Conformational analysis of protein structural ensembles

Claudia Herron Mulet (2029817), Leonardo Amato (2028621) and Matteo Marchesin (1234385)

In this project, we were asked to study the structure and dynamics of the Intrinsically Disordered Protein MKK7. We were given 5 conformational ensembles. On the one hand, our goal was to compute some single conformation features that we could use to compare the individual conformations in the ensemble. On the other hand, we were asked to compare the 5 different ensembles using a global and a local score computed through the ensemble features.

Task 1

The main objective of task 1 is to find relationships between the structures in an ensemble, given as input. The task is divided into 3 main sections: the creation of a feature file,a graph and a pymol image.

Single conformation features

1. **Radius of gyration of the structure**
   For each structure, we compute the radius of gyration using the barycenter formula, instead of the center of mass.
2. **Relative accessible surface area (ASA) for each residue**
   To compute the relative accessible surface area we first compute the half sphere exposure. The number of residues in the upper half sphere (HSE-up) is inversely proportional to the accessible surface area, which means that we just need to compute a min-max normalization of the HSE-up and then invert it in the 0-1 range to get an approximated relative ASA.
3. **Secondary structure classification**
   In each single conformation, the secondary structure (SS) has been predicted for each residue. Because of the lack of hydrogen bond, the DSSP is not the best method to evaluate secondary structure in intrinsically disordered proteins. For this reason, the SS the values of phi/psi angles have been used to classify the residue in the following categorical classes: α-helix left-handed helix, β-strand, polyproline I and II.
4. **Residue distance matrix considering Cα atoms**
   For each conformation, we extract the distance between the alpha carbons of each residue.

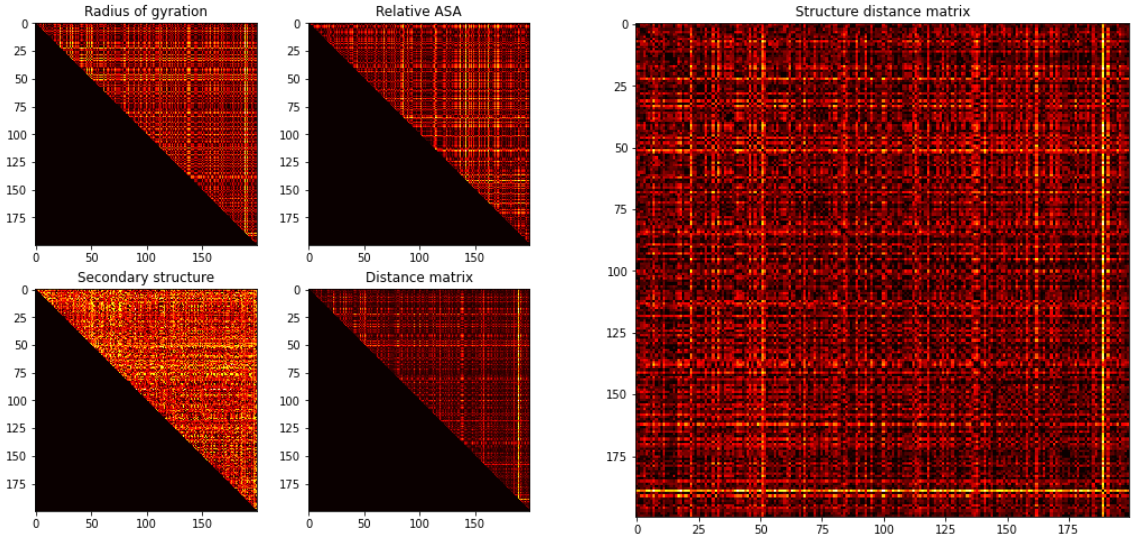
All the features of each structure in the ensemble are then saved in a single json file, to be used in the task2.

Representative conformations graph

To compute the representative we created a feature distance matrix (M,M) between conformations, that combined all the features of a conformation.

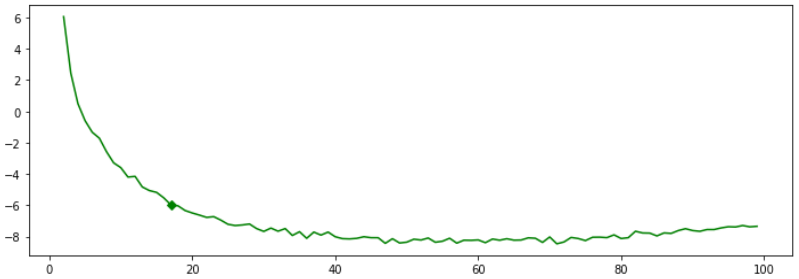
1. The distance between two radii of gyration is the absolute value of the difference.
2. The distance between two rASA vectors is the absolute value of the difference of the mean.
3. The distance between two secondary structure classification vectors is calculated with the SequenceMatcher function.
4. The distance between two residues distance matrix is calculated with the mean square error function.

We standardized each of the four features distance matrices with the mean and the standard deviation and then created a single distance matrix with the simple element wise sum.



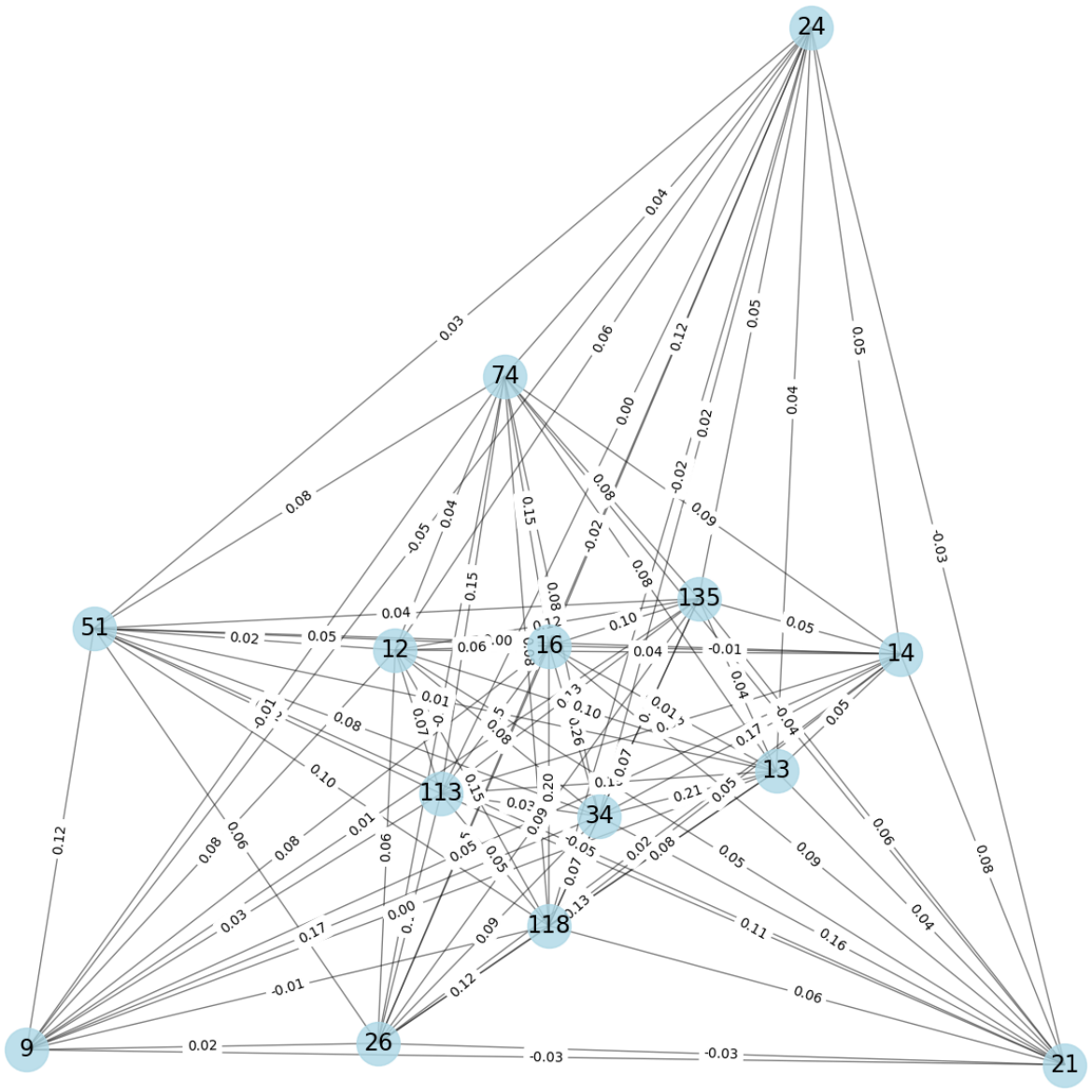
on the left, structure distance matrix for each feature, on the right, all the distance matrix combined

The distance matrix is then passed to the kmedoids function with variable ***k*** in the range *2-200*. For each ***k*** the function computes the mean distance between all the medoids (most central point in each cluster). These values create an elbow function, which we can use to get the optimal number of clusters to use. The elbow point (optimal k) is automatically calculated by checking how the linear approximation of the derivative of the elbow function changes over k.

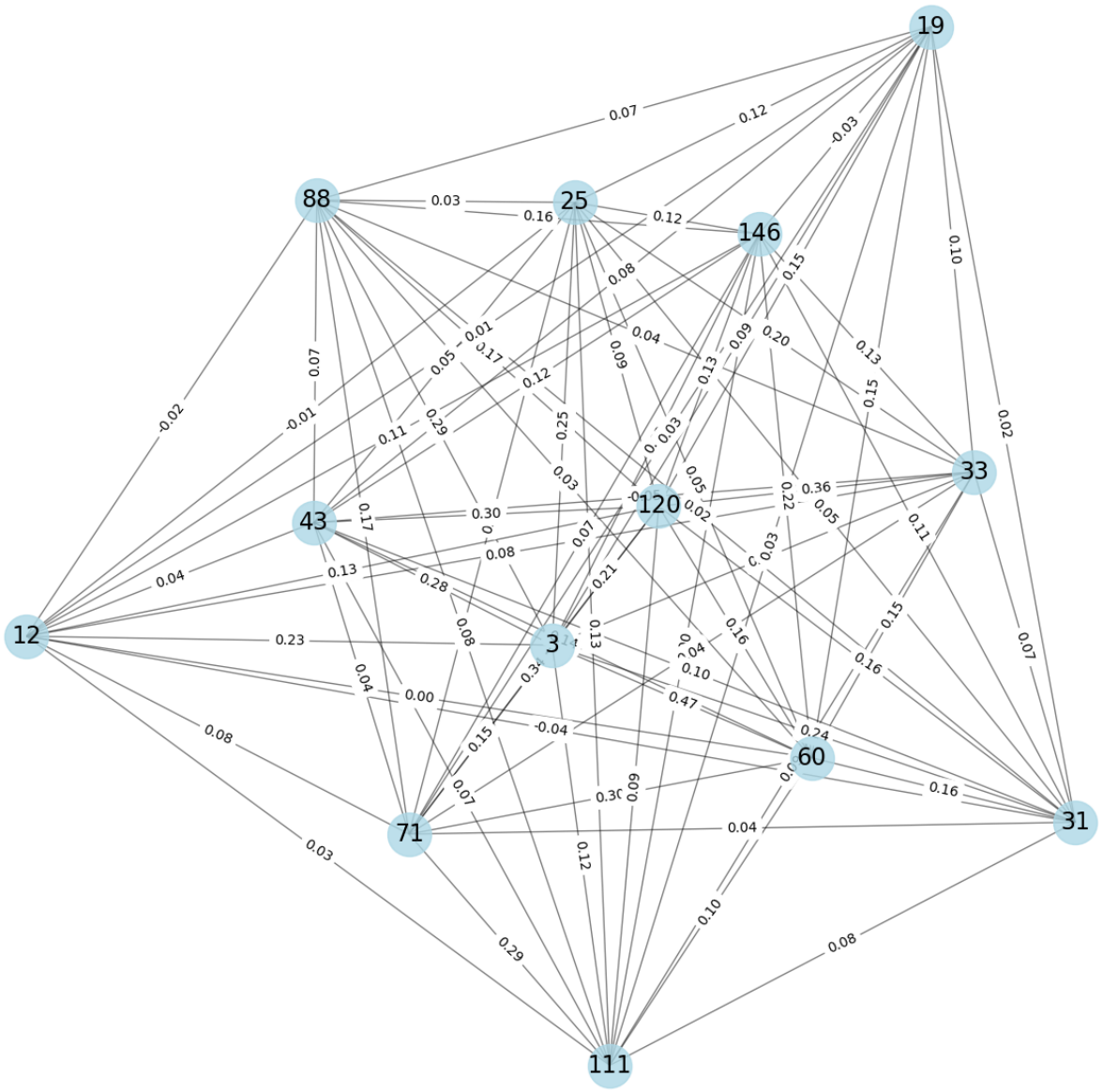


In the end the medoids of each cluster have been selected as representative of the cluster and as a node for the graph. The optimal number of clusters for PED00153 always falls around 12-20.

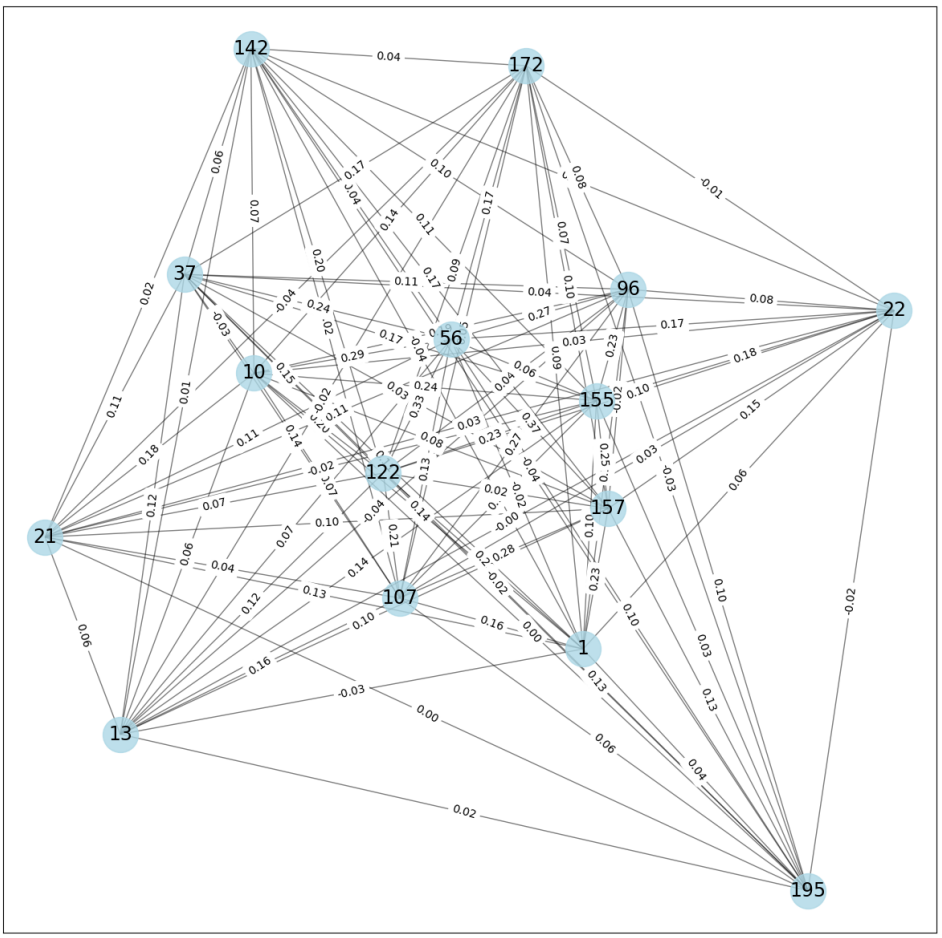
PED153e007 Graph



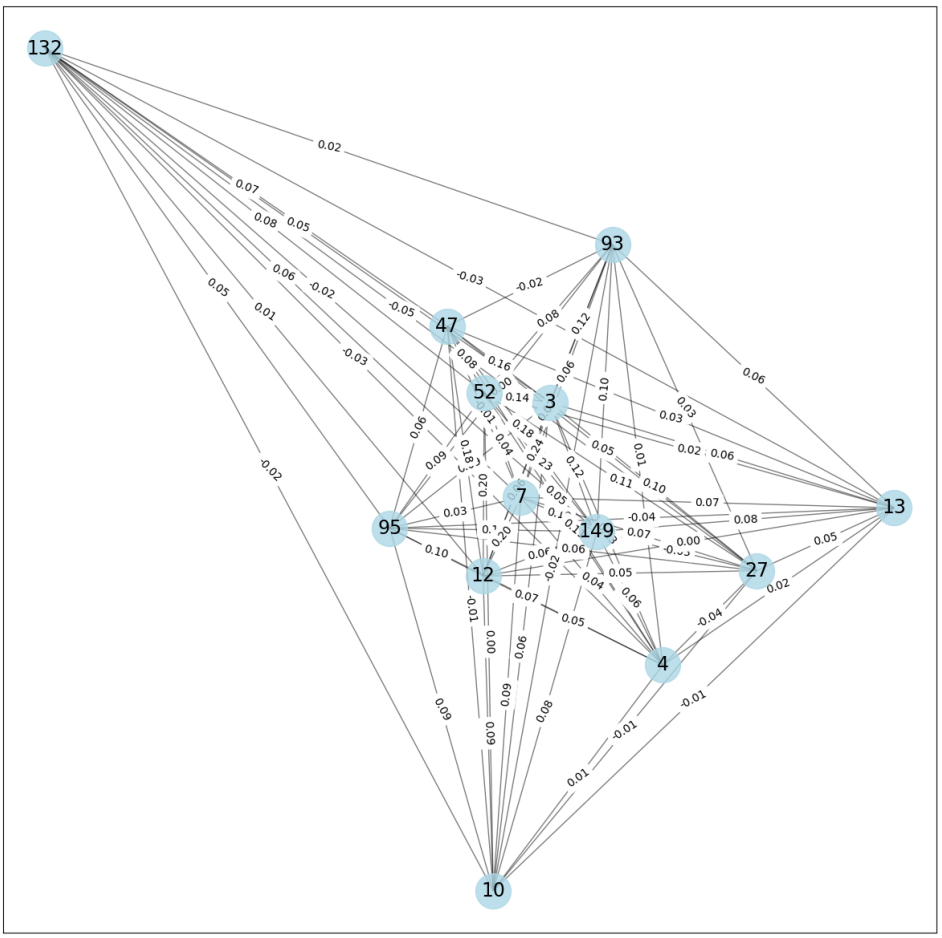
PED153e008 Graph



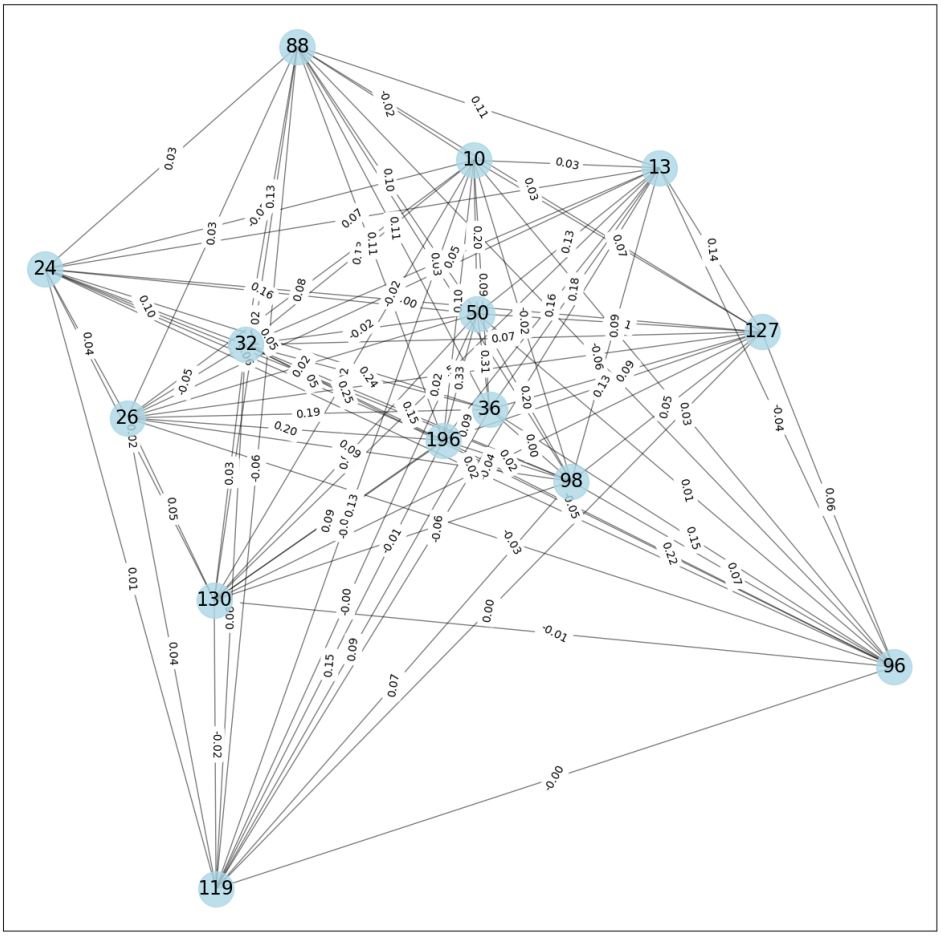
PED153e009 Graph



PED153e010 Graph



PED153e011 Graph

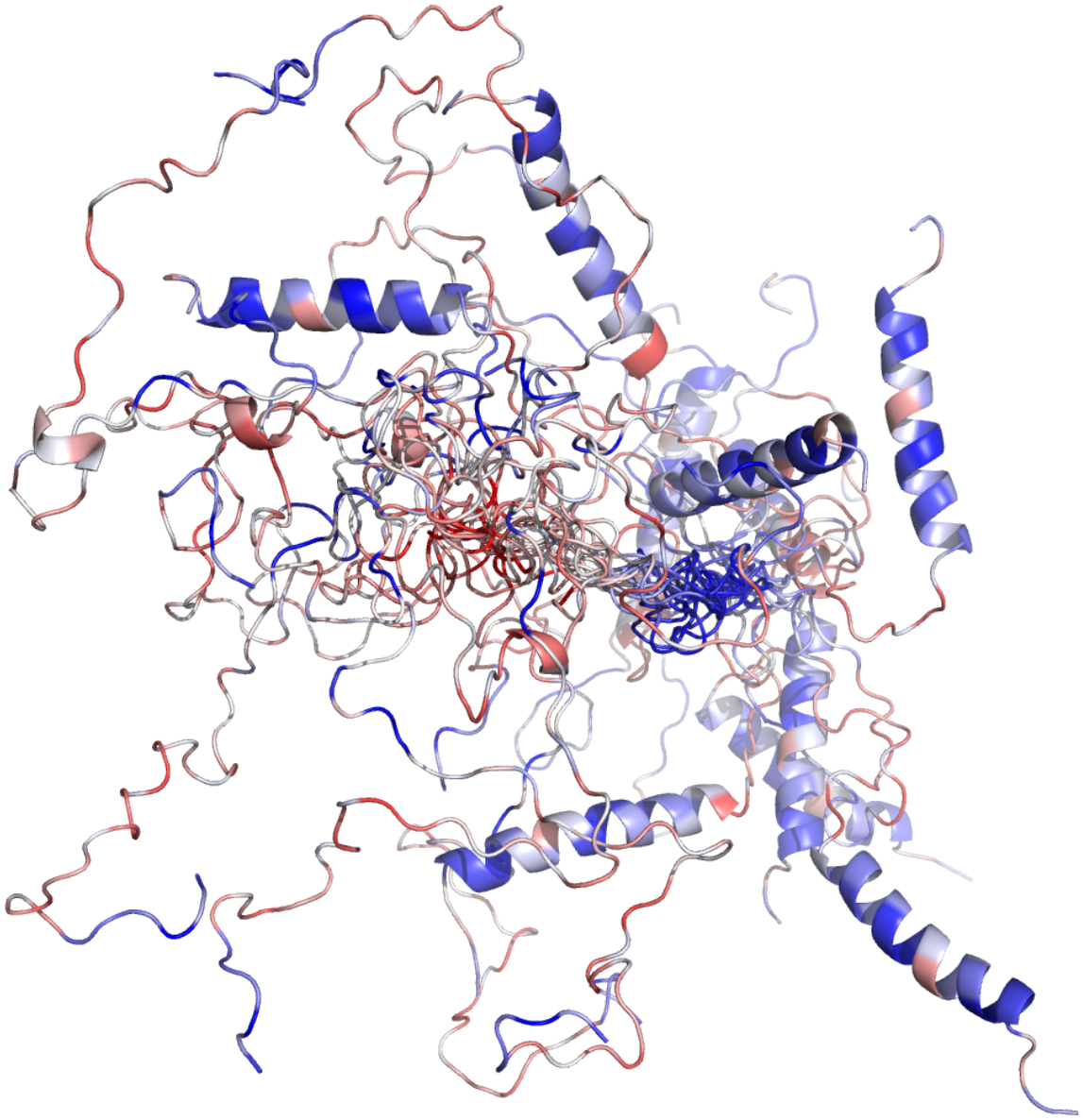


Pymol image

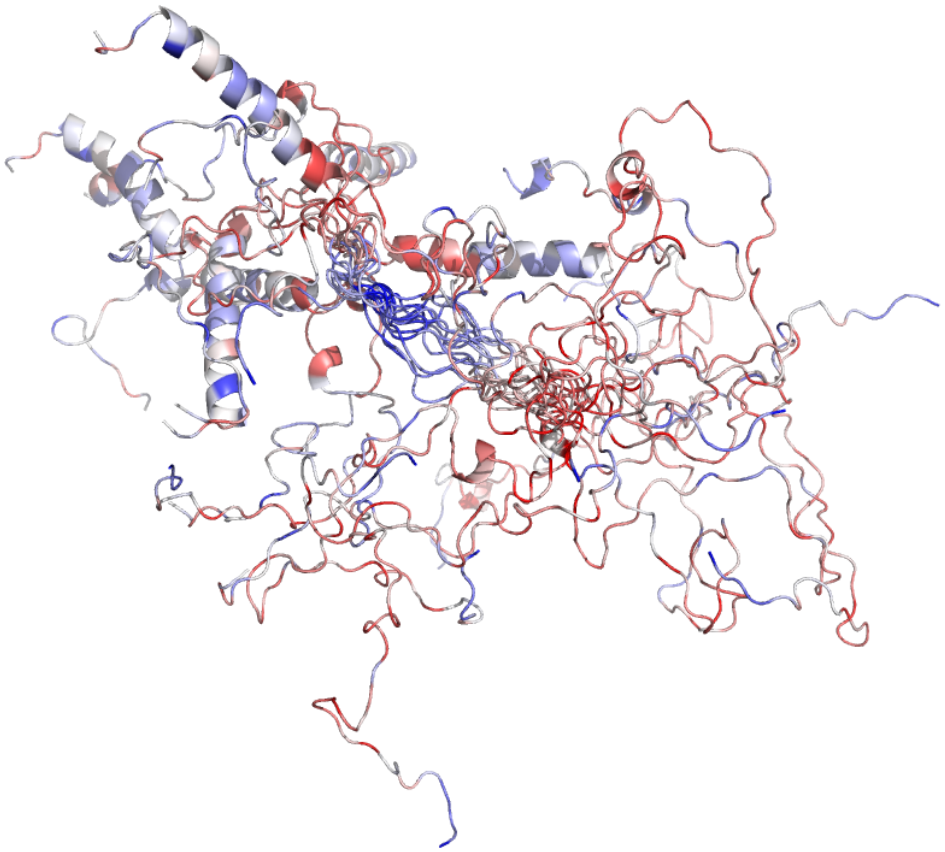
The radius of gyration is a measure of the complete structure and contains no information on single residues, so it was not useful for coloring. The variance of the residues distance matrix is biased towards the central residues, as the distance values for these positions are lower, and hence they vary less. Finally, the secondary structure per residues was a non numeric value and as a result we could not compute a variance metric. Maybe we could compute the entropy (as done in task 2), but we wanted to focus specifically on the variance metric so that we could correctly confront the various features.
The relative ASA is the only feature with a variance that is useful for the problem, because it both contains residue information and local bending information, in the sense that a structure that is more packed together will have a lower relative ASA. We used this metric for both the coloring and translation.
Given the variance of each residue, we then searched the 20 residues segment with lowest average variance, to be used in the nodes alignment.
In the pymol images the blue color will represent low feature variability, while red color will represent high variability.

As it can be seen in the following pymol images, there are two principal areas where the variance is low:
The first is the segment, around 5-25 residues, corresponding to the alpha helix. The second area, around 30-50 residues, correspond to a zone with no secondary structure. From our understanding, we could say that these sections of the protein are the most stable ones, i.e., with less disordered content.

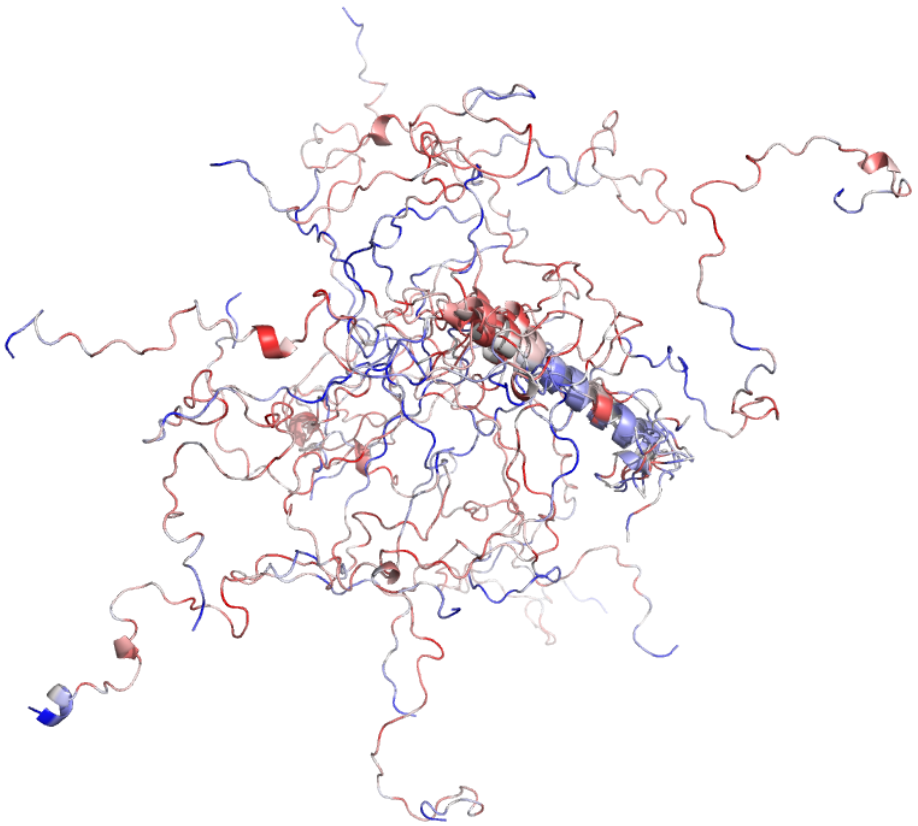
PED00153e007 pymol image



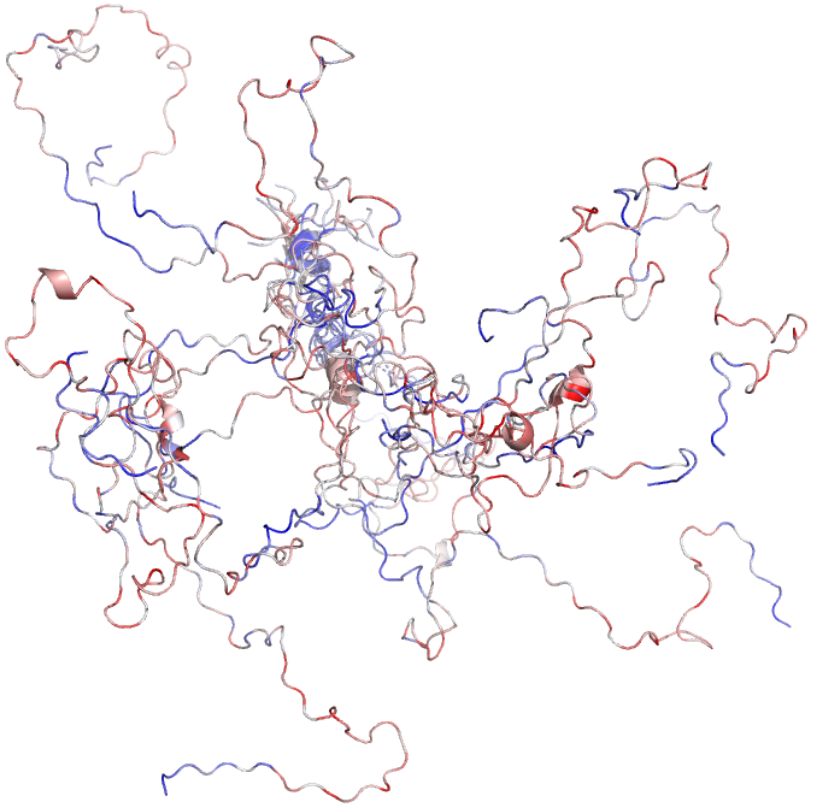
PED00153e008 pymol image



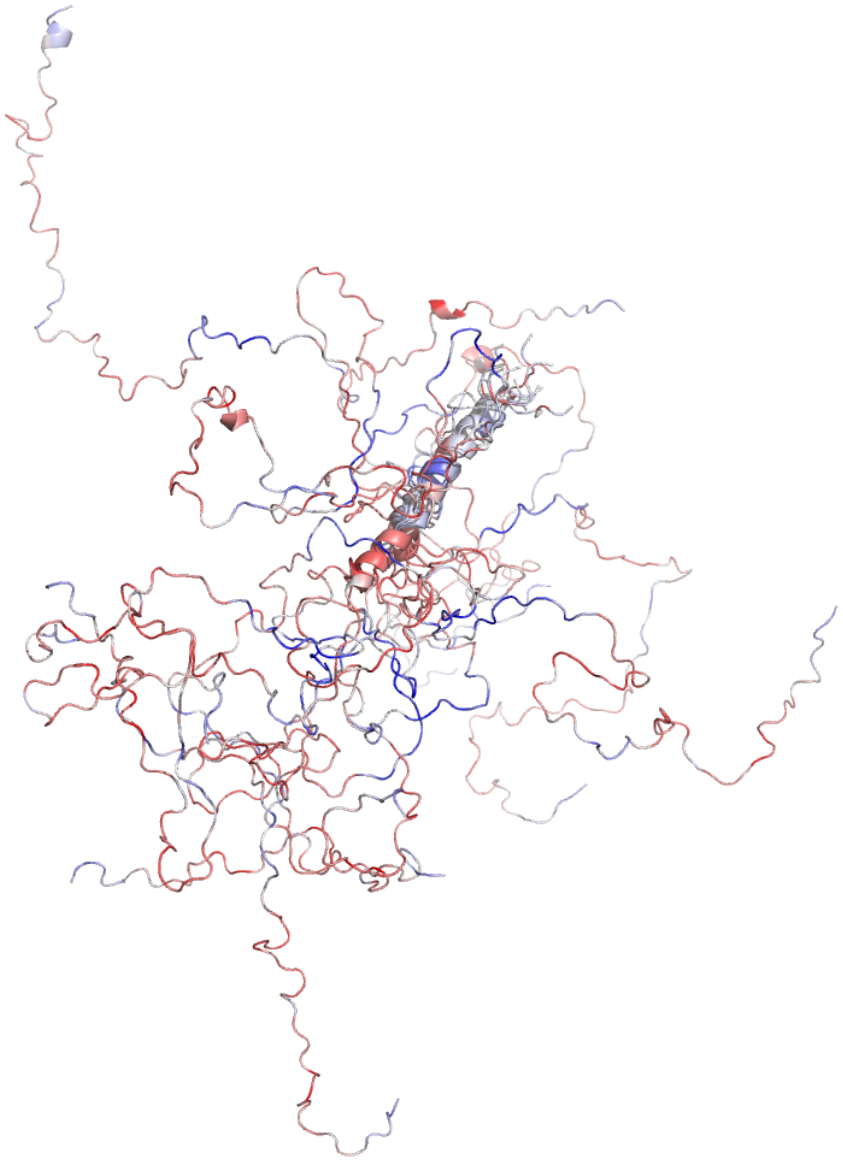
PED00153e009 pymol image



PED00153e010 pymol image



PED00153e011 pymol image



Task 2

Ensemble features

1. **Radius of gyration**
   In this case, we simply retrieve the already computed radius of gyration for each of the conformations in the ensemble.

2. **Secondary structure entropy**
   In order to measure the entropy of the secondary structure for each residues along ensemble conformations, the following formula has been used:

   $$H(X) = -\sum_{x_i \in \chi} x_i \, ln(x_i)$$

   where χ is the set of secondary structure classes, x_i is the frequencies of the i-class for the analyzed residue across all conformations.

3. **Median solvent accessibility**
   The median solvent accessibility for a generic residue *r* has been calculated by simply computing the median of the rASA value of each structure in position *r*.

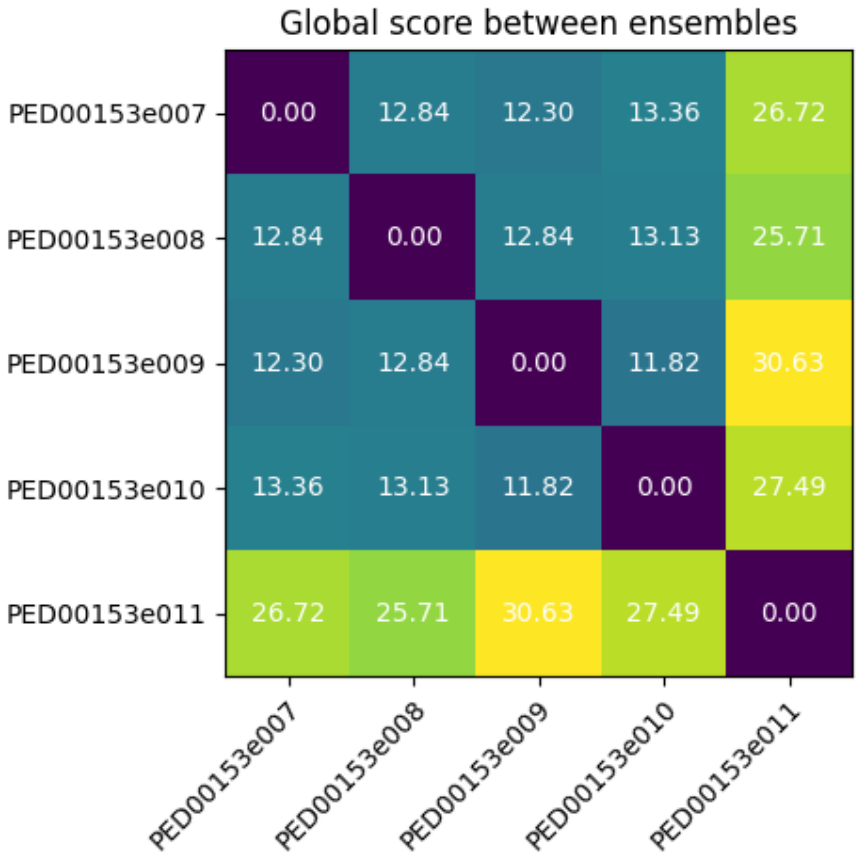
4. **Median RSMD**
   RSMD has been evaluated using equation (5) of Lazar et al., 2020. The only difference with the equation (5) is that the first ensemble conformation was used as reference to evaluate RSMD-value for the rest of conformations, instead to compute RMSD-value across all pairs of conformations.
   Looking at the local Scores Plots (task2 output), it is possible to notice that the comparisons of the ensemble PED00153e010 with all other ensembles show a 10-fold higher peak in the region between residues 16-21. Using pymol, it is possible to notice that there is a big jump between the coordinates of the 20 and 21 residues, which easily explains the peak.

5. **Median distance**
   The median distance of equivalent positions has been calculated by taking all the individual distance matrices and computing the median for each equivalent cell.

6. **Standard deviation of the distance**
   In the case of the standard deviation, we followed the same procedure but computing the standard deviation of each equivalent cell in the distance matrices.

Global score heatmap

To compute the global score between two ensembles, we have implemented the formula in (4) of Lazar et al., 2020 to obtain the *ens_dRMS*. This metric is based on the difference of the median distance of each pair of equivalent positions across ensemble conformations. As a result, the lower the global score, the more similar the ensembles are. In our case, we computed the global scores of the 5 ensembles of the MKK7 IDP and we plotted them in the form of a heatmap:



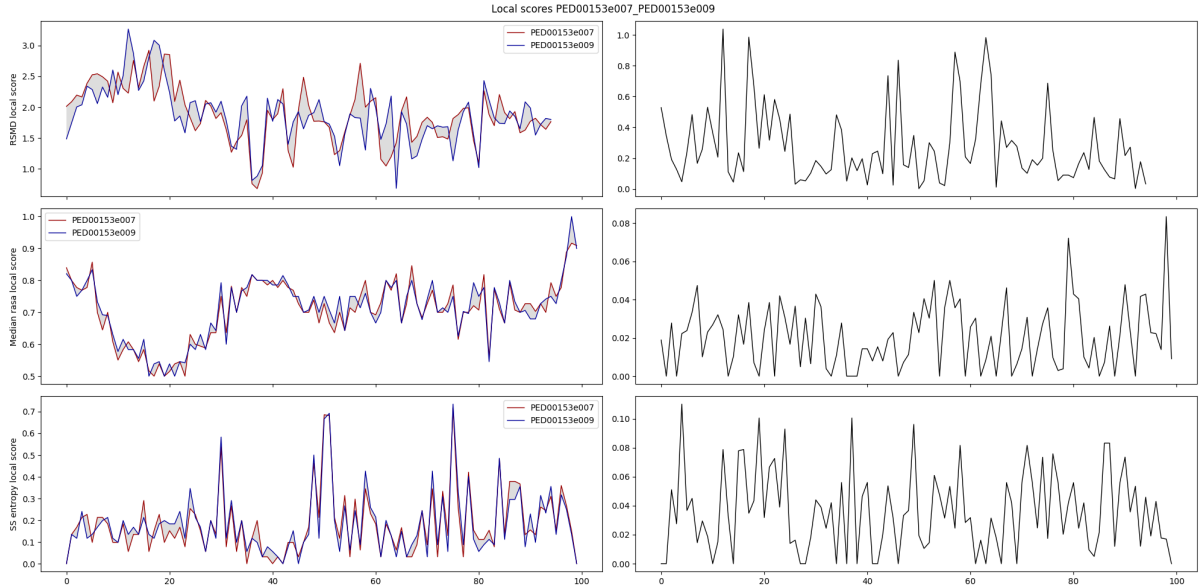
We can observe a pattern in this plot: ensemble 11 has a higher global score with all the other ensembles. This probably indicates that ensembles 7-10 contain structures whose alpha carbons stay within similar coordinates, as the ensemble median differences are more close and hence, the global score is similar.

Local score plot

To compare two ensembles locally we have identified three local scores vectors:

- Confrontation via median RSMD
- Confrontation via median rASA
- Confrontation via secondary structure entropy

In the following plot we display the local plot of the ensembles 007 and 008. On the left column we compare the two feature values along the residue, and on the right column we plot the difference between the two corresponding ensemble features.



The median RMSD seems to be the best one to compare different ensembles, as the other two metrics don't seem to capture significant differences.
As stated before comparaison with the ensemble 010 in case of RMSD produces a peak in the range of residues 16-21.