An Intuitive Approach to Simple Protein Structural Clustering by High Variance Regions

Automatic Clustering of CDK2 Kinase

Yi Yao Tan

Instructor: Thomas Gaillard

Ecole Polytechnique, Paris

Laboratoire de Biochimie

2021

Exploring Intuitive Approaches to Simple Protein Structural Clustering by High Variance Regions

Automatic Clustering of Human CDK2 Kinase

Yi Yao Tan

Introduction

A lot of different methods and algorithms have been devised to optimize protein clustering including the use of domains of topology and geometry to identify similar protein structures. The uses of this field includes that of notably structure based drug design and other structure based biological research, a wide range of automated clustering algorithms have been developed to facilitate this industry. To address this hierarchical clustering is used to group similar proteins. It relies on the input matrix a derived by a specific and adequate distance metric to identify the conformation.

While using the distance based RMSD provides data of structure to structure dissimilarity between different conformations, a simple RMSD matrix lacks to ability to describe the subsequence-wise and symmetrical-wise in shape and position which could be the main identifying elements for a protein's conformation and state. The make up for this loss we explore an intuitive method using the information obtained from the coordinates, in particular, the structurally high variance regions which give accurate information of structural differences of subsequences and use a map of each of them to be a feature in the clustering matrix. Human CDK2 protein was chosen as the protein of study since it is a highly researched protein, with practical applications of clustering in cancer research, crucial in the regulation of the cell-cycle, and has a sizeable amount of experimental data collected on the confirmation structures.

In the end we explore the pros and cons of the methods applied on the high variance regions. Despite the randomness factor by the T-SNE used in mapping the coordinates to lower dimensions, the coordinate based approach consistently performed better than the RSMD in clustering the three groups of the CDK2 protein kinase.

Contents

Method

1.1 Data Annotations

The data list of samples for the human CDK2 protein are first retrieved from Uniprot, code P24941 (LINK UNIPROT) and with a basic shell script a total of 420 pdb files entries of different pdb files are curled from rscb.org. Then using a python script the pdb files are read and annotated storing the name of the entry, chains, the resolution, presence of a cyclin structure for the close and openness of the the individual chains of the conformations, and the phosphorylation of threonine (TPO) at position 160 of the chain to determine the activity. In particular, the presence of cyclin means that the chain is open and otherwise closed and the phosphorylation of threonine means the conformation of active and without, inactive. The scripts then store the annotated data chain by chain then stored into a dictionary with the keys as the name of the chains and values as objects with each of the properties stored within. A total of 531 annotated chains are registered. (EXPLAIN MORE AND CITE CDK2 PAPER)

1.2 Conformation Alignment, RSMD and Coordinate Data

A script is run to pick the chain with the highest resolution and the presence of the natural ligand: ATP in the chain. This choice is then used to align the rest of the chains onto since the highest resolution will give the most accuracy in alignment and since other synthetic ligands, if any, in the other conformations are made to imitate the natural ligand thus making it a good candidate for the rest of the samples to align to. In the case of the experiment 4EOJ chain A was found as the best choice to align the rest of the conformations to, with a resolution of 1.65 angstroms and the natural ligand of ATP.

Conformation alignment uses the software pymol built in function align and iterates through all of the loaded chains and aligns the structures with structural outlier rejection cycles set to the default: 5, rejection cut off default of 2.0 and target = $4EOJ\_A$. This alignment also transforms the structure coordinates to be placed on top of $4EOJ\_A$. All of this is done smoothly since the samples all have the same sequence.

Finally in the same order that the chains are loaded and aligned, a 531 by 531 RMSD matrix is calculated in the same order both in columns and rows the room mean squared derivation of the carbon alpha atoms of residues present in each pair of structures given by:

$$Matrix_{i,j} = \sqrt{\frac{\Sigma_{k=1}^{N}(x_{i,k} - x_{j,k})^2 + (y_{i,k} - y_{j,k})^2 + (z_{i,k} - z_{j,k})^2}{N}}$$

Where $x_{i,k}$ is the x coordinate in the i-th chain of the k-th residue (out of the total residues present in both chains) and N is the total residues present in both chains. Very clearly, the diagonals of the matrix are all 0s since there is no dissimilarity between a conformation and itself. Lastly, coordinate matrices of each conformation is also obtained by pymol's getcoords feature, using multiprocessing for optimization.

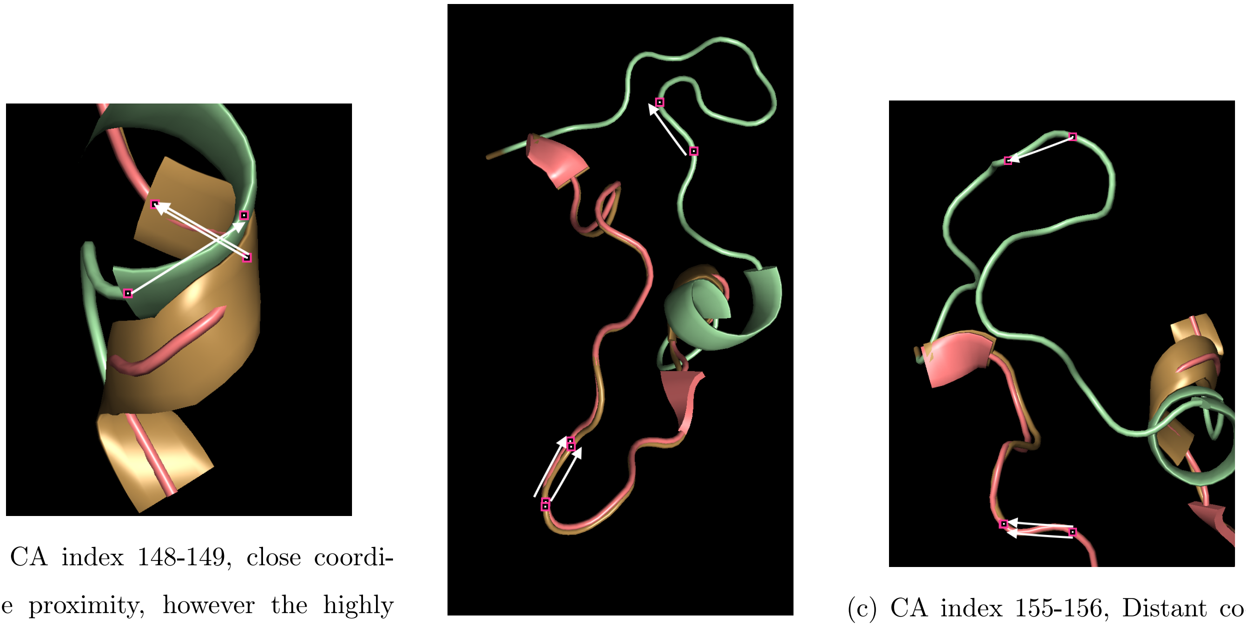
1.3 T-SNE based map clustering technique

1.3.1 Finding high variance regions

The main idea of the algorithm is to find the high variance segments, in terms of the structure, of the sequence and use their coordinates to map to a single value to use as a feature for the clustering algorithms. The method doesn't use a distance matrix and keeps the relative position instead of each segment via a map; this allows for clearer distinction between clusters. For n distinct high variance regions chosen to identify the conformation, there'll be a $number\ of\ samples \times n$ sized matrix.

To define a useful definition of variance of the structure, just by using the variance of coordinates is insufficient to give properties like shape and structure. To define structure, vectors from the coordinate of the carbon alpha from one residue are drawn to the next. Theoretically, unaltered bonds angles and lengths between the residues of corresponding indices of different are parallel and of the same length, thus vector subtraction will give a result near 0; therefore, by using the variance in the defined vectors a variance in shape. Of course the change in a previous bond angle could affect the position of the next, however, in agglomeration the goal of finding a high variance segment is achieved. More specifically the variance of the $residue_k$ to $residue_{k+1}$ vector is defined by the average of the variance of the x, y, and z coordinates.

For example the close up of the activation loop of: 1HCK_A —Open Active, 1QMZ_A in Sand —Open active, 4EOJ_A in Deep Salmon —Closed_inactive :



(a) CA index 148-149, close coordinate proximity, however the highly dissimilar vectors.

(b) CA index 155-156, Distant coordinate proximity highly dissimilar vectors

(c) CA index 155-156, Distant coordinate proximity similar vectors

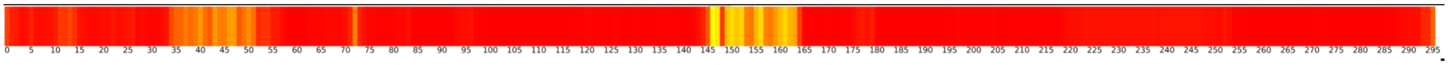


Figure 1.2: Heat map of average x, y, z component variance of residue to next residue. Index 0 is vector from residue 0 to residue 1 (0 indexed). Red = low variance, Yellow = high variance.

To gain proper information of the choice of high variance regions, a function is used to obtain all of the indices with above 90-th percentile variances, then the indices were grouped into subsequences which were ranked by their average variance value from highest to lowest in order were: vectors indices 144 to 164 the activation loop with an average of 2.161 (units), 34 to 54 around the PSTAIRE helix average 1.467 (units), 71 to 73 with 0.886(units), index 0 0.841 (units), 11 to 14 with 0.769 (units) near the glycine rich loop, 293 to 294 with 0.532 (units).

1.3.2 Mapping coordinate to usable feature values

Let m be the number of samples, n be the number of featured segments and, l_i be the length of the i-th feature segment. First to make the matrix with missing residue coordinates all usable, in the case where conformations with missing residues need to be classified, sklearn IterativeImputer is applied upon a matrix of the arrangement of same index from different chains, size of $m \times l_i$, in this case since if a residue is none all of its coordinates are not available, linear regression is not possible to help fill in the row of the matrix, thus IterativeImputer actually imputes using the average of x,y and z of the coordinates of the specific index of the chain over all of the samples. A map is needed to send a matrix $k \times 3$ of coordinates for the segment length of m so that the the values can be used as features to represent the conformation in clustering. Since in this region the coordinates vary a lot of and the shape too, it makes sense to find the average of the coordinates and perhaps similar segment structures would cluster together having similar arrangements of coordinates.

$$map : \mathbf{R}_{k,3} \to \mathbf{R}^3$$

$$A_{k,3} \to V := \frac{\Sigma_{j=1}^{k} A_j}{m}$$

Where A_j is the coordinate of the j-th index of the segment of the length k, this results in a resulting $m \times 3$ matrix where m is the number of samples and each sample has a xyz average coordinate value. Now using the fact that similar structures are already clustered in three dimensions, to to maintain cluster distances on a projection to one dimensional vector of values for all of the structures for the specific segment, t-Distributed Stochastic Neighbor Embedding or T-SNE from the Sci-kit-learn tool kit is used to project the distributions to lower dimensions while preserving inter-clusteral distances giving an $m \times 1$ vector.

T-SNE constructs probability distributions centered around each point given standard deviation derived by user input perplexity. The similarity between the reference point and another is a conditional probability value derived from the probability density beneath a Gaussian distribution centered at point of reference. Once the probabilities are defined they're projected into a lower dimension under with a student-T distribution while minimizing the Kullback-Leiber divergences between the probabilities on each dimension which penalizes on further distances between points therefore prioritizing local structures making T-SNE a natural choice. For the sake of creating more distinct clusters and facilitate hierarchical clustering, the value of perplexity used is 50 the highest value that the paper recommended, however, different perplexity values can be chosen depending on user. (CITE T-SNE PAPER) Finally random initialization is set to PCA for consistency in results. The T-SNE function first projects data using PCA with a randomized svd_solver and then runs gradient descent of the KL divergence on the PCA projection to create the final embedding matrix. (CITE SKLEARN)

When plotted on the final matrix of features conformations with similar segments will cluster together on each dimension, and it can be used to describe well the physical similarities of each structure relative to another via the axes of featured segment. The value vector from each segment can be put together in finally creating an $m \times n$ matrix where m is the number of samples and n is the number of featured segments. Now the $m \times n$ matrix is ready for clustering. **Due to the random initialization of T-SNE, to obtain the final results, one can run multiple trials and find the average. To find the final prediction of a conformation, one can use the most frequent prediction of the conformation across the multiple trials.**

1.4 Clustering

The distance matrices are then put through scipy's hierarchy clustering using ward's algorithm which minimizes the variance in the cluster after merging. (CITE SCIPY) For the purposes of the experiment the dendrogram is automatically cut into 3 using cophenetic distance by sklearn.cluster's AgglomerativeClustering. The result is then plotted using T-SNE projections and analyzed with multi-label statistics.

1.5 Statistics

After the tree cut, a function tests the permutation of the three clusters to fit the three annotated group, the permutation with the most amount of correct predictions is used to access the F1, precision, and recall score using multi-label evaluation metrics.

For the T-SNE based map clustering technique, 20 trials are used, and then the average is found of the 20 trials of each of the F1, Precision, and Recall for each of the groups metrics and then divided by the number of iterations and similarly done for the the cross-classification tables.

In the end the RMSD clustering, T-SNE Based Map clustering, and an additional of dihedral based clustering based on the paper (CITE DIHEDRAL BASED CLUSTERING PAPER HERE) are compared.

Results

The results of only conformations with no missing residues for the full sequence of each of the methods are not shown since only 55 out of 531 have no missing residues. This sample size under represents the total data set. RMSD had a resulting F1 score of 0.396 on the set of 55 conformations while T-SNE based map clustering had an average F1 score 0.445 on 20 runs.

2.1 RSMD results

2.1.1 Complete RMSD CA

The first experiment was done on the complete RMSD matrix of all the conformations and of the whole sequence.

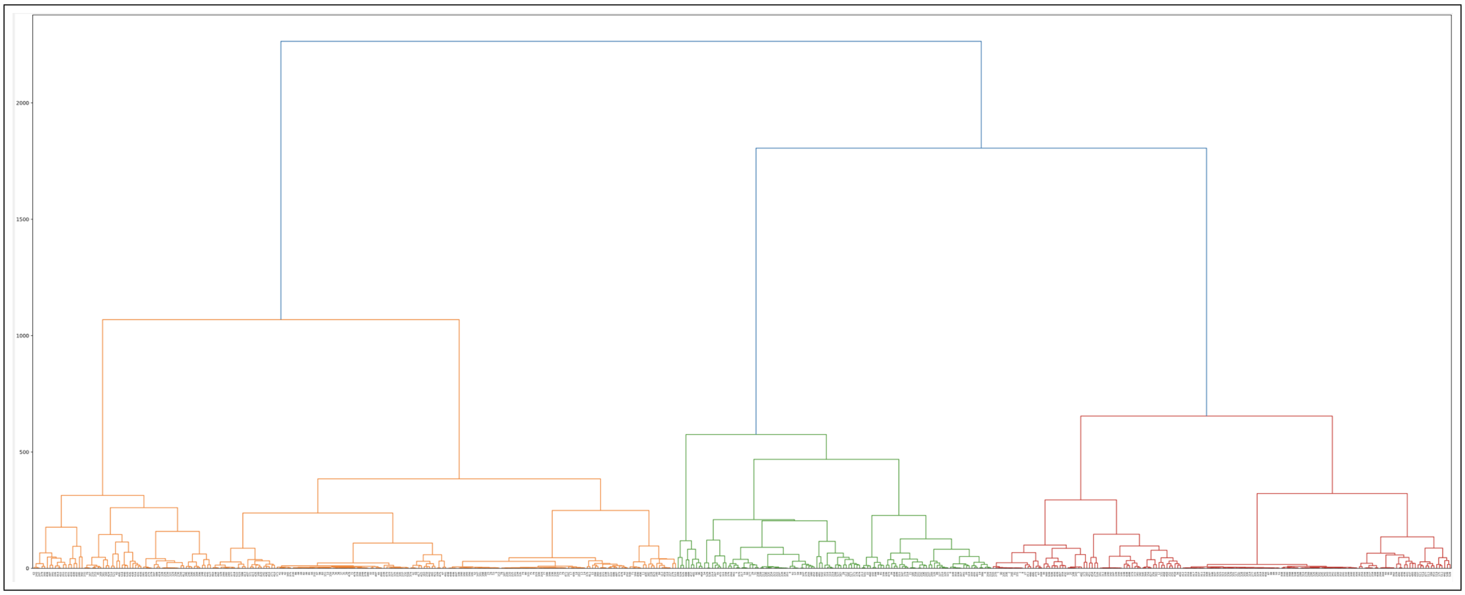


Figure 2.1: dendrogram of the results by ward's algorithm of RSMD CA matrix complete with all the conformations

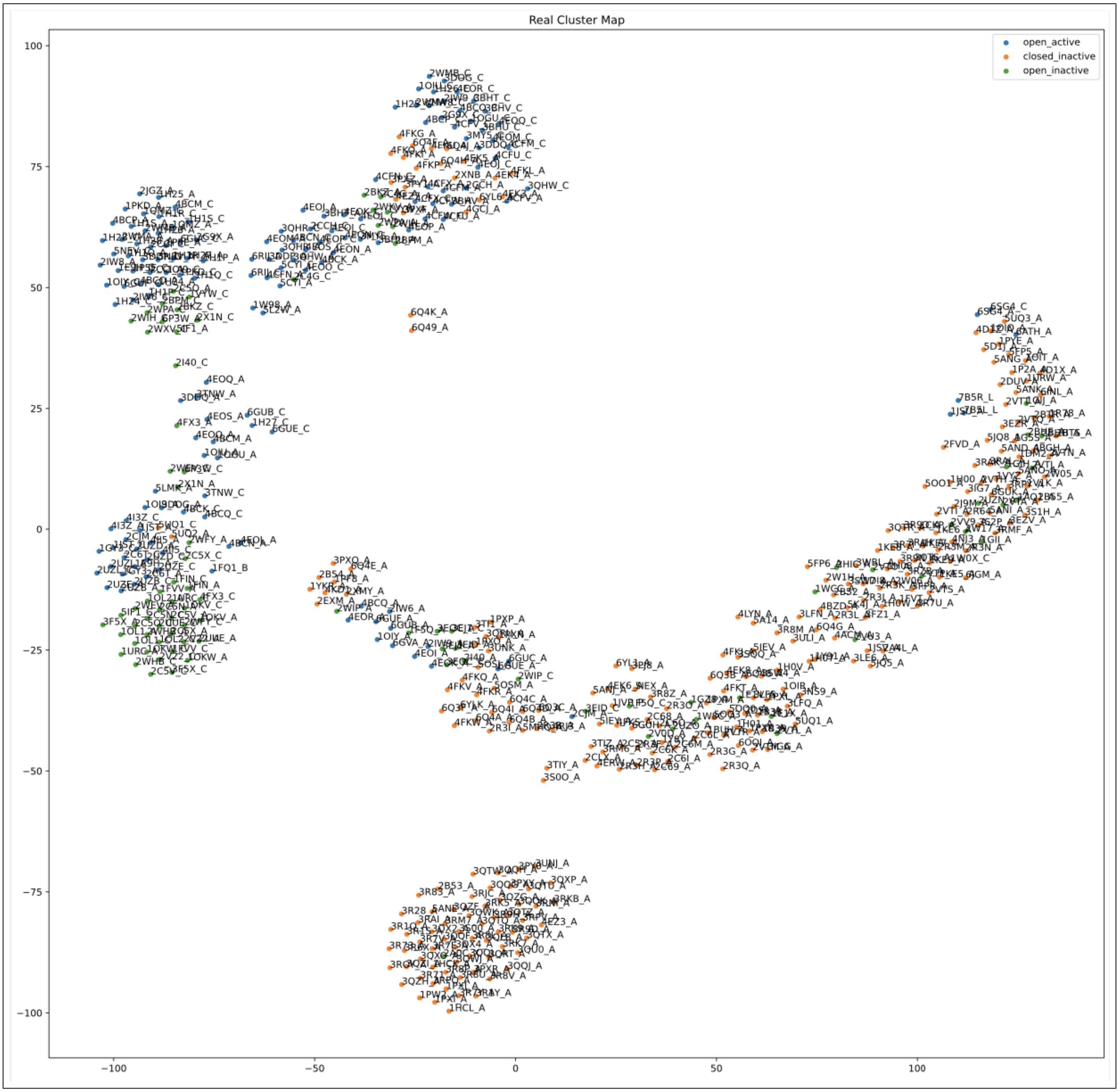


Figure 2.2: T-SNE projection map of the real clusters. It is obvious to see that clusters are not exactly distinct, especially open inactive and open active are well mixed. Especially in the top left all of the groups are evenly mixed. The majority of the group closed inactive on the bottom right is well clustered, but differentiating three distinct clusters isn't very obvious.

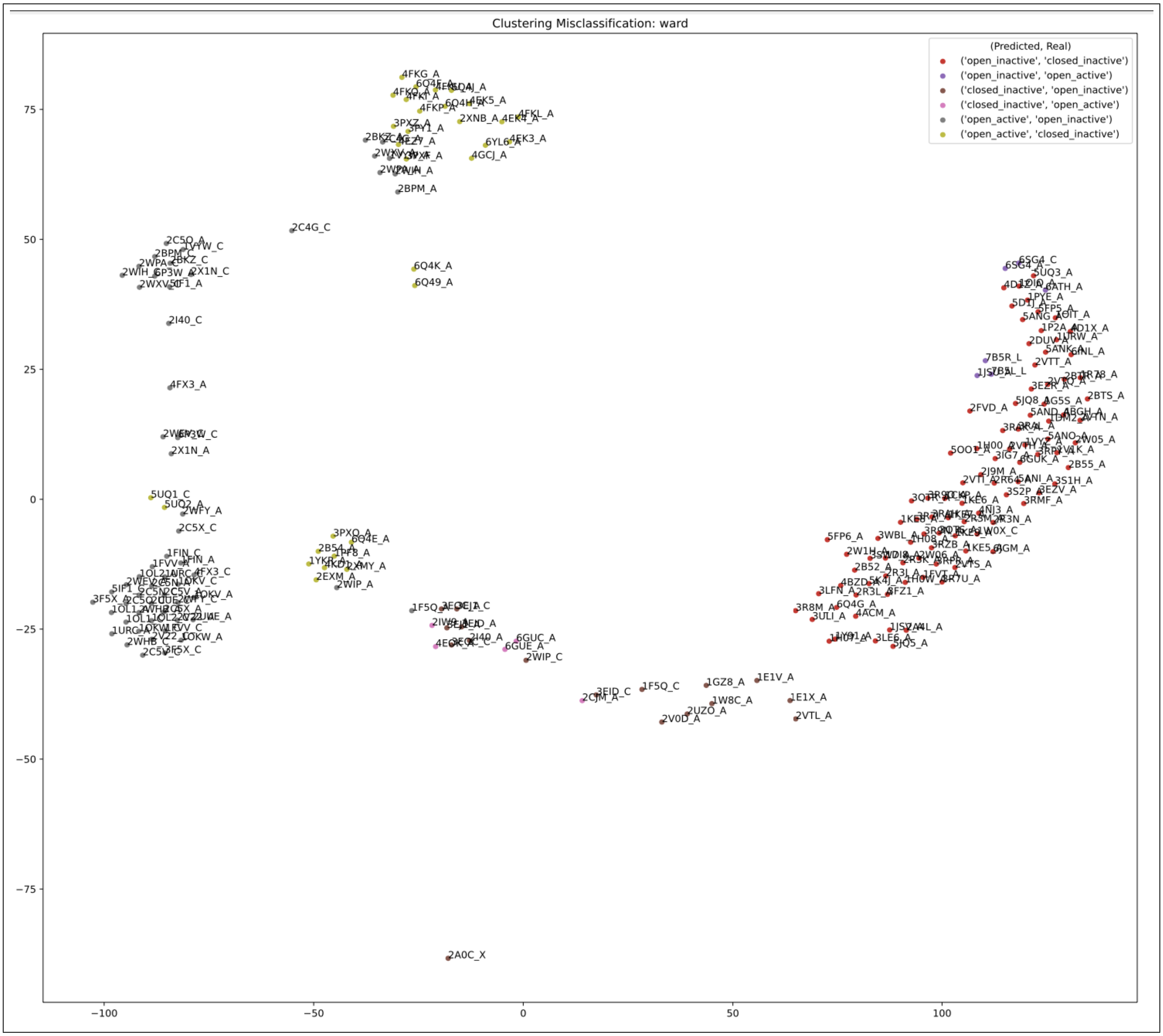


Figure 2.3: Misclassification Clusters map: The largest misclassification cluster can be seen as open inactive is classified as closed inactive up to an extent of the majority are misclassified (on the right side of the map). Beyond that There is another large cluster of misclassified open inactive conformations as open active.

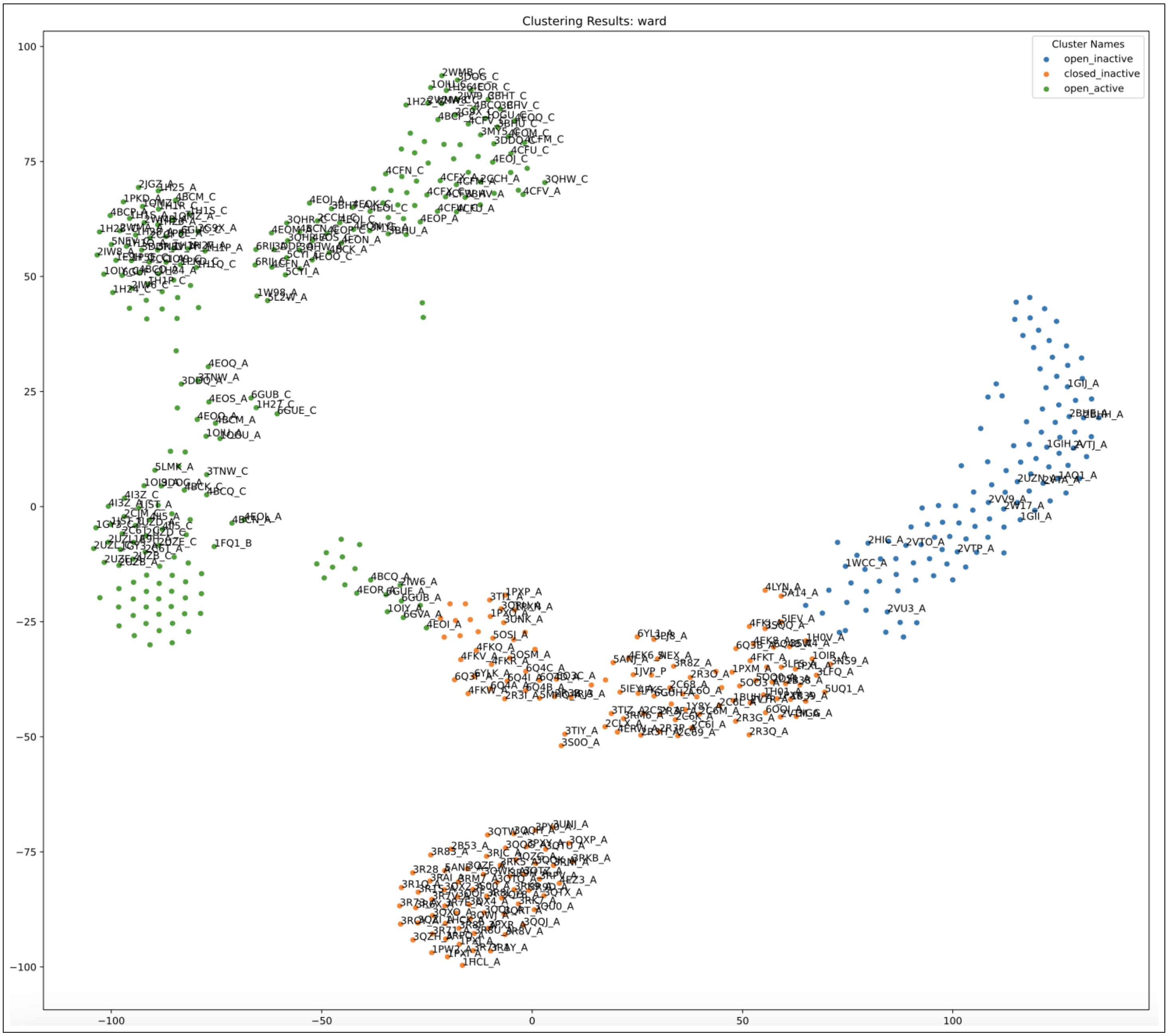


Figure 2.4: Clustering Map: the correctly classified are labeled and incorrect, unlabeled.

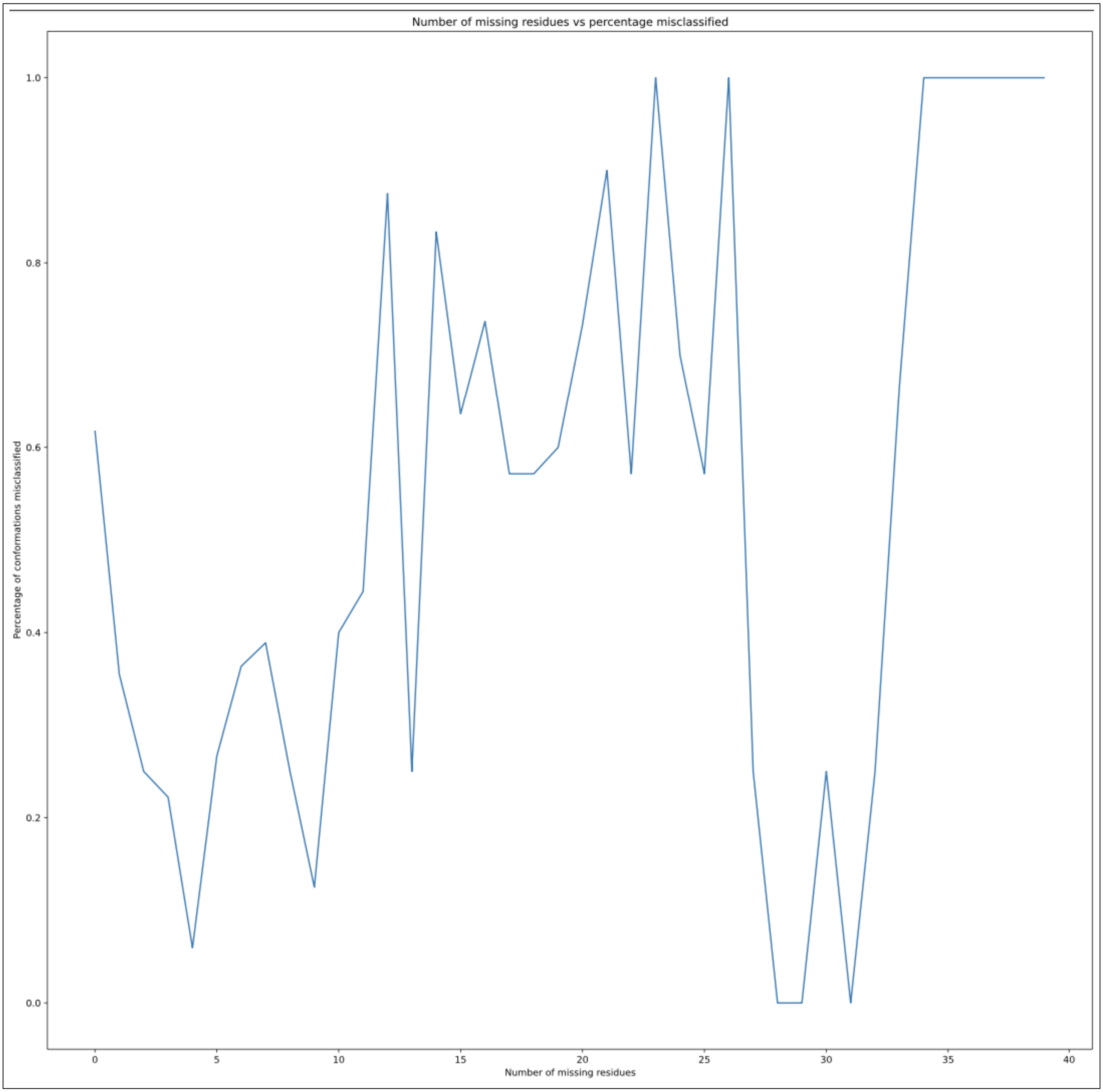


Figure 2.5: Missing residues vs Percentage misclassified: It can be seen that with RMSD matrix of the CA data, there is not much of a correlation between the missing data and the misclassification, thus making it hard to identify the the missing residues a factor of misclassification in incomplete data. A surprising 60% of the conformations with 0 missing residues were misclassified. Either the conformations with missing data skew the projection or rmsd is not a good metric in this case.

Statistics:

	Precision	Recall	F1 score
open inactive	0.135593	0.173913	0.152381
closed inactive	0.872093	0.541516	0.668151
open active	0.626556	0.932099	0.749380
Averages	0.544747	0.549176	0.523304

Table 2.1: RMSD no threshold of missing residues, full sequence, Cross Classification Statistics

	Real open inactive	Real closed inactive	Real open active
Predicted open inactive	16	96	6
Predicted closed inactive	17	150	5
Predicted open active	59	31	151

Analysis:

Open inactive having the some of the qualities of the two other groups is mostly misclassified as the real open active, while closed inactive is classified often as open inactive. Indeed the RMSD does a much better of separating the polar differences between open active and closed inactive. Perhaps if the clustering requirements only asked for two different groups RMSD would perform much better, however having one group with mixed qualities, RMSD failed to distinguish the differences. In general the full RMSD doesn't give very distinct patterns and clusters.

2.1.2 RMSD of CA of Activation Loop, No Missing Residues

For the sake of comparison with the other methods upon the same segment, RMSD is calculated for the specific segment of all the conformations without any missing residues.

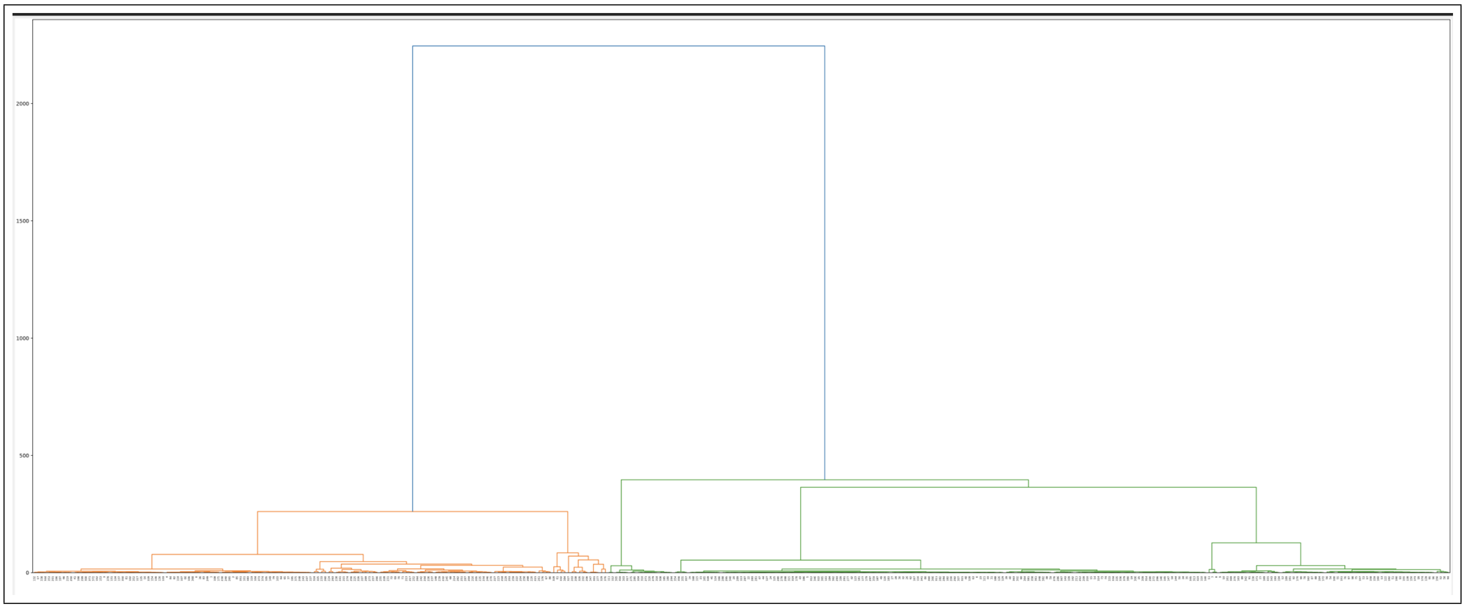


Figure 2.6: Dendrogram of the results by ward's algorithm of RMSD of CA in activation loop no missing residues. It can be seen that the automatic cut gives an orange tree and two sides of the green tree although two layers down on the right side, the cophenetic distance to connect the two sub clusters is sizeable. Leading to the misclassification of open inactive as open active as seen below.

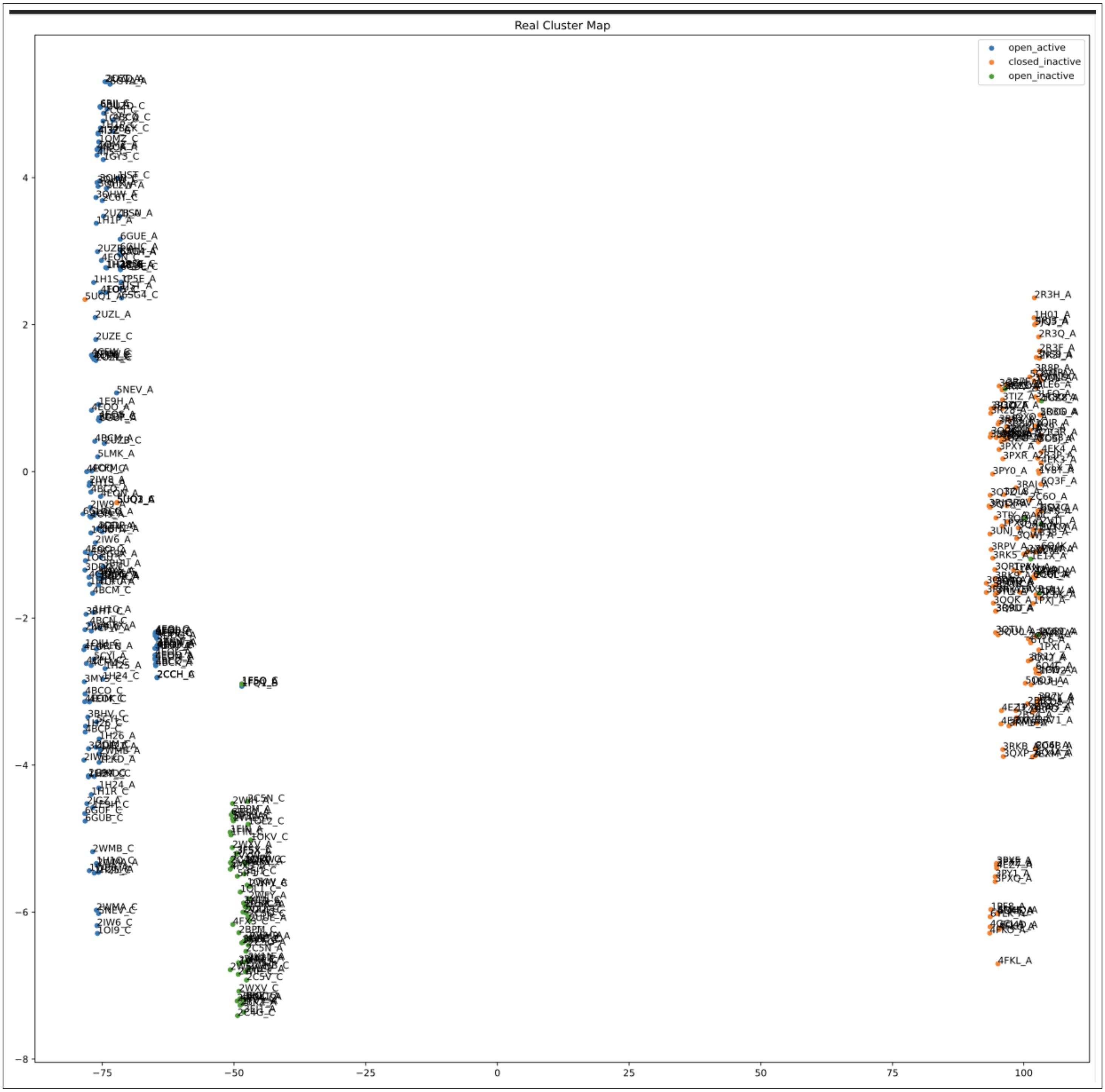


Figure 2.7: Real Cluster T-SNE projected map of RMSD of the activation loop, conformations without missing residues. Clear distinct clusters can be found on the map

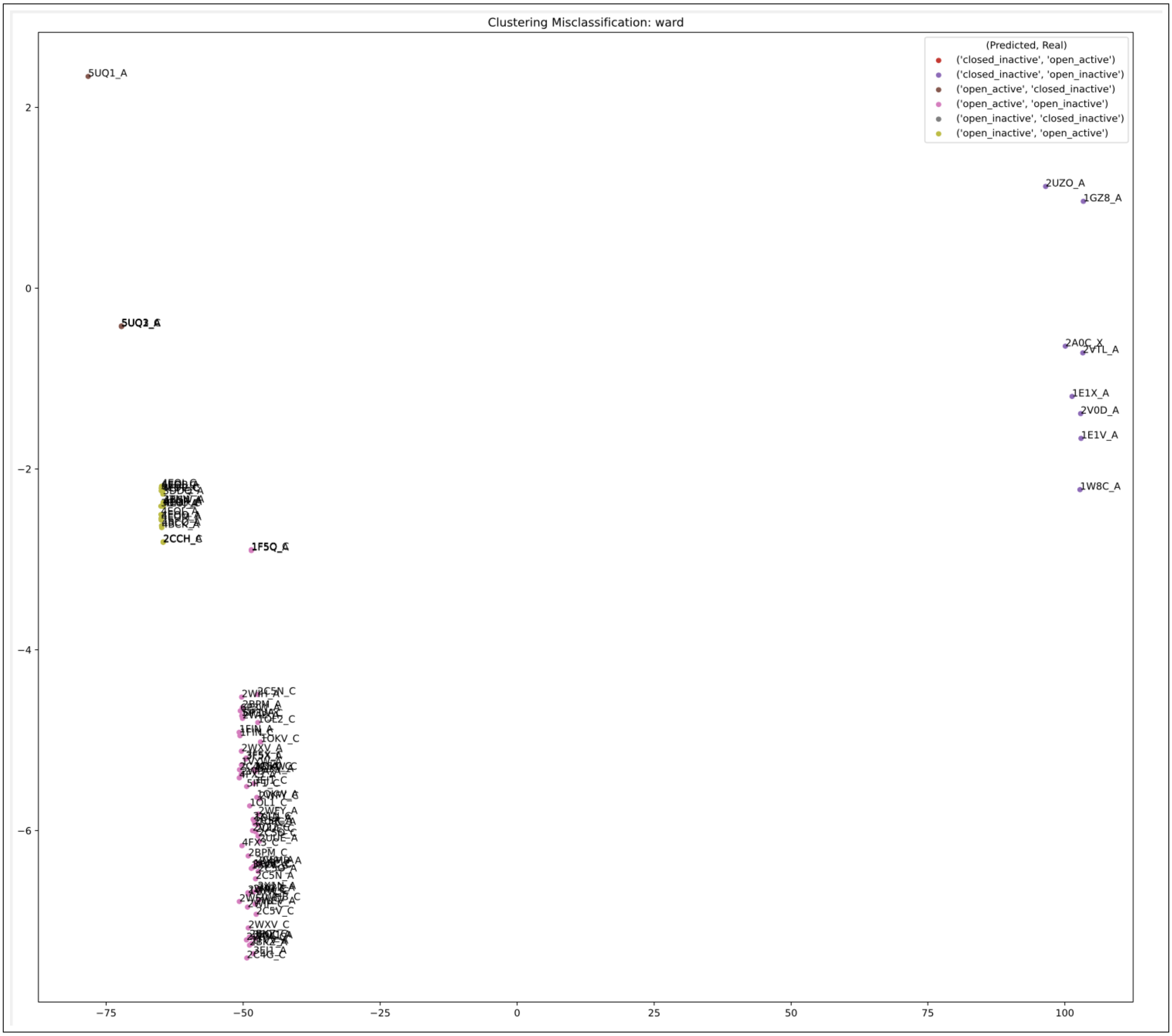


Figure 2.8: Misclassification Map of RMSD CA activation loop no missing residues. Due to the proximity and the large intra-cluster distance of the open active group, the open inactive is completely misclassified as open active. Perhaps a change of clustering distance metric might work better.

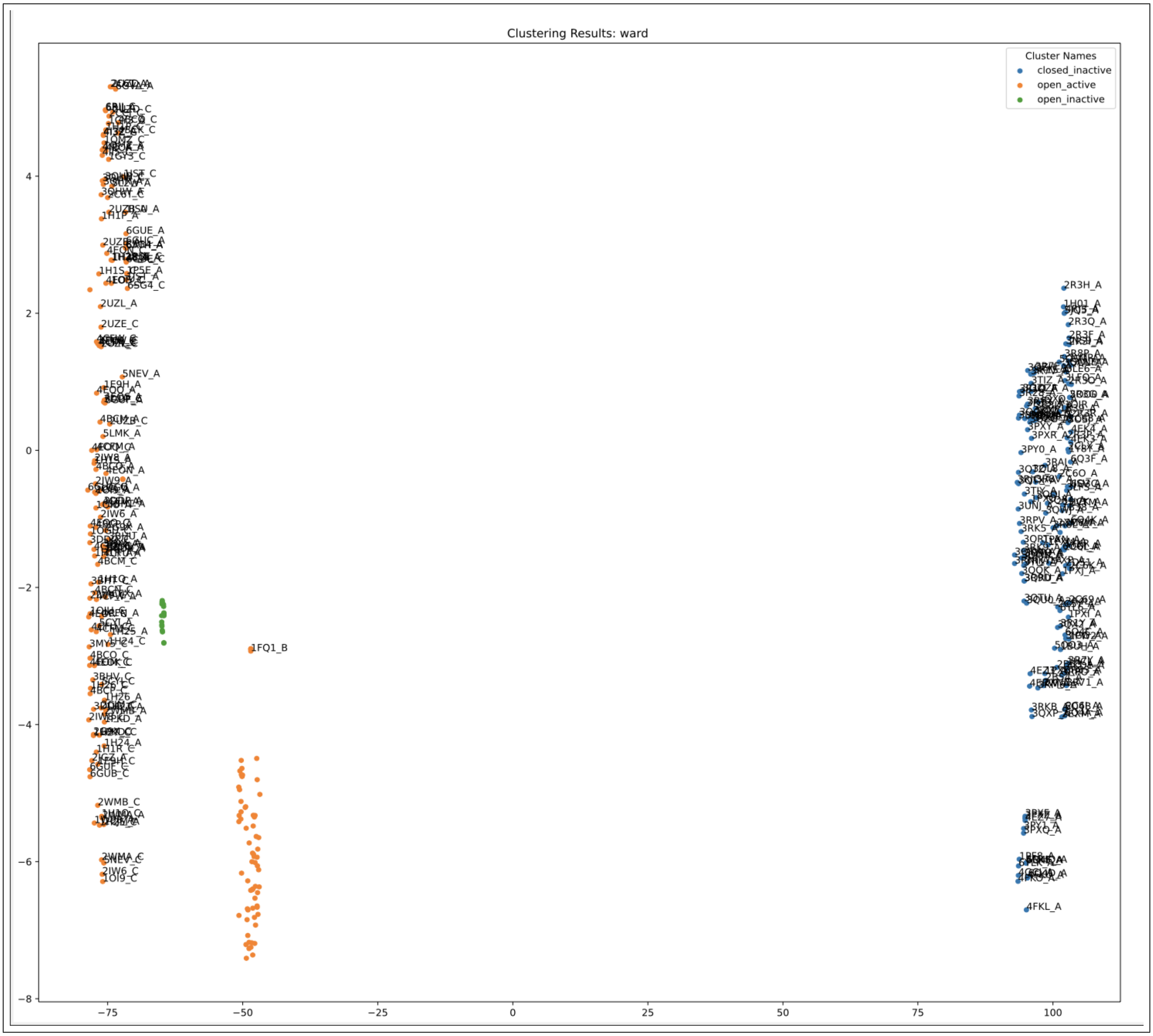


Figure 2.9: Clustering map of the ward's algorithm. A surprising small enclave of open active is misclassified as open inactive

Statistics:

Table 2.2: RMSD Activation Loop with no missing residues evaluation metrics

	Precision	Recall	F1 score
closed inactive	0.948718	0.973684	0.961039
open active	0.672986	0.887500	0.765499
open inactive	0.000000	0.000000	NaN
Averages	0.540568	0.620395	NaN

Table 2.3: RMSD Activation Loop with no missing residues cross-classification table

	Real closed inactive	Real open active	Real open inactive
Predicted closed inactive	148	0	8
Predicted open active	4	142	65
Predicted open inactive	0	18	0

Analysis:

Although the clusters were well defined, due to large intra-cluster distances of open active, open inactive was completely misclassified. Potentially using a kernel projection, different distance metrics, different dendrogram cut, or a different clustering algorithm might work better. Upon further inspection, 5UQ1 A and C, an outlier looks the same as open active on pymol just without the presence of TPO, despite being annotated as closed inactive. The small enclave of conformations mislabelled as open inactive are those which have very similar structures to 4EOJ A (open active) including 4EOJ A itself. Despite the misclassification, the method produced clear clusters which shows that high variance regions were correctly identified and correctly identify each conformation too.

2.2 T-SNE Based Map Clustering results

The plots below describe one trial of the T-SNE Map Clustering with specific parameters. Statistics are of the 20 iterations since T-SNE is based off of random initialization.

2.2.1 Complete T-SNE Based Map based clustering CA

Now we test the ability of the T-SNE map based clustering against that of the full RMSD without the high variance method and imputations for the missing values. Also since both of the plots are one dimensional, y-axis is imputed with 1s and labels are removed due to overcrowding so that the clusters can be visualized.

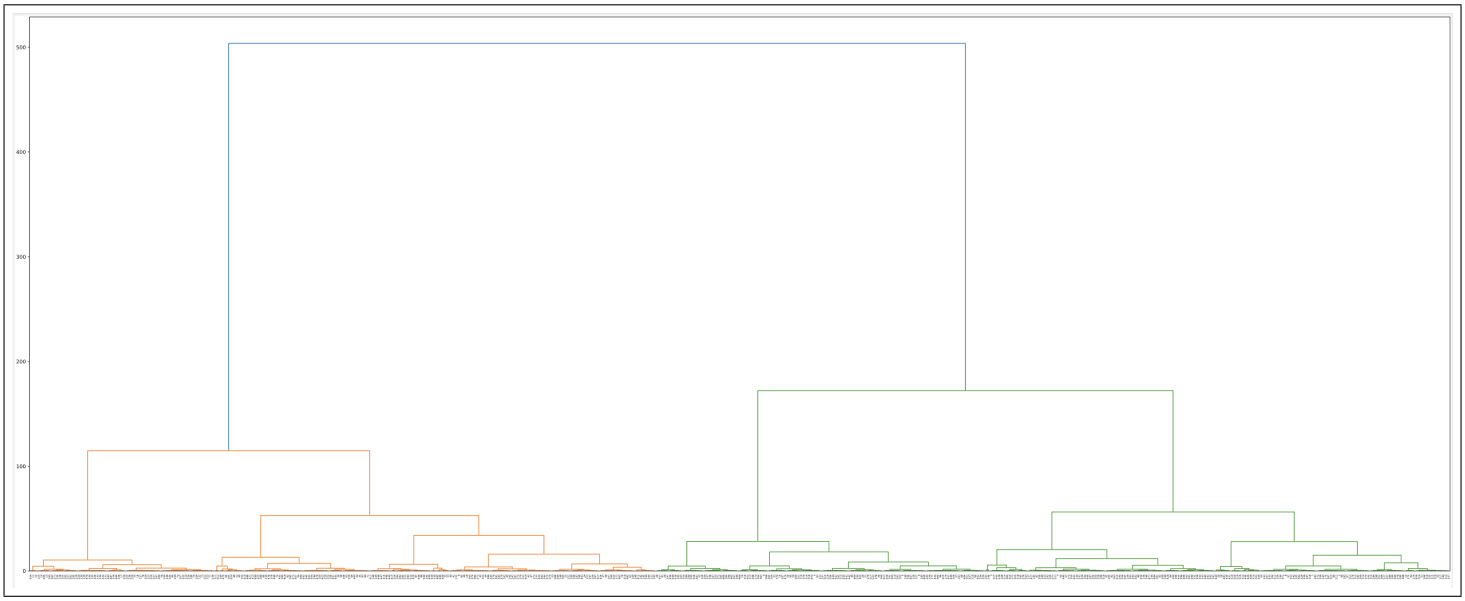


Figure 2.10: dendrogram of the T-SNE based map clustering of Activation loop, No missing residues

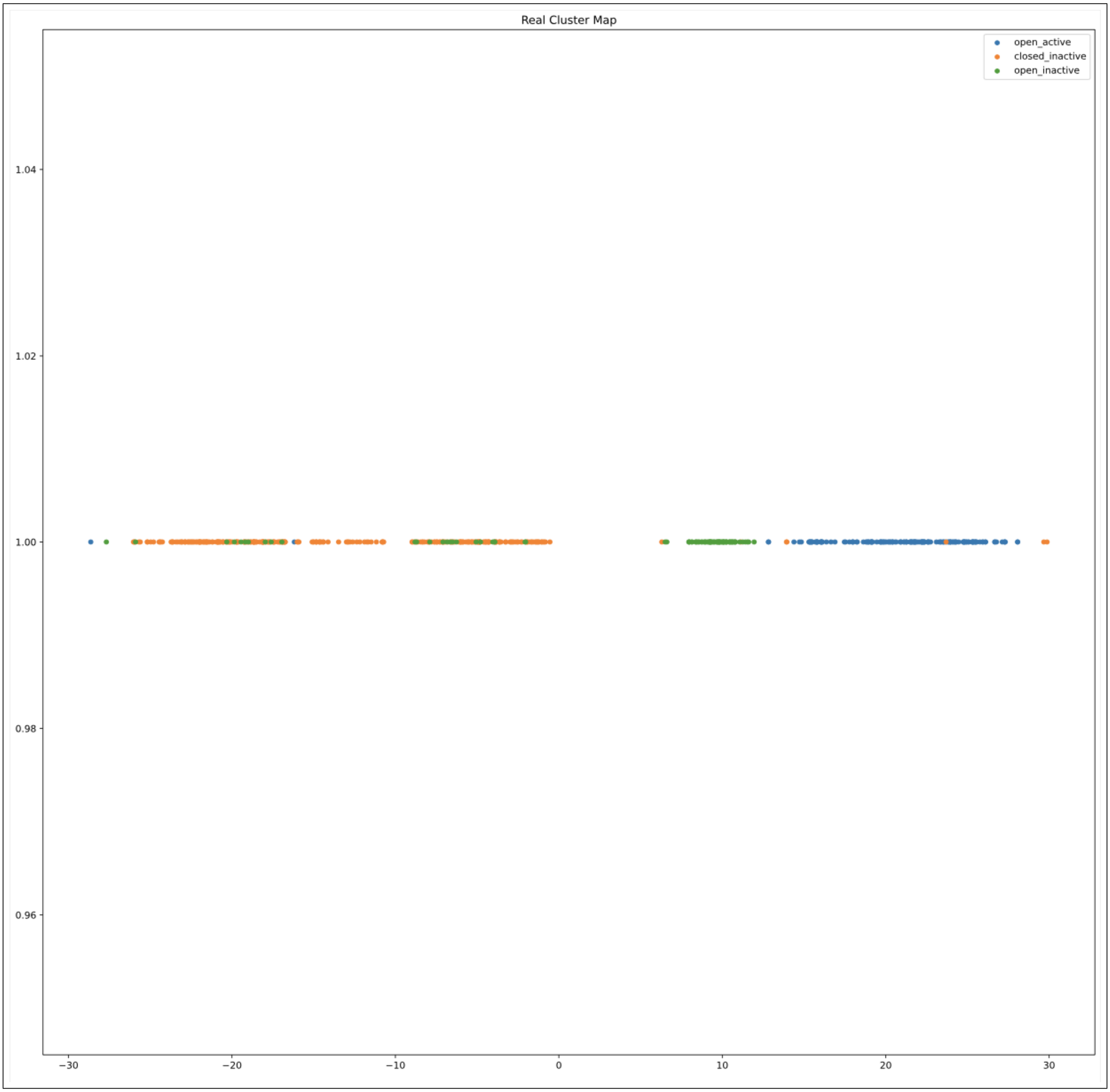


Figure 2.11: Real Cluster T-SNE projected map of T-SNE based clustering of the Activation Loop, No missing residues

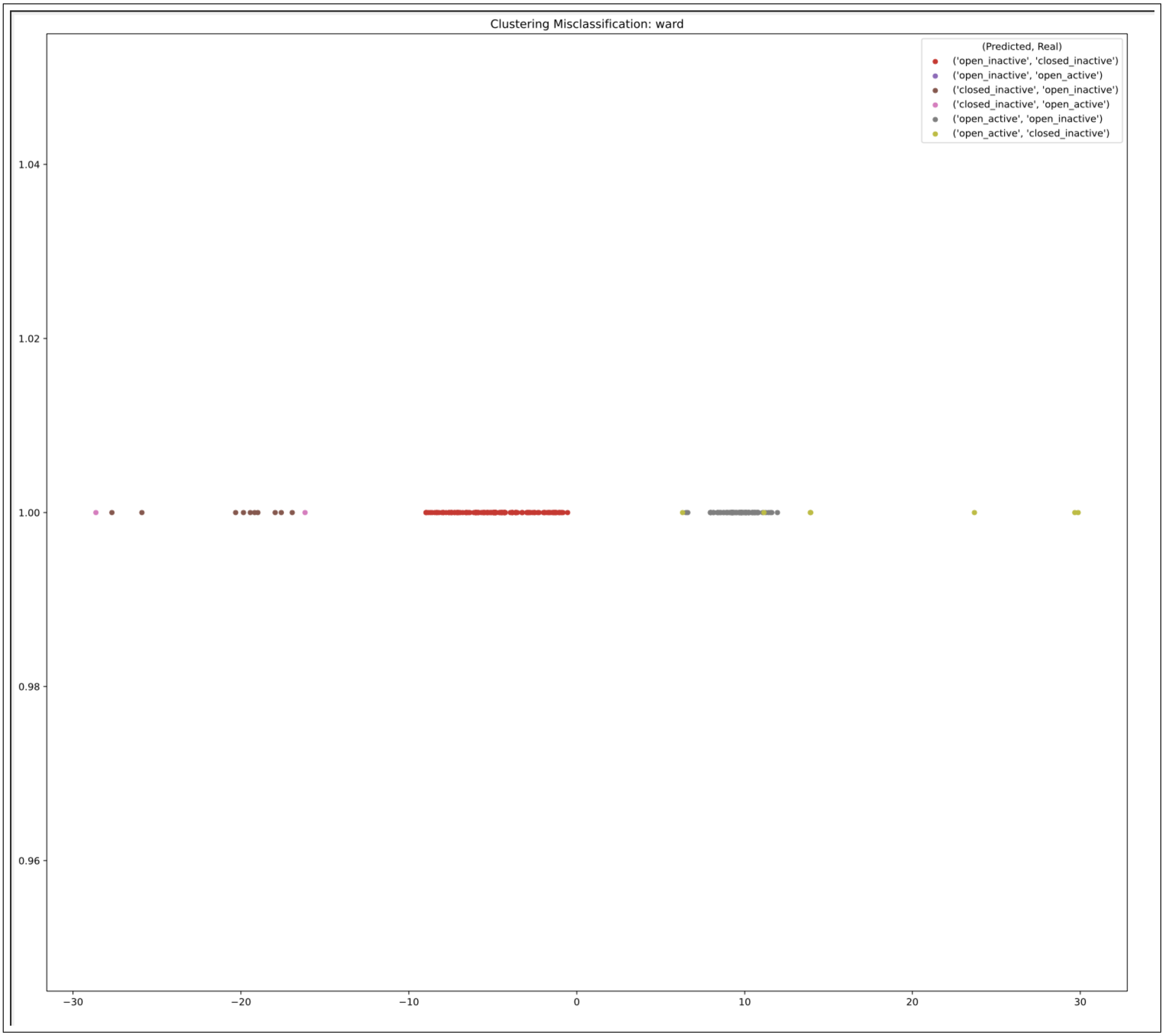


Figure 2.12: Misclassification of T-SNE Based map of full sequence:

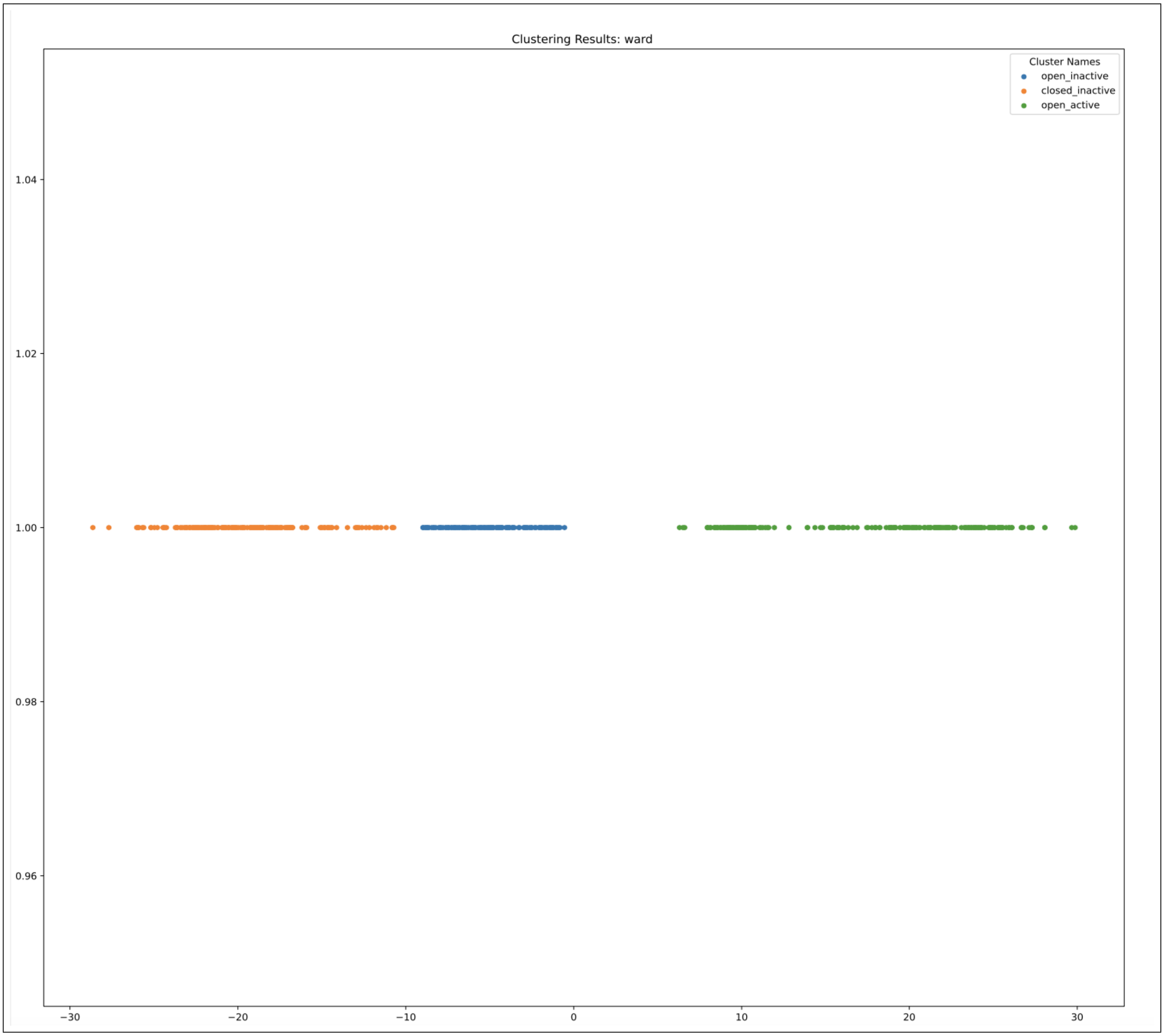


Figure 2.13: Clustering Map of T-SNE Based map clustering of full sequence using ward's algorithm:

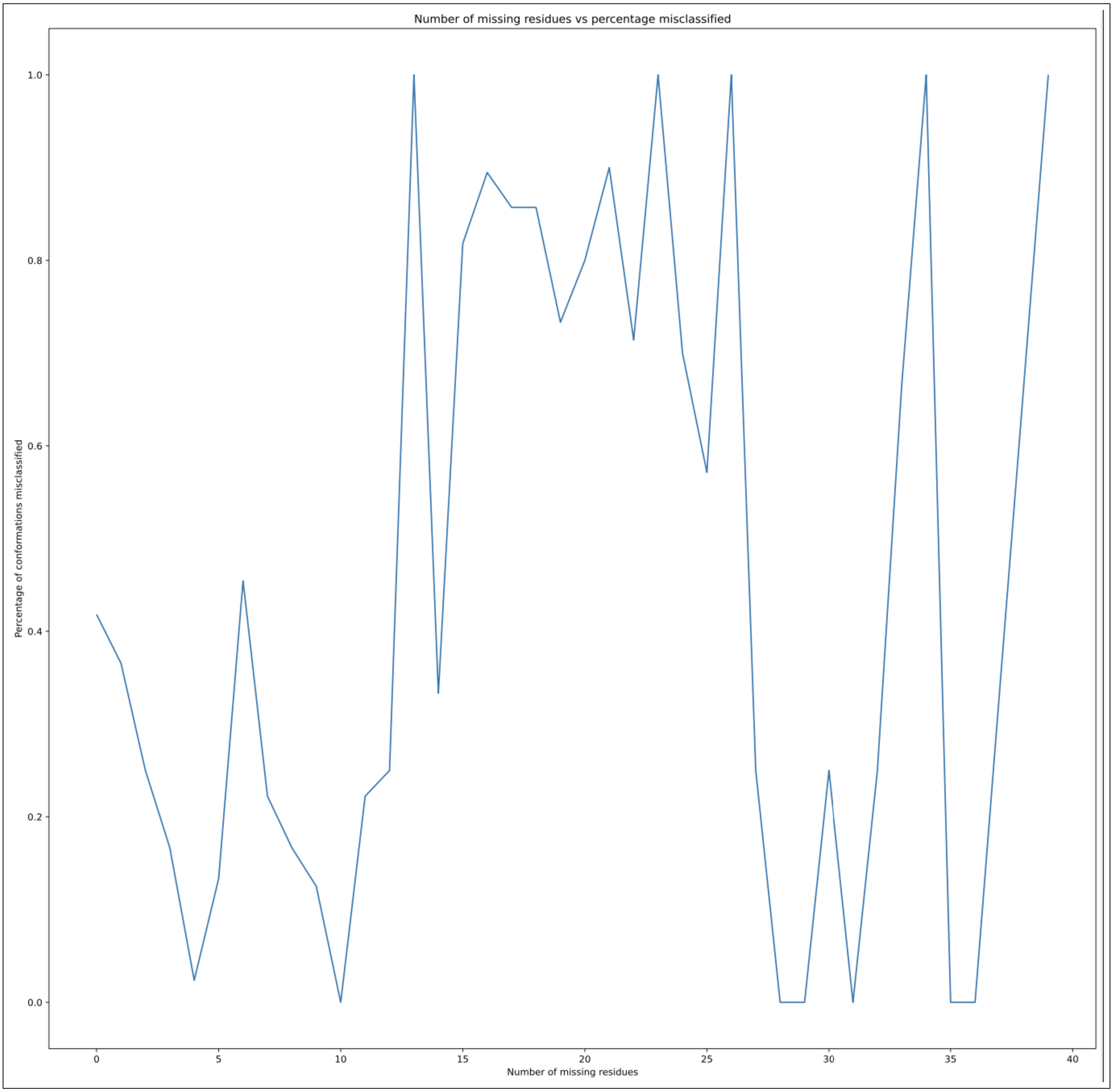


Figure 2.14: Misclassification vs Missing Residues: Indeed this graph resembles that of RMSD for the same full chain, surprisingly high error rates for low numbers of missing residues.

<u>Statistics:</u>

Table 2.4: complete T-SNE based map trials

	Precision	Recall	F1 score
Trial 1	0.584171	0.582976	0.559980
Trial 2	0.575304	0.574047	0.550408
Trial 3	0.573939	0.558055	0.536026
Trial 4	0.580932	0.579366	0.555101
Trial 5	0.576374	0.575600	0.553737
Trial 6	0.567864	0.549631	0.526155
Trial 7	0.573049	0.571990	0.548872
Trial 8	0.565355	0.547560	0.525323
Trial 9	0.577491	0.576454	0.553656
Trial 10	0.572089	0.556851	0.533822
Trial 11	0.575575	0.575251	0.552178
Trial 12	0.575733	0.575600	0.553278
Trial 13	0.583085	0.581773	0.558354
Trial 14	0.581456	0.581773	0.558655
Trial 15	0.579057	0.578861	0.556437
Trial 16	0.572809	0.556851	0.534318
Trial 17	0.578439	0.568536	0.545436
Trial 18	0.583035	0.582122	0.560065
Trial 19	0.580831	0.579715	0.556818
Trial 20	0.583529	0.582976	0.559514

Table 2.5: complete T-SNE based map trials, Average Metrics

	Precision	Recall	F1 score
open inactive	0.117027	0.162500	0.135928
closed inactive	0.931869	0.566787	0.704379
open active	0.682121	0.986111	0.806414
Averages	0.577006	0.571799	0.548907

Table 2.6: complete T-SNE based map trials, Average Cross Classification Statistics

	Real open inactive	Real closed inactive	Real open active
Predicted open inactive	14.95	113.4	0.00
Predicted closed inactive	9.20	157.0	2.25
Predicted open active	67.85	6.6	159.75

Analysis:

The misclassification percentage vs the missing residues graph resembled that of the RMSD for the similar full chain without discarding conformations missing residues. This technique, however, as opposed to the RMSD, did slightly better and created purer and more distinct clusters, especially with open active and open inactive group, it mostly suffered from the shift of the clustering a bit right. Perhaps a change of clustering metric or method would give a higher F1 score.

2.2.2 T-SNE Based Map Clustering of CA of Activation Loop, No Missing Residues

This experiment was run on the highest variance subsequence derived from the technique as mentioned above, after throwing out all of the conformations with more than 1 residue missing within the subsequence in focus.

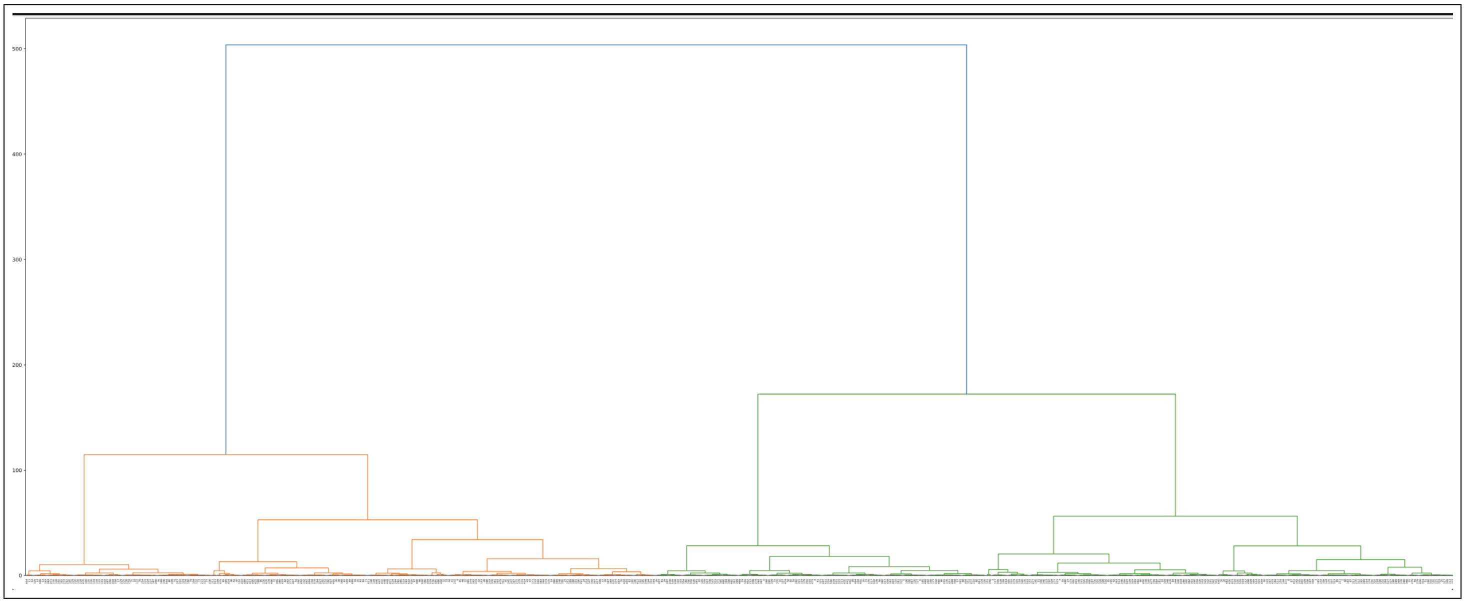


Figure 2.15: dendrogram of the complete T-SNE based map clustering

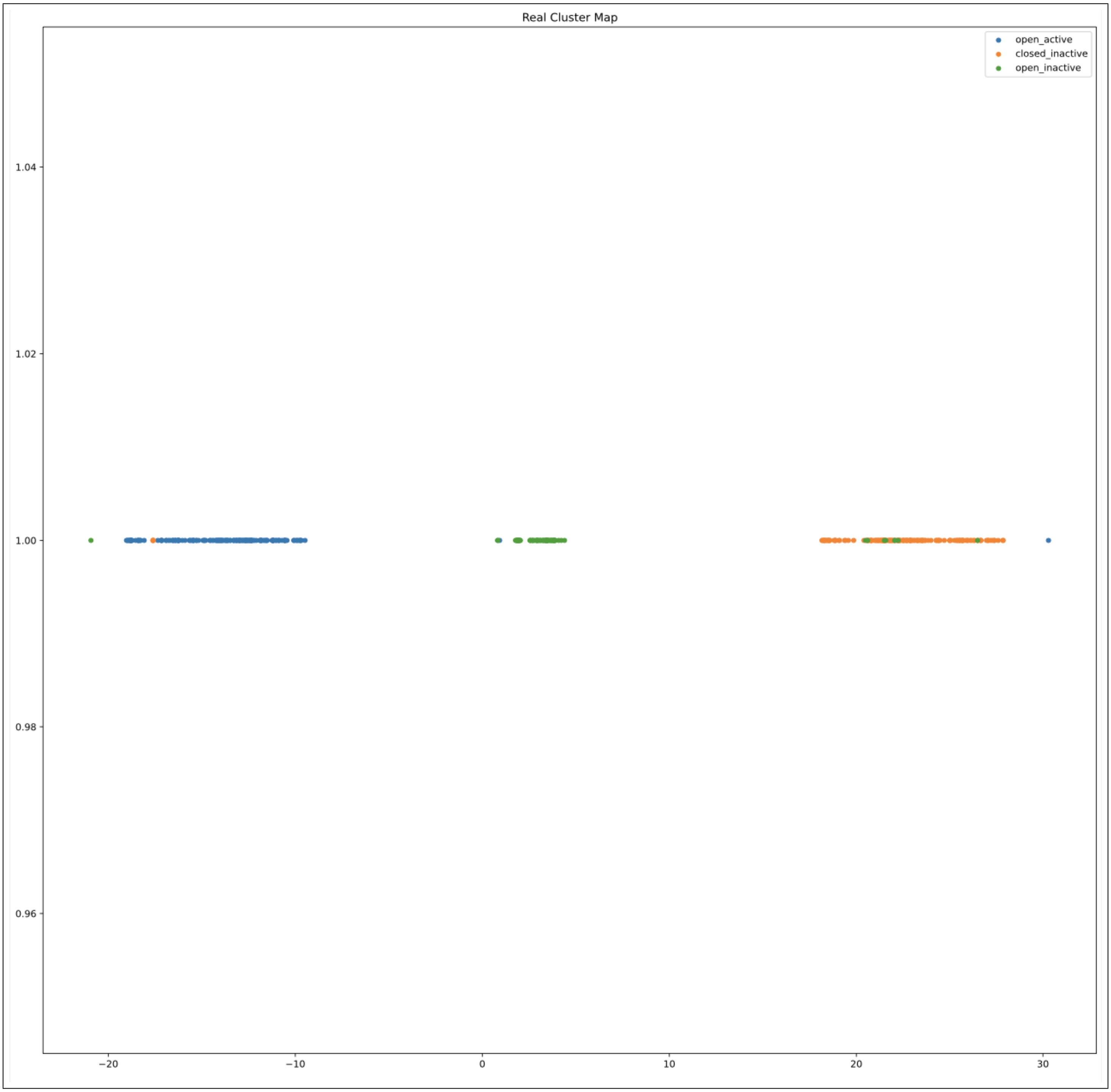


Figure 2.16: Real Cluster T-SNE projected map of T-SNE based clustering of the whole sequence, conformations with imputed values for missing. Clear distinct clusters can be found on the map containing each group. A couple of outliers which make an insignificant portion of the data set. This is a step up from the performance of RMSD, which struggled to create clear clusters.

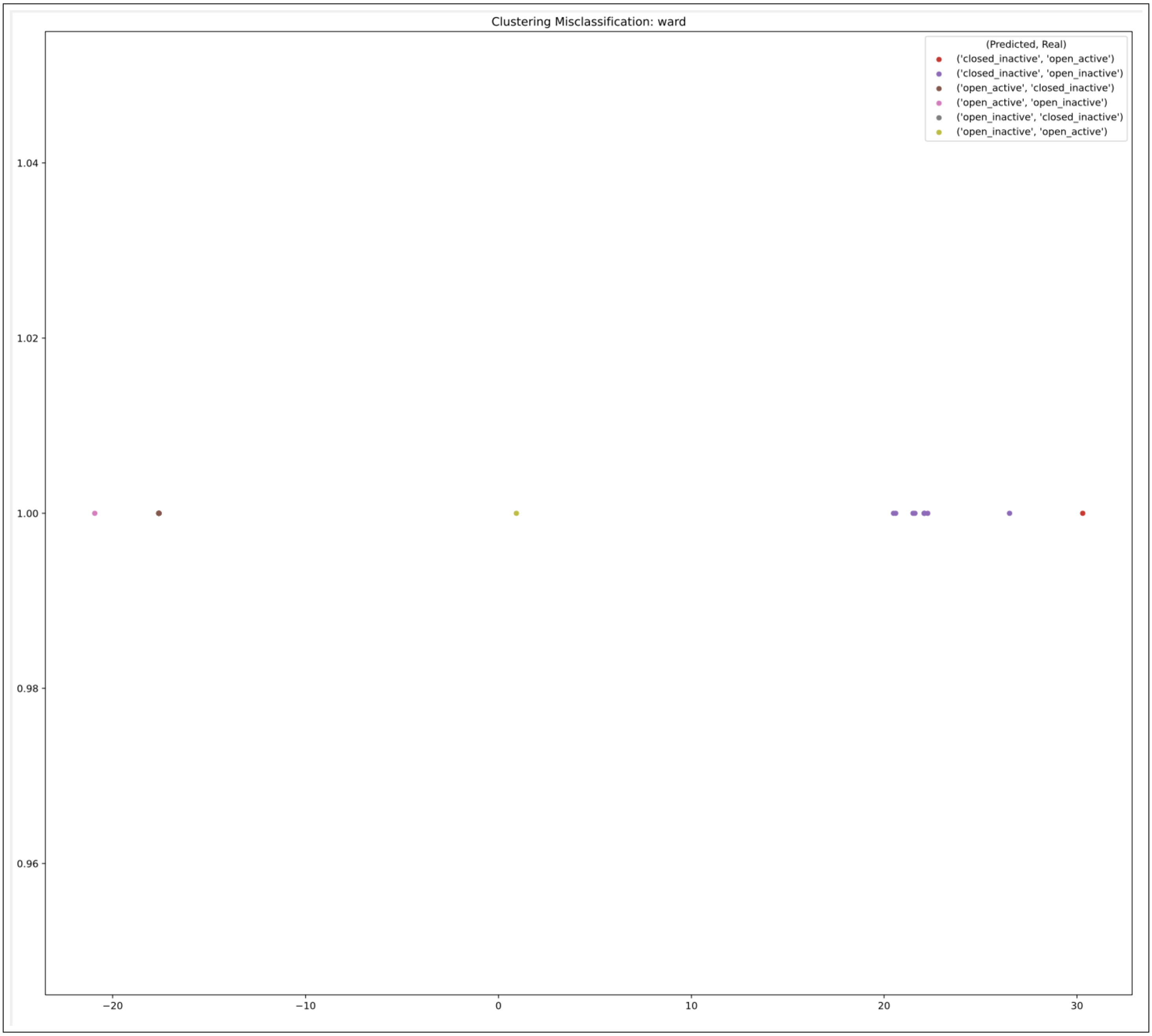


Figure 2.17: Misclassification of T-SNE based map clustering of Activation loop, No missing residues: Misclassification can be clearly seen to be outliers with unexpected properties like 5UQ1 A and C, which was analyzed in the RMSD section above, and 1FQ5 A and C

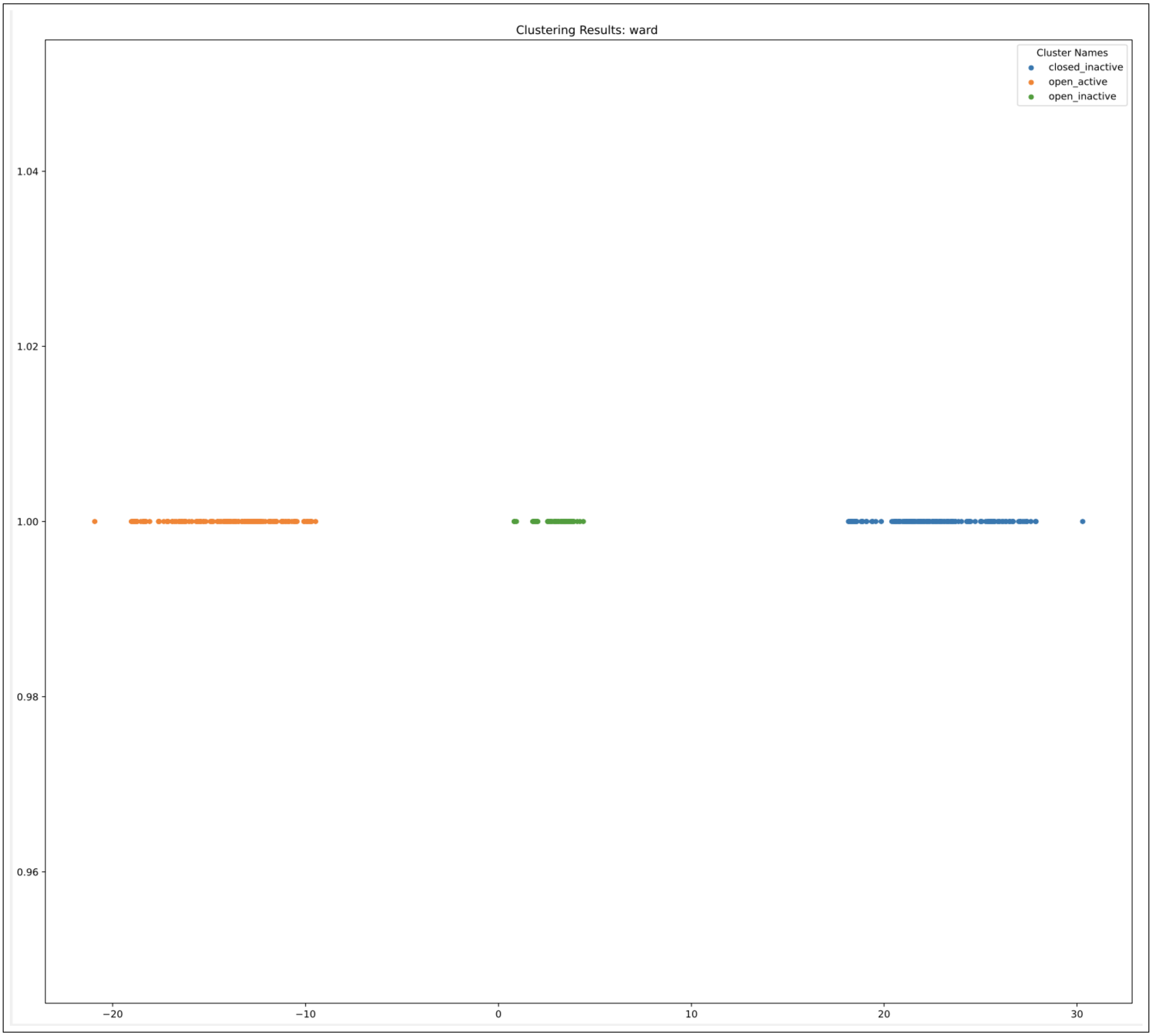


Figure 2.18: Clustering Map of T-SNE Based map clustering of activation loop using ward's algorithm, No missing residues: They seem to be clustered in the exact same way as the annotation real cluster

Statistics:

Table 2.7: Activation Loop with no missing residues, T-SNE Based Map Clustering trials

	Precision	Recall	F1 score
Trial 1	0.962589	0.948339	0.954828
Trial 2	0.957627	0.925218	0.938274
Trial 3	0.945022	0.928461	0.935884
Trial 4	0.954280	0.939676	0.946329
Trial 5	0.949099	0.930364	0.938615
Trial 6	0.932497	0.921592	0.926599
Trial 7	0.946198	0.937523	0.941568
Trial 8	0.946597	0.935220	0.940486
Trial 9	0.953146	0.925798	0.937264
Trial 10	0.941466	0.909906	0.922616
Trial 11	0.941744	0.930325	0.935497
Trial 12	0.939410	0.926197	0.932202
Trial 13	0.955556	0.941689	0.947995
Trial 14	0.963440	0.941290	0.950905
Trial 15	0.955269	0.930364	0.941002
Trial 16	0.951299	0.937303	0.943682
Trial 17	0.959327	0.939387	0.948191
Trial 18	0.942283	0.928680	0.934812
Trial 19	0.943051	0.923894	0.932314
Trial 20	0.900812	0.897148	0.898896
Average	0.944180	0.927715	0.934945

Table 2.8: Activation Loop with no missing residues, Average Metrics

	Precision	Recall	F1 score
closed_inactive	0.931295	0.953289	0.942127
open_active	0.954195	0.974375	0.964127
open_inactive	0.947051	0.855479	0.898581
Averages	0.944180	0.927715	0.934945

Table 2.9: Activation Loop with no missing residues, Average Cross Classification Statistics

	Real closed inactive	Real open active	Real open inactive
Predicted closed inactive	145.50	1.80	8.65
Predicted open active	5.55	156.35	1.80
Predicted open inactive	0.95	1.85	62.55

Analysis:

The results vary by initialization, however, probably by only a few cross classification of samples at at time. Between the groups clear and distinct clusters are formed since Precision Recall F1 score are consistently high for each run. The problem between closed inactive predicted at open active at 5.55 average cross classification is solved. In general the results are consistent across the trials. Beyond this the clusters are seen to be distinct and consistent with the annotated groupings after the T-SNE map. The method proved to do well all around given full data.

2.3 Dihedral Based Clustering

Dihedral based clustering is a method implemented by Protein Structural Statistics. The method uses a distance matrix with the a function of the distance metrics defined by the circular distance, where T is the period. The metric measures dihedral angles within residues which provide an insight to structure. The distance of the k-th dihedral angle of the i-th and j-th conformation is given by:

$$d_k(\theta(i), \theta(j)) = min \begin{cases} |\theta_k(i) - \theta_k(j)| \\ T - |\theta_k(i) - \Theta_k(j)| \end{cases} \tag{2.1}$$

and the Distance matrix is defined by:

$$D(i,j) = \frac{1}{\sqrt{N_{var}}}\sqrt{\Sigma_{k=1}^{N_{var}} d(\theta_k(i), \theta_k(j))^2} \tag{2.2}$$

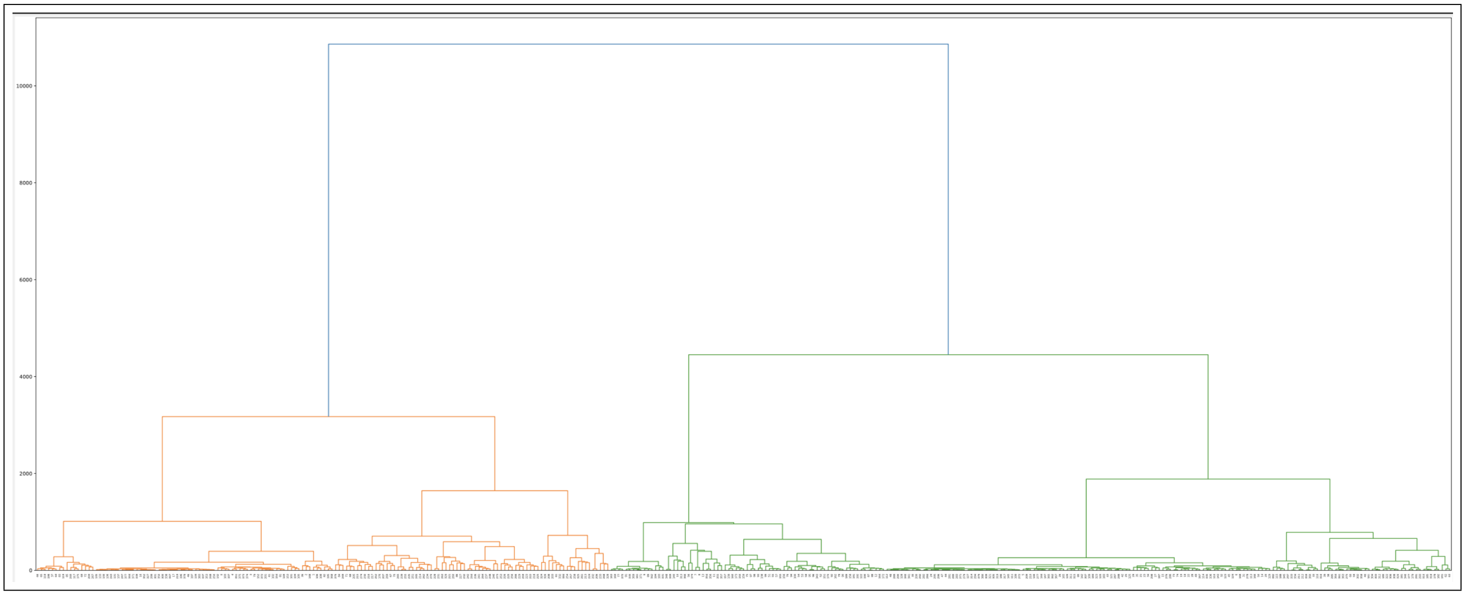


Figure 2.19: Dendrogram of Dihedral-Based clustering of activation loop

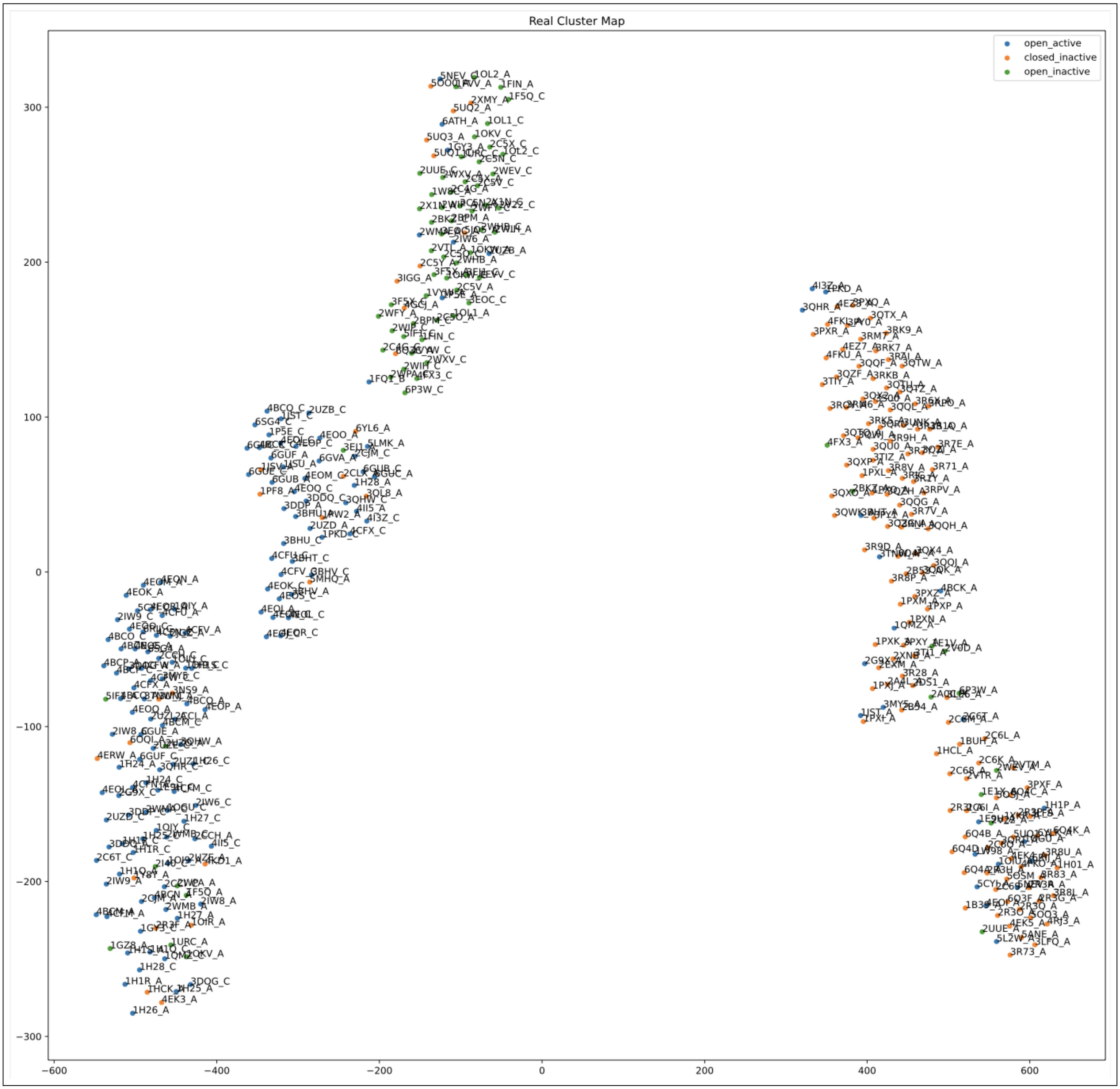


Figure 2.20: Real Cluster of Dihedral-Based clustering of activation loop: Except for the closed inactive on the right the three groups are mixed on the two left clusters. Despite the more distinct clusters on the map, the clustering might be misleading.

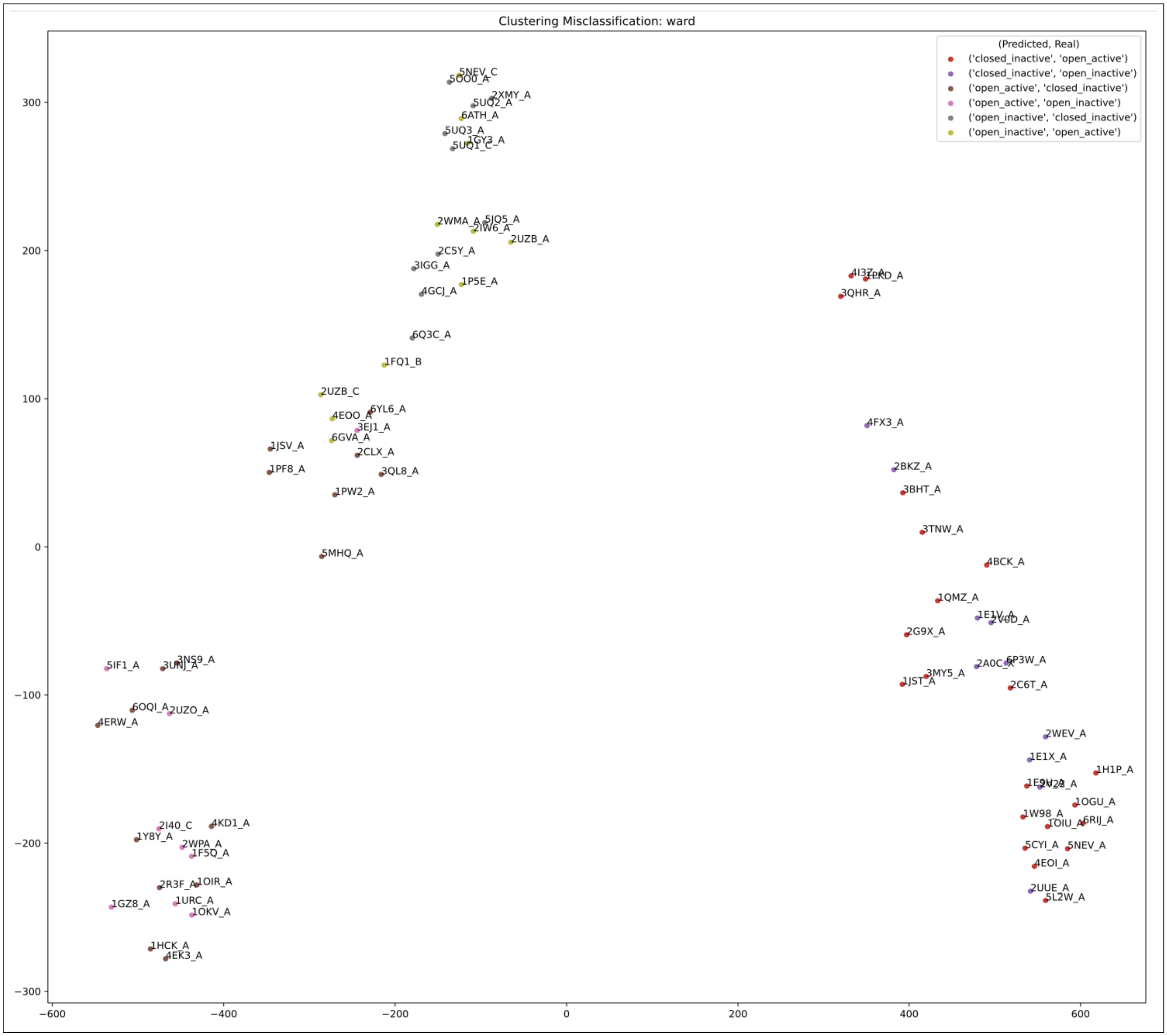


Figure 2.21: Misclassification of Dihedral-Based clustering of activation loop

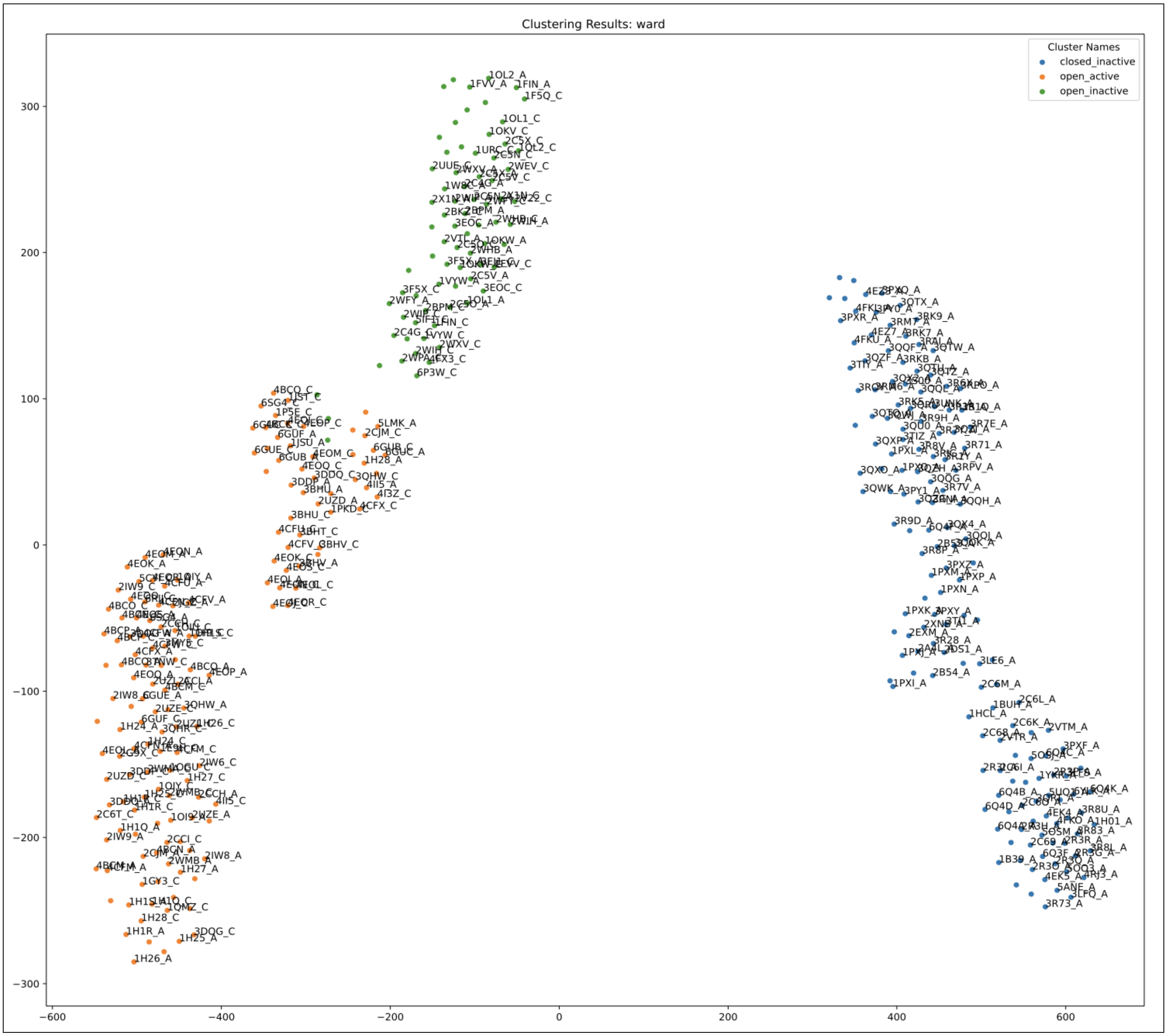


Figure 2.22: Clustering Map of Dihedral-Based clustering of activation loop

Statistics:

Table 2.10: Dihedral-Based Clustering: Activation Loop with no missing residues, Statistics Metrics

	Precision	Recall	F1 score
closed inactive	0.800000	0.821192	0.810458
open active	0.831169	0.800000	0.815287
open inactive	0.720000	0.739726	0.729730
Averages	0.783723	0.786973	0.785158

Table 2.11: Dihedral-Based Clustering: Activation Loop with no missing residues, Cross Classification Statistics

	Real closed inactive	Real open active	Real open inactive
Predicted closed inactive	124	21	10
Predicted open active	17	128	9
Predicted open inactive	10	11	54

Analysis:

Unlike the RMSD of the activation loop, which sacrificed precision for recall or the F1 score of one group for another, since the dihedral based matrix was evenly spread with a higher concentration of each annotated group in each cluster, the dihedral-based classification method did well all-around since it gives insight into shape and structure within a conformation. However, clusters were not as distinct as the previous two methods which used functions of physical coordinates of the conformation that played a larger role than bond angles.

Conclusion

3.1 Summary of methods

The RMSD matrix and T-SNE based map clustering both gave consistent results and clustered the conformations well. The RMSD based clustering although less costly and more stable, occasionally was thwarted by the high intra-clusteral distances although the conformations were well clustered.

The T-SNE based map, on the other hand, despite, being a bit more costly requiring multiple trials, and certain user inputs, much more defined distinct clusters based on position instead of relative position as RMSD which showed fruition in out performing both RMSD and dihedral based clustering in the statistical metrics. Some part of the distinctness of the clusters can be attributed to extra T-SNE projection with the high perplexity which with the Kullback-Leiber divergence penalization, created clear neighborhoods for each cluster. In both cases the methods applied on specific regions with high structural variance both performed much better as it amplified the differences previously suppressed by the rest of the sequence which were pair-wise similar.

The T-SNE based map clustering, in general provides user flexibility— able to control for certain high variance regions, to pick and choose each segment and include them as features and has consistently out performed the other methods. It has some variable parameters like clustering perplexity and initialization, which can however, be generalized since high perplexity will always form distinct clusters and multiple runs can give an average result.

RMSD is straight-forward and robust, however, occasionally doesn't create clear distinct clusters, given that we don't know the actual groupings in actual applications, can be problematic.

Dihedral-based is straight-forward and robust as well, giving a well balanced result. On the other hand it doesn't cluster the conformations as well as the previous two methods often mixing different groups in the same cluster, this is penalized in the statistical metrics.

3.2 Discussion

The method for T-SNE based map clustering is very rough and many more optimizations can be built upon it. For example currently the imputation method is very primitive only taking the averages of the coordinates of the same indices in other conformations this can lead to the formation of an artificial cluster made from the averages of two groups. Perhaps a method which might take into account the coordinates of the neighboring residues in a weighted manner to impute a more precise result.

On top of this the average coordinate, is quite an efficient map in this case, yet in order to deal with symmetrical substructures which have the same average coordinates a more sophisticated map can be devised. A proposition of using the current methods to address the problem is to again cut the subsequence into half and use each half as a feature instead as all together, however, this isn't very automated.

As seen throughout both of the T-SNE based map clustering and RMSD matrix, despite the correct clustering of the methods, intra-cluster distances tend to hamper with the results causing a proportion of the correctly clustered group to be misclassified. Perhaps different distance metrics or clustering techniques could be used to address this issue.