Supplementary Materials

AFflecto: A web server to generate conformational ensembles of flexible proteins from AlphaFold models

Mátyás Pajkos[1], Ilinka Clerc[1], Christophe Zanon[1], Pau Bernadó[2], Juan Cortés[1]

[1] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.*
[2] *Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, Montpellier, France.*

1 Identification of conditionally folded secondary structural elements

1.1 Experimentally verified information

To develop the method for identifying conditionally folded secondary structural elements (SSEs), datasets of natively folded and conditionally folded regions were first established based on experimental data. DSSP [1] was then used to identify SSEs in each region of both datasets, identifying natively and conditionally folded SSEs. The exact process is outlined below.

The natively folded SSE dataset was based on 329 PDB entries collected Lui *et al.* [2] to establish a high-quality structure set without any disorder-to-order transitions. SSE assignment by DSSP was applied to 327 of the 329 entries (excluding 3GOT and 1C53), considering only SSEs of at least 3 residues in length and classified as H, B, E, G, I secondary structure type. This resulted in a total of 5716 SSEs, representing the natively folded SSEs and serving as true positives in the conditionally folding SSE identification method.

To collect conditionally folded SSEs, we used the dataset of experimentally verified conditionally folders (CFs) compiled in a recent study on the systematic identification of intrinsically disordered conditionally folded regions using AlphaFold2 [3]. This dataset included known CFs with PDB files from five databases: Disordered Binding Sites (DIBS), Mutual Folding Induced by Binding (MFIB), DisProt, Molecular Recognition Feature (MoRF), and FuzDB [4, 5, 6, 7, 8]. The known CFs were mapped to rigid regions predicted by AFflecto, and only those rigid segments that overlapped with the known CFs by at least 90% were retained. This resulted in 84, 169, 169, and 31 rigid regions from DIBS, MFIB, DisProt, and MoRF, respectively, with no rigid regions covered by FuzDB CFs at the 90% threshold. Following this, similar to the natively folded dataset, SSEs were identified within the retained rigid regions using DSSP, yielding a total of 1388 SSEs for the conditionally folded SSEs dataset, which was used as true negatives.

1.2 Analysis of tertiary contact numbers between SSEs

Next, for both datasets, the tertiary contacts formed by each SSE with other SSEs were calculated. Concretely, for each SSE, the number of contacts with other SSEs was counted, and the per-residue contact number was determined by dividing this number by the length of the SSE (requiring at least two contacts between SSEs; otherwise the per-residue contact number is zero). In the natively folded dataset, 5,716 SSEs were analyzed. In the conditionally folded (CF) SSE dataset, contact numbers were calculated for 1,388 SSEs. The distributions of per-residue contacts for both datasets are summarized in Fig. 1a, clearly showing that fewer contacts are more typical for CF SSEs than for natively folded ones.

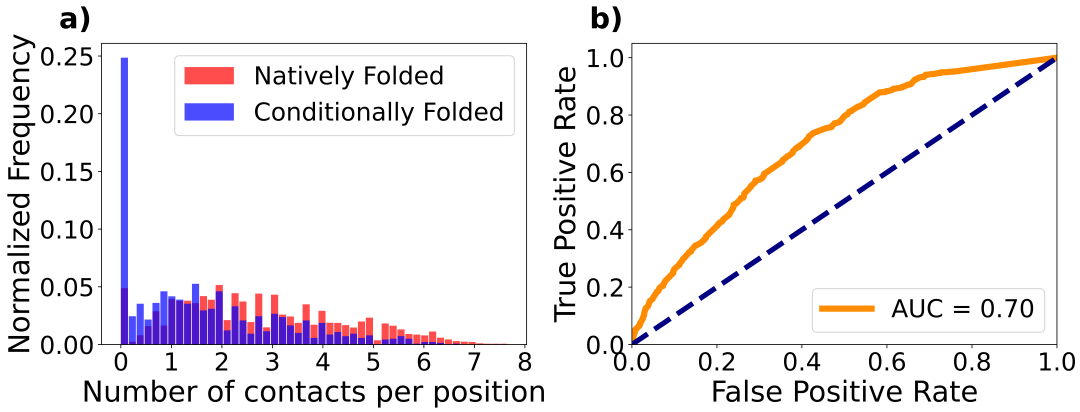

Figure 1: Comparison of Contact Distributions and Classification Performance for Natively and Conditionally Folded SSEs (a) Comparison of the average number of contacts formed between SSEs in the natively folded and conditionally folded datasets. (b) Receiver Operating Characteristic (ROC)

1.3 Application of contact number as a discriminator between natively and conditionally folded SSEs

Based on the per-residue contact number distributions of conditionally and natively folded SSEs, we assessed the effectiveness of using the calculated tertiary contacts as a discriminative feature to distinguish natively folded SSEs from conditionally folded ones. To achieve this, we performed a Receiver Operating Characteristic (ROC) analysis, which evaluates the performance of a binary classification model. In the ROC analysis we applied a sliding cutoff on the per-residue contact number to classify the SSEs as 1 (natively folded, when the contact number is higher than the cutoff) or 0 (conditionally folded, when the contact is lower than the cutoff), and compared the results to the ground truth. We calculated the Area Under the Curve (AUC), a metric that quantifies the overall ability of the model to distinguish between the two classes (a value of 1 indicating perfect discrimination). Our analysis yielded an AUC of 0.70 (Fig. 1b). We also calculated the F1 score, a harmonic mean of precision and recall that measures the balance between false positives and false negatives in the classification. The maximum F1 score obtained was 0.89, with a corresponding true-positive rate of 0.94, demonstrating that the model effectively captures the majority of positive cases. However, the false-positive rate was relatively high at 0.69. Despite this, in our scenario, capturing true positives is crucial, thus the trade-off with false positives is acceptable. For the 0.89 F1 score max, the optimal cutoff on the per-residue contact number was determined to be 0.474. This cutoff is used as the default parameter on the web server.

2 Structural investigation of the Tau N-terminal region

For the structural investigation of the C-terminal tail of the Tau40 protein, we truncated the full-length PDB file to retain only the last 32 residues (410–441) and generated 1000 conformers using the default settings of AFflecto. To identify structural patterns within the ensemble, contact-based clustering was performed using WARIO [9]. The clustering resulted in two clusters, containing 44.8% and 22.5% of the conformers for clusters #0 and #1, respectively (the remaining 32.7% of the conformations were not clustered). For each cluster, a cluster-specific contact map and DSSP-based heatmap were generated (Fig. 2). In cluster #0, inter-residue contacts were identified in the middle of the tail sequence (see contact map), suggesting that this region exhibits a shared partially structured element among the conformers. The DSSP-based structural descriptor (heatmap) revealed that this partially structured element predominantly appears as an α-helix. In contrast, no partially structured elements were observed in cluster #1, indicating that this conformer population is represented as random coils within the ensemble. These analyses clearly show that the conformational ensemble generated by AFflecto captures the helical propensity of this region.

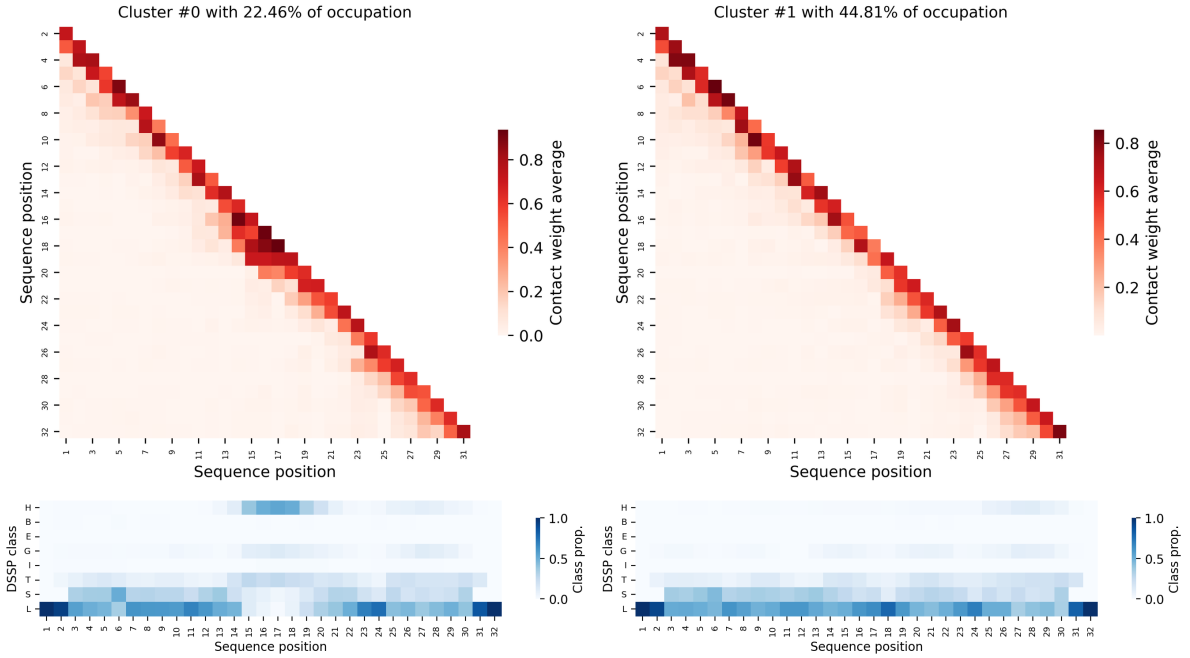

Figure 2: Cluster-specific contact map and DSSP-based heatmap descriptor for clusters #0 and #1 (left and right, respectively). The sequence numbering in the figures starts from 1, corresponding to residue 410 of the analyzed tail region.

References

[1] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 12, 2577–2637.

[2] Liu, Y., Wang, X., and Liu, B. (2021). RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. *Brief Bioinform* **22**, 2, 2000–2011.

[3] Alderson, T. R., Pritišanac, I., Kolarić, , Moses, A. M., and Forman-Kay, J. D. (2023). Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc Natl Acad Sci USA* **120**, 44, e2304302120.

[4] Schad, E., Fichó, E., Pancsa, R., Simon, I., Dosztányi, Z., and Mészáros, B. (2018). Dibs: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**, 3, 535–537.

[5] Fichó, E., Pancsa, R., Magyar, C., Kalman, Z. E., Éva Schád, Németh, B. Z., Simon, I., Dobson, L., and Tusnády, G. E. (2024). MFIB 2.0: a major update of the database of protein complexes formed by mutual folding of the constituting protein chains. *Nucleic Acids Res*, gkae976.

[6] Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., Pajkos, M., Lazar, T., Peña-Díaz, S., Santos, J., Ács, V., Farahi, N., Fichó, E., Aspromonte, M. C., Bassot, C., Chasapi, A., Davey, N. E., Davidović, R., Dobson, L., Elofsson, A., Erdős, G., Gaudet, P., Giglio, M., Glavina, J., Iserte, J., Iglesias, V., Kálmán, Z., Lambrughi, M., Leonardi, E., Longhi, S., Macedo-Ribeiro, S., Maiani, E., Marchetti, J., Marino-Buslje, C., Mészáros, A., Monzon, A. M., Minervini, G., Nadendla, S., Nilsson, J. F., Novotný, M., Ouzounis, C. A., Palopoli, N., Papaleo, E., Pereira, P. J. B., Pozzati, G., Promponas, V. J., Pujols, J., Rocha, A. C. S., Salas, M., Sawicki, L. R., Schad, E., Shenoy, A., Szaniszló, T., Tsirigos, K. D., Veljkovic, N., Parisi, G., Ventura, S., Dosztányi, Z., Tompa, P., Tosatto, S. C. E., and Piovesan, D. (2022). Disprot in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res* **50**, D1, D480–D487.

[7] Yan, J., Dunker, A. K., Uversky, V. N., and Kurgan, L. (2016). Molecular recognition features (morfs) in three domains of life. *Mol Biosyst* **12**, 3, 697–710.

[8] Hatos, A., Monzon, A. M., Tosatto, S. C. E., Piovesan, D., and Fuxreiter, M. (2022). FuzDB: a new phase in understanding fuzzy interactions. *Nucleic Acids Res* **50**, D1, D509–D517.

[9] González-Delgado, J., Bernadó, P., Neuvial, P., and Cortés, J. (2024). Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins. *Bioinformatics* **40**, 11 (10), btae627.