-	-
6svc	β_3	6	β_3	0.66	0.66	β_3	0.70	0.70	β_3	0.79	0.8	β_3	0.78	0.79
2evq	β_2	6	β_2	0.84	0.86	β_2	0.82	0.86	β_2	0.89	0.93	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	0.70	coil	0.55	0.55	-	-	-
6r2x	Cα	6.5	Cα	0.75	0.9	Cα	0.74	0.91	α	0.8	0.92	Cα	0.8	0.9
7b2f	αCα	6.5	$\alpha C, alpha$	0.77	0.82	$\alpha C\alpha$	0.79	0.84	$\alpha C\alpha$	0.79	0.84	$\alpha C\alpha$	0.77	0.83
6s0n	α-like	6.8	α	0.76	0.76	α	0.79	0.78	coil	0.67	0.67	-	-	-
7li2	β_2-like	7	β_2, β_2-like	0.62	0.62	β_2, β_2-like	0.64	0.64	coil	0.6	0.6	coil	0.58	0.58
1wbr	αC	7	α	0.69	0.7	α	0.70	0.70	α	0.69	0.7	α	0.68	0.69
1pgbF	β_2	7	β_2	0.77	0.77	β_2	0.79	0.78	$\beta_2, ext - untr$	0.83	0.83	β_2	0.91	0.91
MEAN				0.73	0.75		0.74	0.76		0.74	0.76		0.76	0.78
STDEV				0.07	0.09		0.06	0.08		0.10	0.12		0.08	0.09
MEDIAN				0.75	0.76		0.74	0.75		0.74	0.76		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

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

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the "Initiative d'Excellence" program from the French State, Grant "DYNAMO", ANR-11-LABX-0011, and by INSERM U1133 recurrent funding.

ACKNOWLEDGMENTS

The authors thank Andrew Doig for helpful discussions about pKa.

FIGURE CAPTIONS

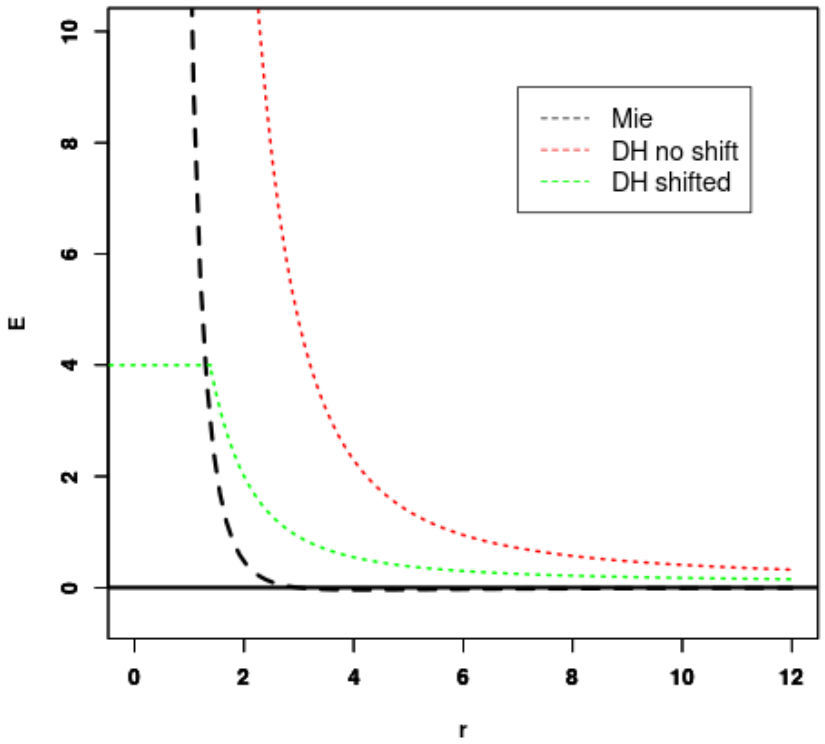

Figure 1. Fitting Debye-Hückel (DH) energy to Mie potential. The shift of the unshifted DH potential (red) is set so that the Mie (black) and shifted DH potential (green) cross for some energy threshold (4 kcal/mol in this case).

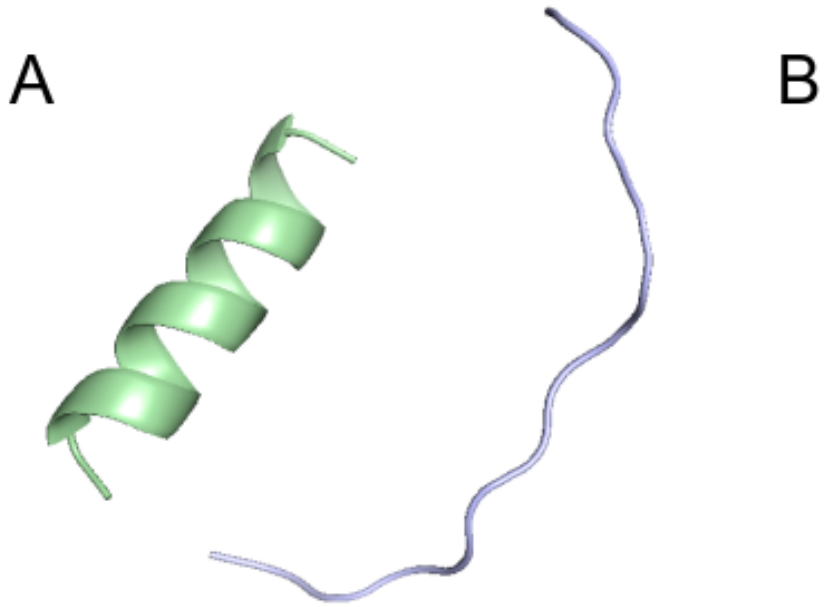

Figure 2. PF-DH Conformational ensemble of (K)15 as a function of pH. A: pH 7.4 B: pH 13. Only the lowest energy model (rank 1) is depicted.

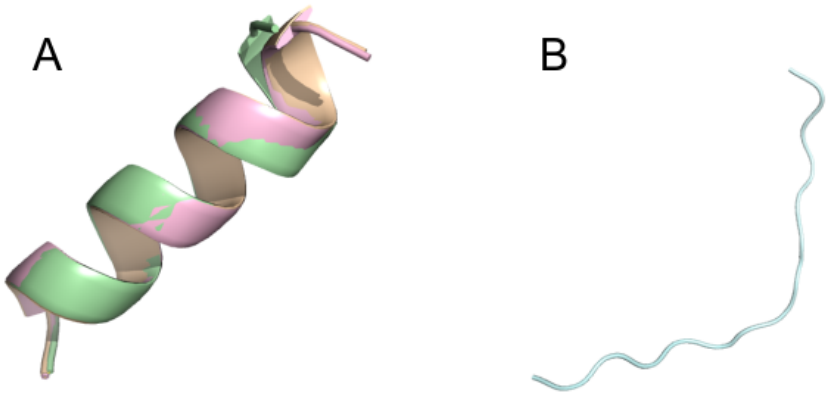

Figure 3. Conformational ensemble of (E)15 at pH7.4. A: PF-noDH, AlphaFold2 and TrRosetta B: PF-DH. Only the lowest energy (rank 1) model is depicted.

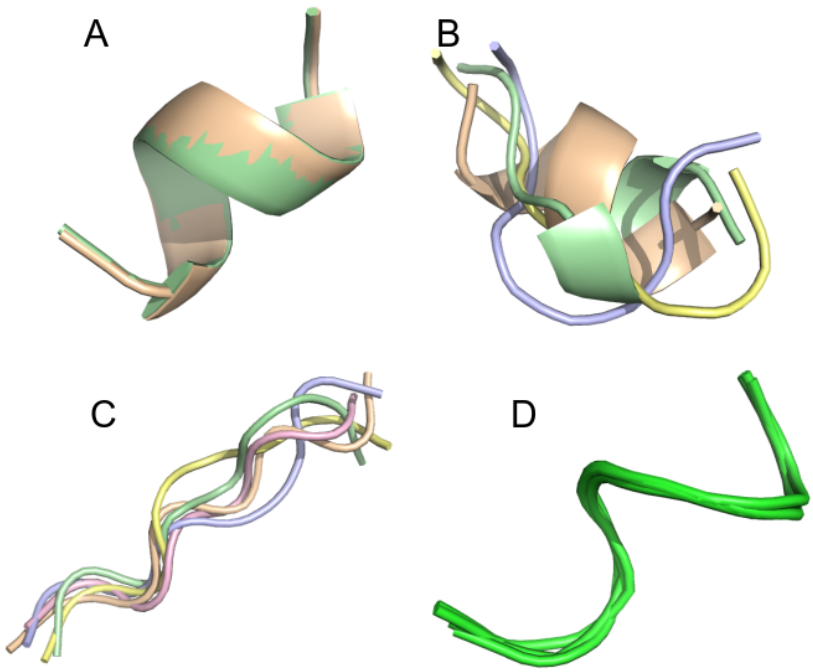

Figure 4. Conformational ensemble of 6nm3.
A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: NMR structure. For A, B and C, the 5 predicted models are depicted. For D, all models provided in the PDB are depicted.

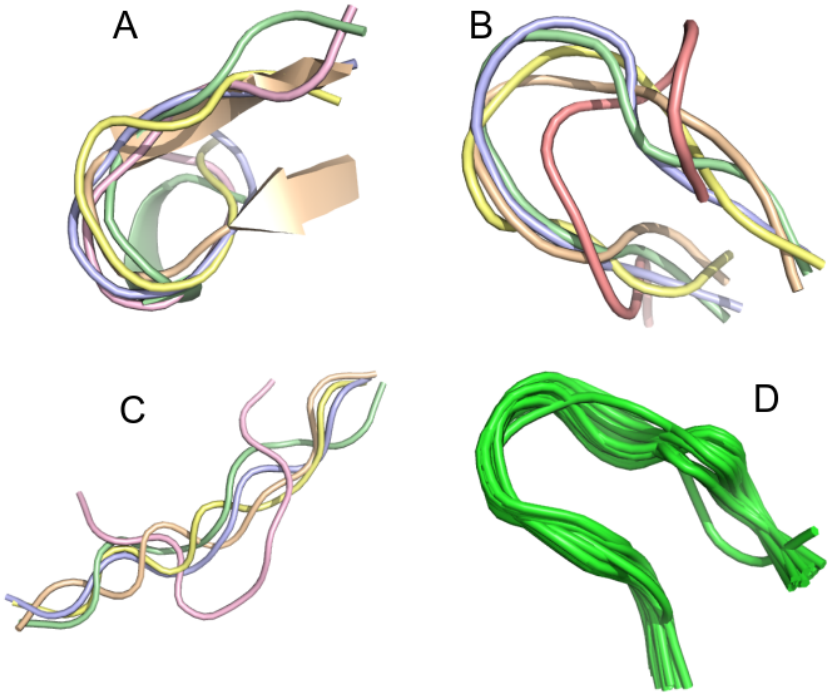

Figure 5. Conformational ensemble of 1egs.
A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: NMR structure. For A, B and C, the 5 predicted models are depicted. For D, all models provided in the PDB are depicted.

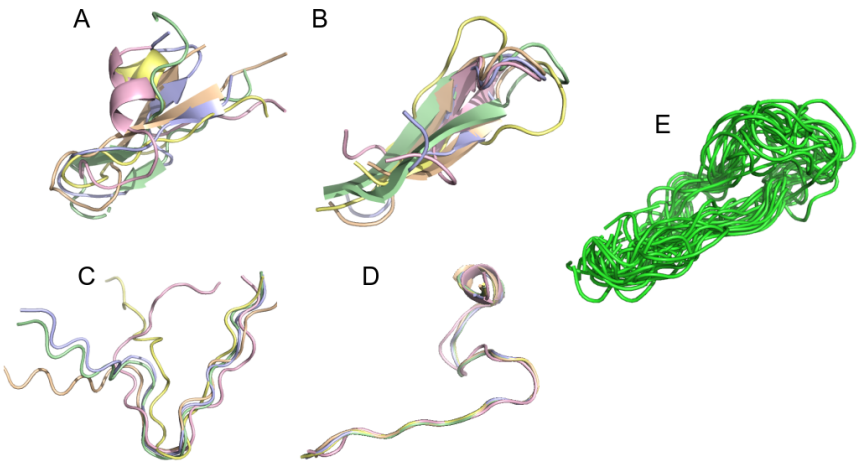

Figure 6. Conformational ensemble of 7li2.
A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D:TrRosetta, E: NMR structure. For A, B, C and D the 5 predicted models are depicted. For E, all models provided in the PDB are depicted.

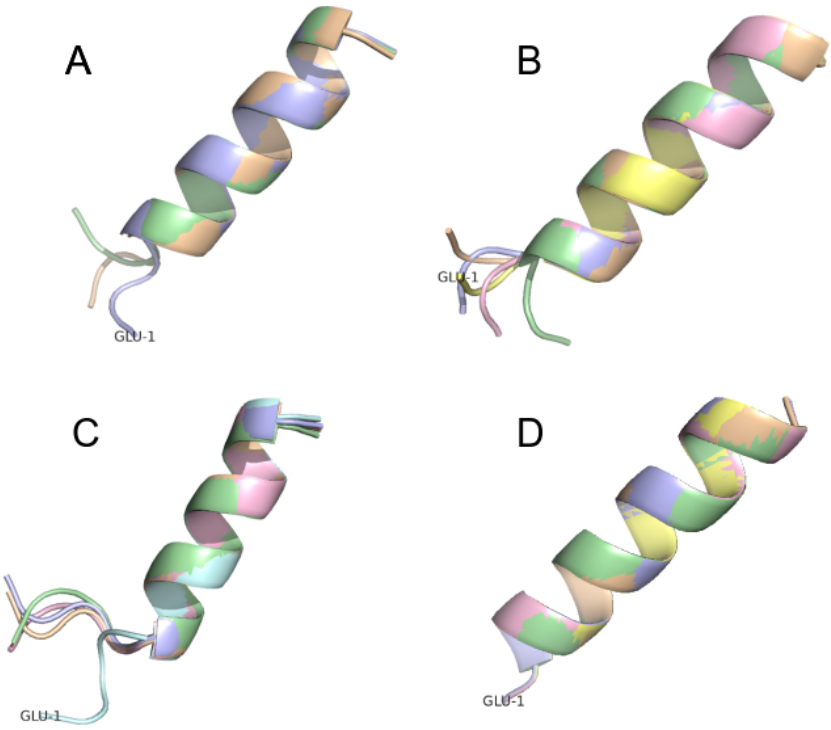

Figure 7. Conformational ensemble of pep17. A: PF-noDH, B: PF-DH - pH 2, C: AlphaFold2 and D: TrRosetta. For each method, the 5 predicted models are depicted.

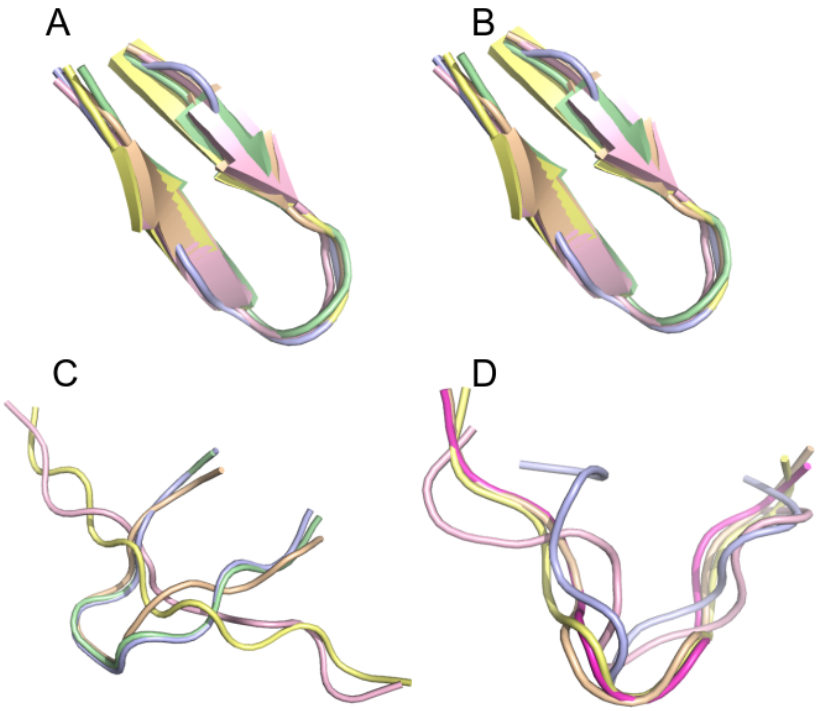

Figure 8. Conformational ensemble of pep10. A: PF-noDH, B: PF-DH at pH 4.3, C: AlphaFold2, D: TrRosetta. For each method, the 5 predicted models are depicted.

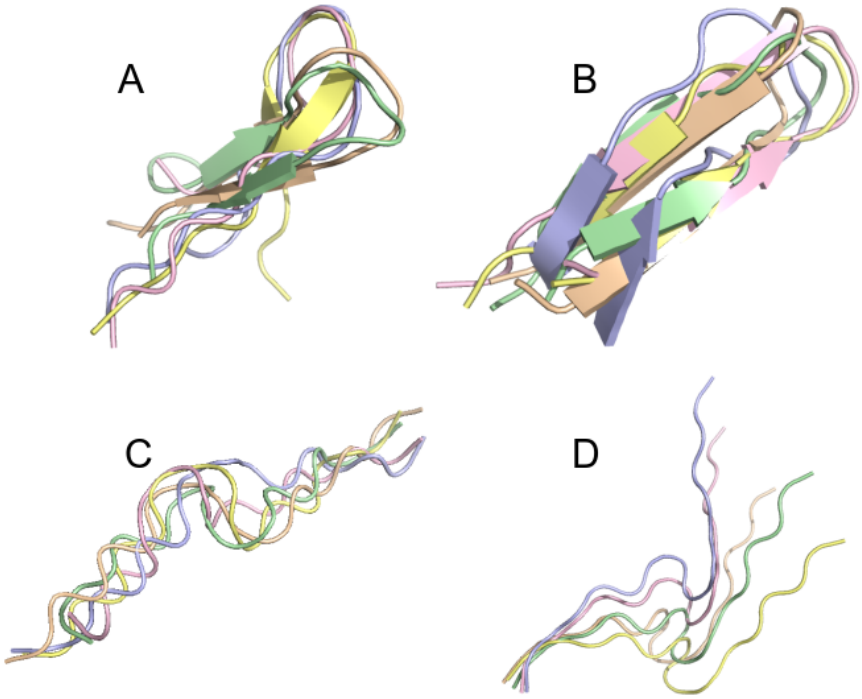

Figure 9. Conformation ensemble of tau-fragment at pH 7. A: PF-noDH, B: PF-DH, C: AlphaFold2 and D: TrRosetta. For each method, the 5 predicted models are depicted.