Study of GSK3 active site and clustering approaches

Algorithms in Structural Biology 2023

Instructors: Ioannis Emiri & Evagelia Chrysina

Marina Thalassini Filippidou 7115152200032

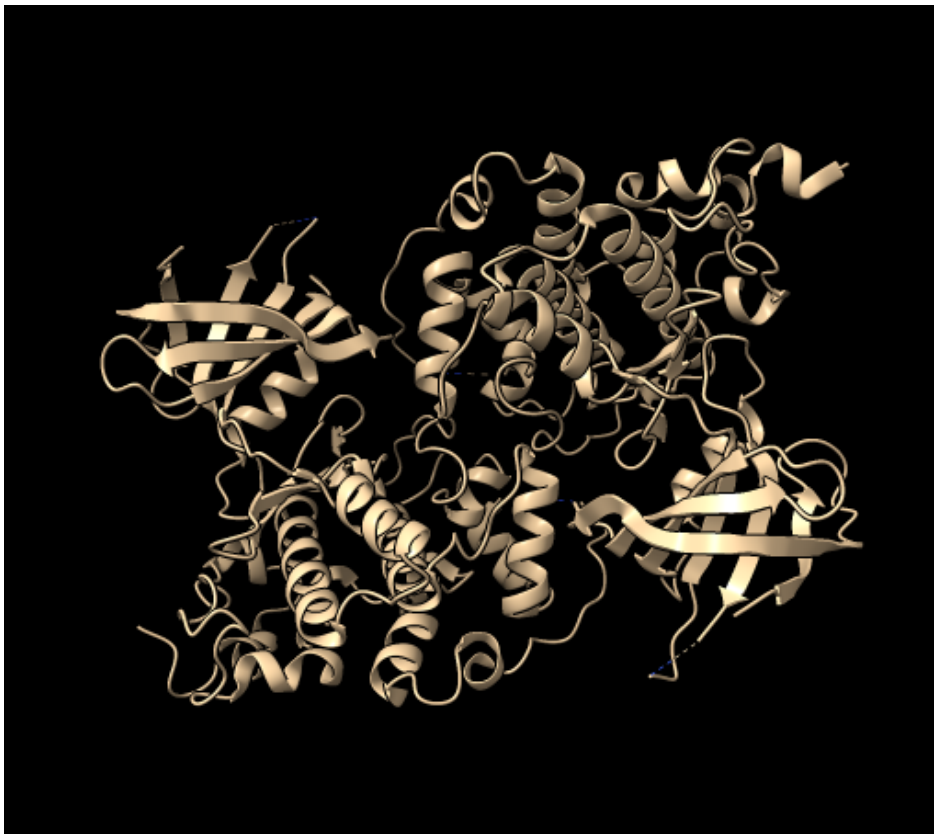


Fig1. Visualization of 8djd.pdb using ChimeraX

Contents

Abstract

This research focuses on exploring the active sites of GSK3 (glycogen synthase kinase 3) kinases, which play a crucial role in enzyme function and have implications for drug development. We compare the performance of two advanced prediction tools in identifying GSK3 binding sites, using known binding sites from the literature as a benchmark. Our computational pipeline is based on the work of Kostantina Roka and generates all the possible conformations of the binding site of a specific protein given its active site, a set of pdb strucures and the pdb files of the amino acid rotamers. Clustering algorithms are then applied to analyze and identify structural patterns within the conformations. By following this systematic approach, our study aims to enhance our understanding of GSK3 active sites and contribute to future drug design endeavors in the field of enzymology.

Introduction

Protein kinases play a pivotal role in cellular signaling by facilitating the transfer of the terminal phosphoryl group from nucleoside triphosphates to specific amino acid residues in protein substrates. Among the various amino acids, serine and threonine have been identified as the primary acceptors of phosphorylation by protein kinases. This phosphorylation process serves as a key regulatory mechanism in numerous cellular pathways.

It is worth noting that dysregulation of protein kinases has been implicated in a wide range of human malignancies. Mutations, chromosomal rearrangements, and gene amplification events can lead to aberrant activity and expression levels of protein kinases, disrupting normal cellular signaling processes. Consequently, this dysregulation often contributes to the development and progression of various human disorders, including cancer.

Given their pivotal role in cellular signaling and their association with human diseases, protein kinases have emerged as important therapeutic targets. Inhibition of aberrant protein kinase activity has shown promising potential for the development of novel treatments for human disorders, including targeted therapies for cancer. The identification and characterization of specific protein kinases involved in disease pathways provide valuable insights for designing effective therapeutic strategies aimed at restoring normal cellular signaling and combating human disorders.

Glycogen synthase kinase 3

Glycogen synthase kinase 3 (GSK-3) is a protein kinase that plays a crucial role in a wide range of essential biological processes. Its involvement in glucose regulation, apoptosis, protein synthesis, cell signaling, cellular transport, gene transcription, proliferation, and intracellular communication highlights its significance in various disease pathways and makes it a promising target for therapeutic interventions and medical imaging.

GSK-3 is expressed in all tissues and belongs to the protein kinase family responsible for transferring phosphate groups from ATP to specific target substrates. Acting as a serine/threonine kinase, GSK-3 phosphorylates amino acid residues of its substrates, primarily serine and threonine. This phosphorylation process serves as a regulatory mechanism, modulating complex biological processes such as glucose metabolism, cell signaling, cellular transport, apoptosis, proliferation, and intracellular communication. As our knowledge of GSK-3 expands, we anticipate the discovery of additional roles in various biological processes. Given its involvement in diverse cellular pathways, GSK-3 has attracted attention as a valuable target for drug development and medical imaging in numerous diseases. Over the past two decades, the fascination with GSK-3 and its diverse actions has grown exponentially, as it appears to impact nearly every aspect of cellular signaling and is implicated in an unparalleled number of disease processes.

GSK-3 gained recognition as a therapeutic target for diabetes following its discovery in insulin signal transduction pathways in the mid-1990s. One of its key functions is regulating glycogen synthase (GS) activity, an enzyme responsible for converting glucose into glycogen. Dysregulation of GSK-3 has been associated with several diseases, including diabetes, inflammation, cancer, Alzheimer's disease, and bipolar disorder. In individuals with type II diabetes and obese animal models, increased expression and activity of GSK-3 have been observed, leading to impaired insulin-mediated glycogen synthesis and glucose homeostasis. Inhibitors targeting GSK-3 have shown potential anti-diabetic effects in laboratory studies and animal models. However, developing selective GSK-3 inhibitors remains challenging due to the need for specificity toward a kinase involved in multiple pathways with numerous substrates, which may lead to unintended side effects and toxicity.

Methods

Dataset

Our study utilized two distinct datasets to facilitate the investigation of kinases in the GSK3 family and their active sites. The first dataset comprised 33 specific kinases of the GSK3 family obtained from the Protein Data Bank (PDB). This dataset primarily served as the basis for active site prediction. For the subsequent steps in the pipeline, we employed a more extensive dataset. This comprehensive dataset consisted of all the kinases belonging to the GSK3 family available in the PDB. By incorporating a broader range of kinases, we aimed to gain a more comprehensive understanding of the similarities and differences within the GSK3 family's active sites.

To prepare the datasets for analysis, we downloaded the corresponding PDB files. To ensure consistency and facilitate comparisons, we employed the software tool COOT to superimpose the structures. This process enabled us to generate an average structure using a python script that served as a representative template for further analyses and comparisons within the pipeline. For rotamers, we utilized the Richardson's Penultimate Rotamer Library. From this library, we selected and isolated the specific amino acids relevant to our analysis of the GSK3 kinase family's active sites. This enabled us to explore the conformational variations and flexibility within these crucial amino acids.

We can explore some of our amino acid rotamers pdb file by using Chimera visualization software.

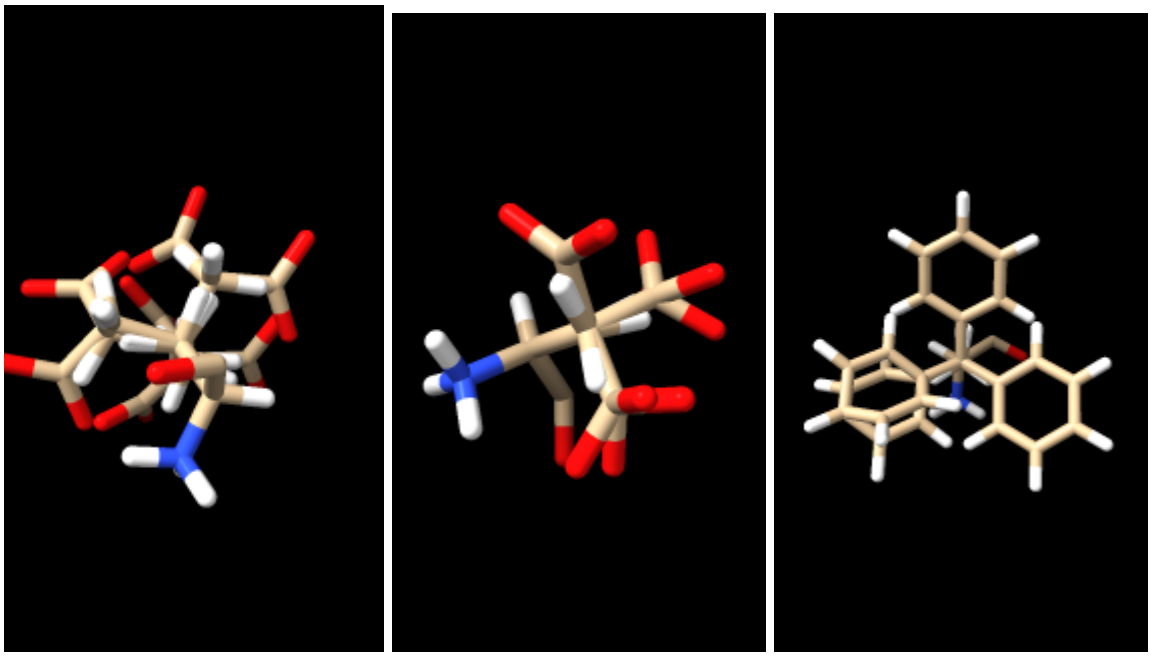


Fig2. Glutamic acid (GLU), Aspartic acid (ASP),Phenylalanine (PHE)

Retrieval of Binging site

To retrieve the real binding site, we relied on a specific publication titled "The Structure of Phosphorylated GSK-3β Complexed with a Peptide, FRATtide, that Inhibits β-Catenin Phosphorylation." This study described the binding interaction between GSK-3 and a peptide called FRATtide, which acts as an inhibitor of β-catenin phosphorylation. The experimental data, including the structure factors and coordinates for the GSK-3β/FRATtide complex, were deposited in the Protein Data Bank under the accession code 1gng.

To access more information about the specific PDB file, we utilized the PDBe-Knowledge Base (https://www.ebi.ac.uk/pdbe/pdbe-kb/). By searching for the code 1gng on the website, we obtained two results corresponding to the FRATtide (FRAT1_HUMAN, Q92837) and GSK-3 kinase (GSK3B_HUMAN, P49841). To obtain detailed annotations for the GSK-3 kinase, we selected "View Details" and then navigated to the "Annotations" section. From there, we downloaded the "All Annotations" CSV file (P49841-annotations.csv), which provided us with information about the amino acids comprising the catalytic center. By searching for the rows corresponding to "Ligand binding sites," we could extract the necessary details about the binding site's specific amino acids.

2964	P49841,Ligand binding sites,ADP,63,63,Type: Ligand binding site Range: GLY63 - GLY63 Ligand: ADP Observed in 2 entries; e.g. 4nm3;
2965	P49841,Ligand binding sites,ADP,64,64,Type: Ligand binding site Range: ASN64 - ASN64 Ligand: ADP Observed in 2 entries; e.g. 4nm7;
2966	P49841,Ligand binding sites,ADP,65,65,Type: Ligand binding site Range: GLY65 - GLY65 Ligand: ADP Observed in 2 entries; e.g. 4nm3;
2967	P49841,Ligand binding sites,ADP,67,67,Type: Ligand binding site Range: PHE67 - PHE67 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2968	P49841,Ligand binding sites,ADP,70,70,Type: Ligand binding site Range: VAL70 - VAL70 Ligand: ADP Observed in 3 entries; e.g. 4nm7;
2969	P49841,Ligand binding sites,ADP,83,83,Type: Ligand binding site Range: ALA83 - ALA83 Ligand: ADP Observed in 2 entries; e.g. 4nm0;
2970	P49841,Ligand binding sites,ADP,85,85,Type: Ligand binding site Range: LYS85 - LYS85 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2971	P49841,Ligand binding sites,ADP,110,110,Type: Ligand binding site Range: VAL110 - VAL110 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2972	P49841,Ligand binding sites,ADP,132,132,Type: Ligand binding site Range: LEU132 - LEU132 Ligand: ADP Observed in 1 entry; e.g. 4nm5;
2973	P49841,Ligand binding sites,ADP,133,133,Type: Ligand binding site Range: ASP133 - ASP133 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2974	P49841,Ligand binding sites,ADP,134,134,Type: Ligand binding site Range: TYR134 - TYR134 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2975	P49841,Ligand binding sites,ADP,135,135,Type: Ligand binding site Range: VAL135 - VAL135 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2976	P49841,Ligand binding sites,ADP,138,138,Type: Ligand binding site Range: THR138 - THR138 Ligand: ADP Observed in 3 entries; e.g. 4nm7;
2977	P49841,Ligand binding sites,ADP,141,141,Type: Ligand binding site Range: ARG141 - ARG141 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2978	P49841,Ligand binding sites,ADP,185,185,Type: Ligand binding site Range: GLN185 - GLN185 Ligand: ADP Observed in 4 entries; e.g. 4nm3;
2979	P49841,Ligand binding sites,ADP,186,186,Type: Ligand binding site Range: ASN186 - ASN186 Ligand: ADP Observed in 3 entries; e.g. 4nm7;
2980	P49841,Ligand binding sites,ADP,188,188,Type: Ligand binding site Range: LEU188 - LEU188 Ligand: ADP Observed in 4 entries; e.g. 4nm7;
2981	P49841,Ligand binding sites,ADP,200,200,Type: Ligand binding site Range: ASP200 - ASP200 Ligand: ADP Observed in 5 entries; e.g. 4nm0;
2982	P49841,Ligand binding sites,ANP,6,6,Type: Ligand binding site Range: ARG6 - ARG6 Ligand: ANP Observed in 3 entries; e.g. 1o6l;
2983	P49841,Ligand binding sites,ANP,9,9,Type: Ligand binding site Range: SER9 - SER9 Ligand: ANP Observed in 3 entries; e.g. 1o6l;
2984	P49841,Ligand binding sites,ANP,63,63,Type: Ligand binding site Range: GLY63 - GLY63 Ligand: ANP Observed in 2 entries; e.g. 1pyx;
2985	P49841,Ligand binding sites,ANP,64,64,Type: Ligand binding site Range: ASN64 - ASN64 Ligand: ANP Observed in 2 entries; e.g. 1pyx;
2986	P49841,Ligand binding sites,ANP,65,65,Type: Ligand binding site Range: GLY65 - GLY65 Ligand: ANP Observed in 2 entries; e.g. 1pyx;
2987	P49841,Ligand binding sites,ANP,67,67,Type: Ligand binding site Range: PHE67 - PHE67 Ligand: ANP Observed in 1 entry; e.g. 1j1b;
2988	P49841,Ligand binding sites,ANP,70,70,Type: Ligand binding site Range: VAL70 - VAL70 Ligand: ANP Observed in 2 entries; e.g. 1pyx;
2989	P49841,Ligand binding sites,ANP,83,83,Type: Ligand binding site Range: ALA83 - ALA83 Ligand: ANP Observed in 2 entries; e.g. 1pyx;

Fig3. P49841-annotations.csv

From the csv file we can access information about the amino acids and respective positions of the binding site for various ligands as we see in the picture above. We selected specifically the 17 amino acids and positions of the binding site with the ADP molecule and extracted the information in a txt file in the following format in order to use it for the generation of the conformations of the binding site.

GSK3_activesite.txt

```
63 GLY
64 ASN
65 GLY
67 PHE
70 VAL
83 ALA
85 LYS
110 VAL
132 LEU
133 ASP
134 TYR
138 THR
141 ARG
185 GLN
186 ASN
188 LEU
200 ASP
```

Fig3. PSK_activesite.txt

Binding Site prediction tools

In our quest to explore the active sites of kinases in the GSK3 family, we initiated our investigation by examining two widely used active site prediction tools: P2Rank and Deep Pocket. Previous related work has extensively utilized P2Rank, making it a natural choice for our comparative analysis. However, given the advancements in the field, we sought to determine if Deep Pocket could offer improved performance in identifying active sites. The decision to include Deep Pocket in our study was further reinforced by its open-source nature, allowing us to leverage its capabilities without any limitations. Additionally, the scores reported in Deep Pocket's publication indicated superior predictive accuracy compared to P2Rank, further motivating our choice to include Deep Pocket as a key component in our active site analysis.

Deep Pocket’s Workflow and published results

We focused our attention on DeepPocket, an innovative framework that combines geometry-based software with deep learning techniques. This unique approach integrates 3D convolutional neural networks to enhance the accuracy of active site identification. DeepPocket builds upon the initial pocket identification performed by Fpocket, a well-established geometry-based tool. However, DeepPocket takes the process a step further by employing neural networks to rescore these identified pockets, ensuring a more precise and refined selection. Furthermore, DeepPocket goes beyond traditional pocket identification by segmenting the identified cavities on the protein surface, providing a more granular analysis of the active site architecture. This combination of geometry-based software and deep learning in DeepPocket offers a promising avenue for improving the accuracy and reliability of active site prediction, making it a valuable tool for our study.

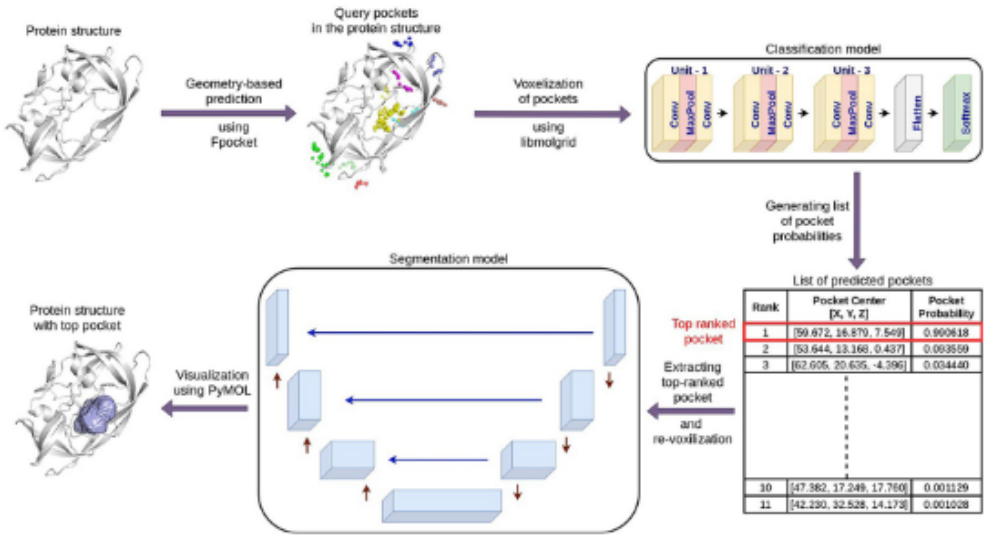


Fig5. Deep Pocket’s pipeline

DeepPocket Workflow

The DeepPocket framework simplifies the process of active site prediction through a straightforward workflow. The first step involves preparing the input structure by removing heteroatoms and solvent molecules using the Biopython47 library, ensuring a clean and focused analysis. Next, Fpocket is employed to identify potential pockets within the protein structure. The barycenter, calculated for each predicted pocket, serves as a representative point within the pocket and acts as a candidate pocket center. To rank these candidate pocket centers, constant-sized grids are placed at each barycenter, and a CNN scoring function evaluates their features. This ranking process helps prioritize the most promising pockets. The top-ranked centers then undergo a CNN segmentation model, refining the analysis and generating the final pocket structure. This segmentation step provides a more detailed and accurate representation of the active site.

Based on the published results available, Deep Pocket demonstrates superior performance compared to p2rank across various datasets. The table depicting Deep Pocket's results clearly indicates its efficacy in active site prediction. In nearly all cases, Deep Pocket outperforms p2rank, showcasing its superiority in identifying and predicting active sites. These findings further support our decision to utilize Deep Pocket as a primary active site prediction tool in our study.

Table 1. DCA Results Comparison

	COACH420		HOLO4K		SC6K	
	Top-n	Top-(n + 2)	Top-n	Top-(n + 2)	Top-n	Top-(n + 2)
Fpocket	35.09%	51.25%	36.34%	51.53%	23.99%	37.23%
Deepsite	53.07%	53.07%	51.65%	51.67%	52.94%	65.41%
Kalasanty	63.51%	65.18%	61.21%	62.63%	61.75%	61.75%
P2Rank	**68.24%**	75.48%	70.6%	80.05%	62.9%	75.74%
DeepPocket	67.96%	**79.94%**	**73.36%**	**82.97%**	**64.58%**	**83.01%**

Fig6. Deep Pocket's results

Utilization of Deep Pocket

Deep Pocket's impressive performance initially sparked our enthusiasm; however, the challenges encountered during its execution on our local computer dampened our spirits. The

Deep Pocket package, available on GitHub, had several dependencies, including Fpocket, PyTorch, libmolgrid, Biopython, and other commonly used Python packages. Each of these dependencies had its own set of requirements, adding complexity to the setup process.

We encountered a significant challenge with one specific requirement of libmolgrid, namely CUDA. CUDA is a parallel computing platform and application programming interface developed by NVIDIA, designed to harness the power of GPUs (Graphics Processing Units). Unfortunately, our local computer lacked the necessary NVIDIA GPU with the required specifications for CUDA.

To overcome this limitation, we embarked on modifying the code to enable it to run on a CPU instead of a GPU. Through persistent efforts, we successfully modified and debugged the code, making it compatible with our local PC. Despite the initial challenges, our perseverance paid off, as we were ultimately able to utilize Deep Pocket for our study.

P2rank

P2Rank is a machine learning-based method specifically designed for the prediction of binding sites using protein structures. This approach utilizes a Random Forests classifier to infer the ligandability of local chemical neighborhoods situated near the protein surface. These neighborhoods are represented by specific near-surface points and characterized by aggregating physico-chemical features projected from neighboring protein atoms onto these points. Through the prediction algorithm, P2Rank identifies points with high predicted ligandability, which are subsequently clustered and ranked to generate a comprehensive list of binding site predictions. For our work, we made use of the freely available P2Rank tool, accessible at https://prankweb.cz/.

Deep Pocket vs P2rank

Two distinct approaches were employed for predicting the binding site. The first approach involved providing the prediction algorithm with the average structure derived from all the PDB files. In contrast, the second approach involved feeding the algorithm with individual PDB files and subsequently determining the intersection of their respective predictions. These alternative strategies aimed to explore different avenues and perspectives in accurately identifying the binding sites within the GSK3 kinase family.

In our comparison of the results, we utilized the combination of the initial residue name and residue sequence number to determine the accuracy of the predictions. To consider a prediction correct, both the predicted name and position of the amino acid needed to match the true values. With this criterion in mind, Deep Pocket failed to predict any of the correct amino acid-sequence number combinations using either of the two methods. On the other hand, P2Rank achieved an impressive 70% accuracy by predicting 12 out of 17 correct combinations when utilizing the intersection of the predictions. However, when using the prediction based on the average pocket, P2Rank did not yield any correct predictions.

An additional improvement for the validation of these results would involve shifting the focus from individual amino acid-position combinations to identifying patterns of amino acid residues within the predicted binding sites. Instead of solely evaluating the correctness of individual predictions, we can analyze the presence and similarity of specific residue patterns. By examining the presence and consistency of these patterns in the predictions, we can gain insights into the overall quality and reliability of the tools.

Furthermore, another approach to assess the accuracy of the predictions is to calculate the root mean square deviation (RMSD) distance between the predicted binding sites and the real binding site structure. The RMSD provides a quantitative measure of the structural similarity between the predicted and real binding sites. A lower RMSD value indicates a closer match between the predicted and real structures, indicating higher accuracy in the prediction.

By incorporating these additional measures, including the analysis of residue patterns and the calculation of RMSD distances, we can further enhance our understanding of the performance and reliability of the prediction tools in accurately identifying the binding site.

Although these approaches could provide valuable insights into the quality of the predictions and the similarity between the predicted and real binding sites, they were not included in the current comparison due to time constraints. Future research endeavors may consider incorporating these measures to further improve the assessment of prediction tools and enhance our understanding of their accuracy and reliability in predicting binding sites.

Conformations Creation

For the creation of the possible conformations of the binding site we utilized the pipeline constructed by our colleague Kostantina Roka which follows the following steps:

1. Superposition on average structure and creation of rotamers' set

2. Reducing the rotamers' number based on their appearance frequency in the starting pdbs

3. Create all valid conformations of the active site (Validity: No steric clashes)

4. Creating vectors with coordinates and chi angles for clustering

We used the command line interface (cli) to run the python scripts for the creation of the conformations.

First we confirm that all the pdbs contain all the rotamers of the active site by running:
confSpace pdbprocessing --filepath <path/to/superposed_pdbs> --activesite <path/to/active_site.txt>
If not then we would need these files.

Then in order to superpose the rotamers to the average structure and remove the hydrogen atoms form the rotamer pdbs if they exist we run:
confSpace rotamersprocessing --filepath <path/to/superposed_pdbs> --activesite <path/to/active_site.txt>

After this step we can observe the superposed rotamers with no H inside the directory *<path/to/code/results/AlignedRotamers/average/noH>*
We can also observe the superposed rotamer's weights in the directory *<path/to/code/results/rotamers_weights>*

Finally to create the final coordinates and angles dataset which we will use for clustering we we run:
confSpace conformationcreation --filepath <path/to/average/noH>> --activesite <path/to/active_site.txt>

The final step of our pipeline, which involves generating diverse conformations of the binding site, can be computationally intensive, especially when dealing with a large number of amino acids in the active site. In our attempt to run the pipeline for GSK-3 with an active site of 17 amino acids, we encountered resource limitations that prevented us from completing the required computations.

To overcome this challenge and still provide insights into the clustering methods applicable to this type of data, we utilized the results obtained from the conformations of Glycogen Phosphorylase. By leveraging this dataset, we were able to demonstrate some of the potential clustering techniques that can be employed in analyzing and understanding the structural characteristics of enzyme active sites.

It is important to acknowledge that the computational demands of our pipeline can vary depending on the size and complexity of the active site. In cases where the resources are limited, alternative approaches or optimizations may be necessary to achieve the desired results.

Clustering

To perform comprehensive clustering analysis, we implemented two distinct pipelines: one for clustering the coordinates and another for clustering the angles. In both cases, we employed weighted and unweighted versions of the k-means algorithm to partition the data into clusters. The purpose of using weighted clustering was to consider the frequency of appearance of rotamers in each cluster.

After clustering, we visualized the results using two popular dimensionality reduction techniques: t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection). These methods allowed us to project the high-dimensional data into a lower-dimensional space, facilitating the visualization of clusters and potential patterns.

Additionally, we conducted cluster analysis to investigate the dominant rotamers within each cluster. By examining the composition of rotamers in each cluster, we gained insights into the prevalence and distribution of specific rotamers across the clusters.

By executing these separate pipelines and analyzing the results, we aimed to gain a deeper understanding of the clustering patterns and dominant rotamers associated with the coordinates and angles of the protein data. This comprehensive approach enabled us to explore and interpret the structural characteristics of the data from multiple perspectives.

Angles Dataset

The angle dataset consisted of 69119 rotamers each one characterized by 26 angles. To facilitate the analysis, we first scaled the data to ensure consistent ranges across the angles. Subsequently, we applied dimensionality reduction techniques, specifically UMAP and t-SNE, to reduce the dimensionality of the data while preserving its underlying structure. This allowed us

to visualize the reduced data in two-dimensional plots, providing insights into the clustering and patterns present within the dataset. By employing these dimensionality reduction methods, we aimed to gain a better understanding of the relationships and similarities among the rotamers based on their angle characteristics.

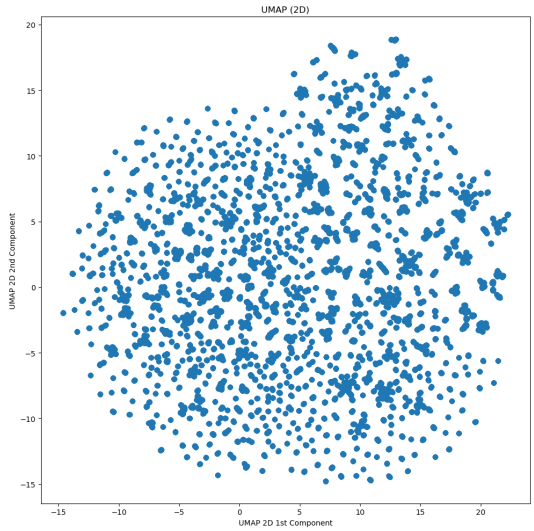


Fig7. UMAP on angles dataset

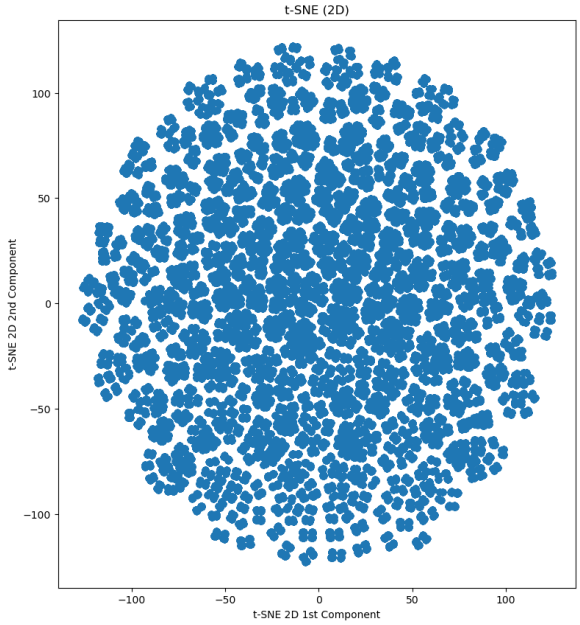


Fig8. t-SNE on angles dataset

Upon analyzing the results of t-SNE visualization, we notice an intriguing characteristic of the data: it appears to exhibit an underlying symmetry. The t-SNE algorithm has successfully captured this symmetry in the reduced-dimensional representation of the dataset. This symmetry implies that there are distinct patterns or similarities among the rotamers, which are effectively reflected in the t-SNE plot. The identification of such symmetry can provide valuable insights into the structural organization and relationships within the dataset, potentially uncovering meaningful clusters or subgroups of rotamers that share similar angle characteristics. Further investigation of this symmetry could yield valuable information for understanding the conformational space and behavior of the protein.

Weighted Clustering

After obtaining the weights for each conformation by multiplying the weights of individual rotamers, we proceeded to perform weighted k-means clustering. To determine the optimal number of clusters for our data, we constructed an elbow plot. The plot was generated by varying the number of clusters (k) in the range of 2 to 11. The elbow plot allowed us to visualize the relationship between the number of clusters and the corresponding sum of squared distances (inertia). By examining the plot, we aimed to identify the "elbow point" where the decrease in inertia becomes less significant as the number of clusters increases. This elbow point serves as an indication of the optimal number of clusters to be chosen for our analysis.

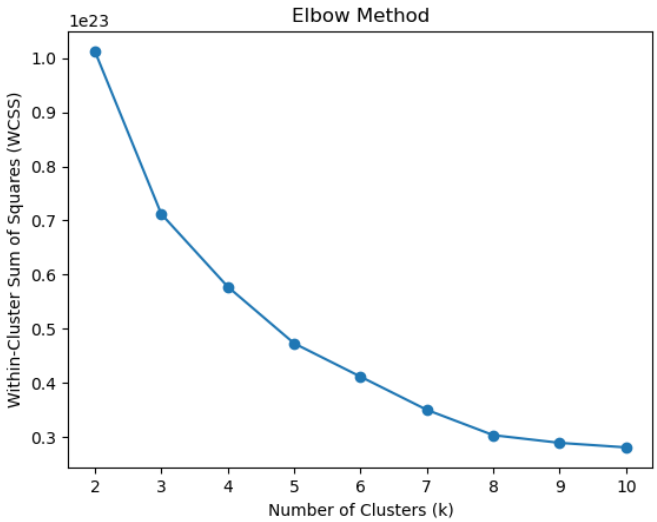


Fig9. Elbow method for weighted kMeans on angles dataset

After determining that 8 was the optimal number of clusters based on the elbow plot, we proceeded to apply the weighted k-means algorithm. The sizes of the resulting clusters are [34560, 11519, 8640, 5760, 4320, 2160, 2160] which suggests that the first cluster is significantly bigger than the others and the final 3-4 clusters can be treated as outliers. Nevertheless we may have some interesting information to extract from all of them.
By visualizing the cluster centers using Multi-Dimensional Analysis we can arrive to the same conclusion as we observe that the first three are close together while the rest are further apart.

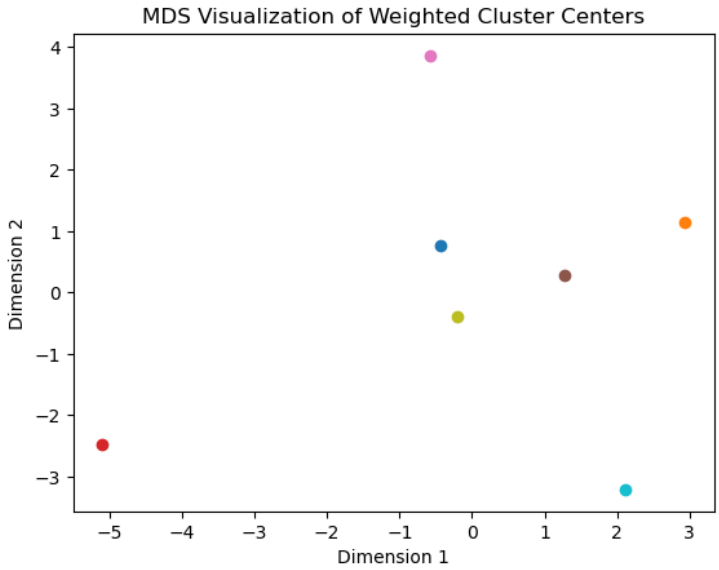


Fig10. Cluster centers visualization for weighted kMeans on angles dataset

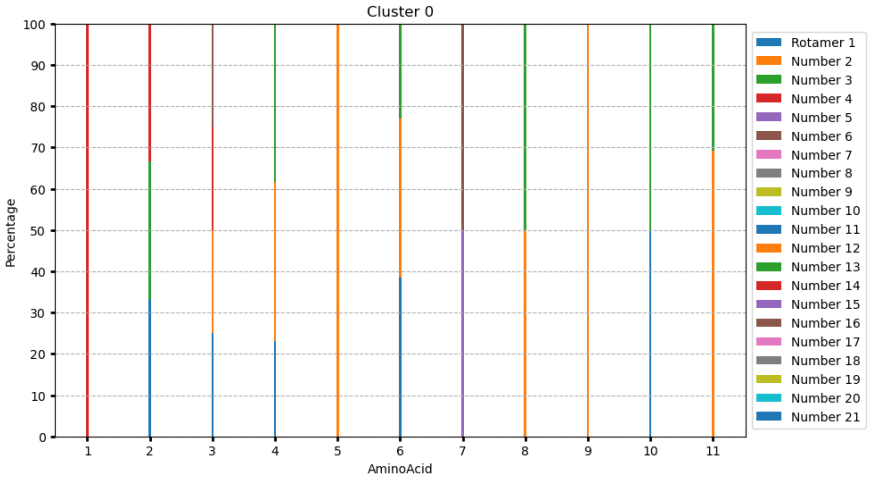
Upon analyzing the consistency of each cluster, we observed that certain clusters were predominantly characterized by specific amino acid rotamers. In other words, these clusters exhibited a significant prevalence of particular conformations associated with certain amino

acids. This finding suggests that certain amino acid rotamers have a higher likelihood of belonging to specific clusters, indicating a potential correlation between the amino acid composition and the cluster assignments.

More specifically, we have identified specific rotamers that dominate certain clusters, and we have summarized these findings in the following table and accompanying graphs.

Cluster #	Amino Acid: Rotamer
0	GlU88: 4
3	GlU88: 1 ASN284: 2
6	ASN282: 3

Table1. Dominant amino acid rotamers in clusters of weighted kMeans (angles)



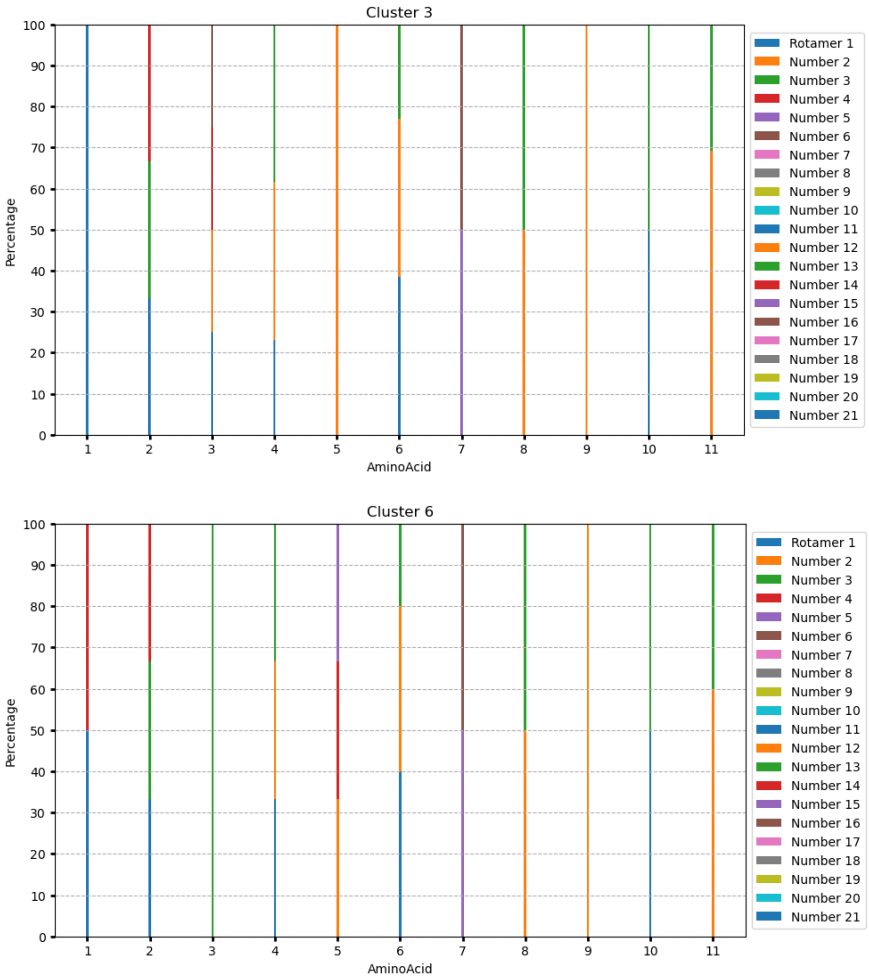


Fig11. Clusters 0,3 and 6 for weighted kMeans on angles dataset

In the following pictures we can see the two rotamers of GLU and ASN

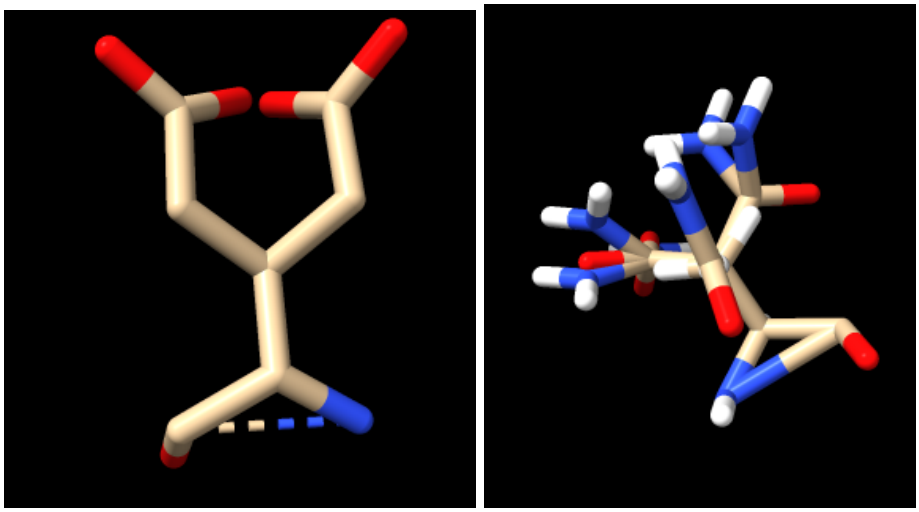


Fig12.rotamers of GLU and ASN

The observation of a significant dominance of either of the two rotamers of glu88 in clusters 0 and 3 is intriguing because, based on their frequency of appearance, they are expected to occur with similar probabilities. This indicates that there might be other factors influencing the clustering of these rotamers in specific clusters. Exploring the structural or functional implications associated with these rotamers could provide valuable insights into their preferential associations with certain clusters.

Upon analyzing the plots, we notice that all the clusters consistently contain the second rotamers of his341. This finding aligns with our prior knowledge of the rotamer's frequency of appearance, which exhibits a 100% probability for the second rotamer. Hence, the results from the plots confirm that his341 predominantly adopts the second rotamer across all clusters.

The numbers mentioned in the plot are the sequence of aminoacids of the active site of glycogen phosphorylase so 1:GLU, 2:LEU, 3:ASN, 4:ASP, 5:ASN, 6:PHE, 7:ARG, 8:ASP, 9:HIS, 10:HIS, 11:THR

We can further visualize our results by utilizing the t-SNE and UMAP dimensionality reduction techniques on the reduced data, along with the labels generated by the weighted k-means algorithm. This visualization provides a clearer understanding of the clustering patterns and how the rotamers are distributed within each cluster.

Fig13. Weighted clustering results visualized on t-sne reduced data (angles)

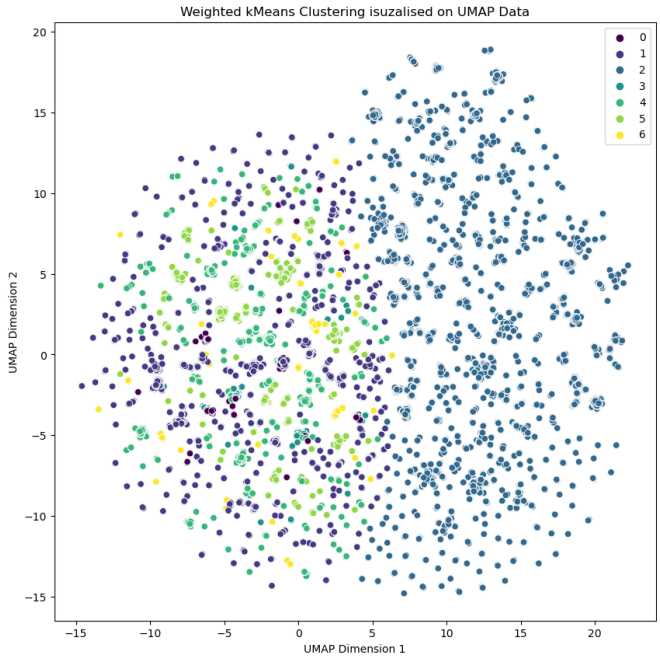


Fig14. Weighted clustering results visualized on umap reduced data (angles)

Upon examining the visualization of the reduced data using t-SNE and UMAP, we can observe that there is one prominent cluster that appears distinct and separate from the other clusters. However, the remaining clusters exhibit a degree of intermixing, indicating that their boundaries are less clearly defined.

Unweighted Clustering

In addition to the weighted k-means clustering, we also conducted unweighted k-means clustering to compare the results of the two methods. Through the elbow method, we determined that the optimal number of clusters for the unweighted k-means clustering is also 8.

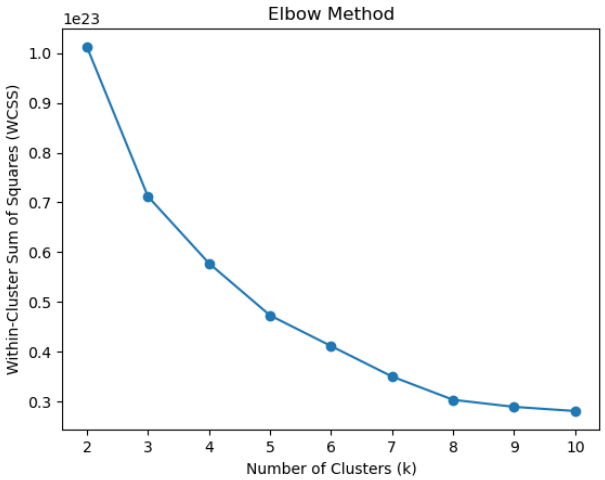


Fig15.Elbow method for unweighted kMeans on angles dataset

The sizes of the resulting clusters are 11520, 11519, 8640, 8640, 8640, 8640, 5760, 5760 which means that they cannot be clearly distinguished as main clusters and outliers although we can observe three distinct sizes of around (11000,8600,5700) points.
Upon visualizing the cluster centers using MDA we can arrive at the same conclusion being that there are no distinct outliers present.

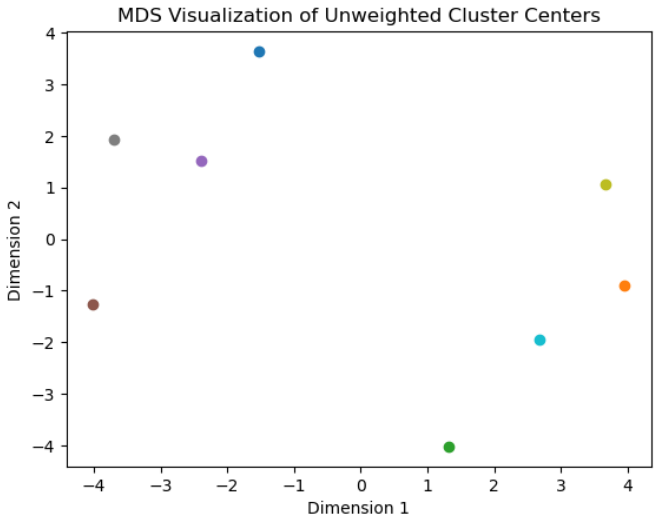


Fig16.Cluster centers visualization for unweighted kMeans on angles dataset

After performing the same cluster analysis we obtained the following results for the dominant rotamers of each cluster.

Cluster #	Amino Acid: Rotamer
0	PHE285: 3
2	ASP283: 5
4	ASP283: 5

Table2. Dominant amino acid rotamers in clusters of unweighted kMeans (angles)

In the following picture we can see the rotamers of ASP and PHE

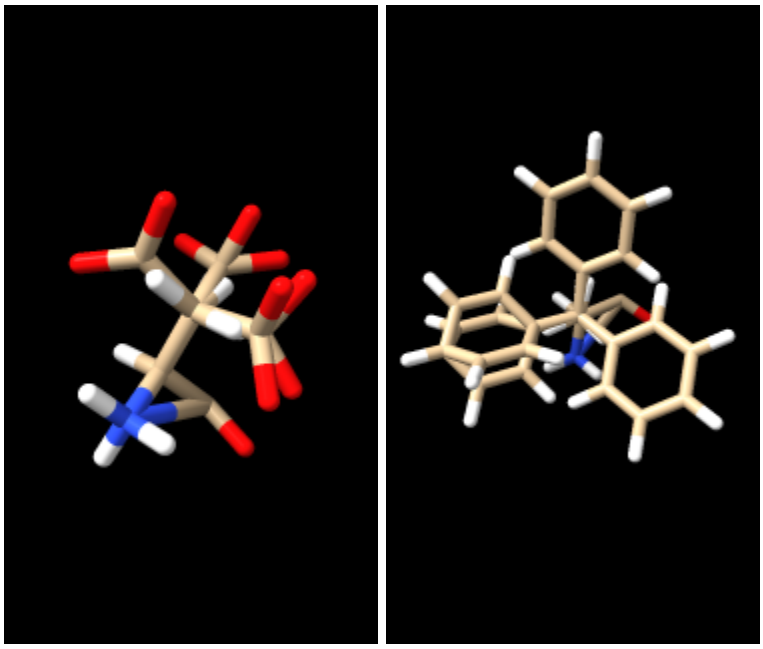


Fig17. Rotamers of ASP and PHE

We can also visualize the consistency of the above mentioned clusters.

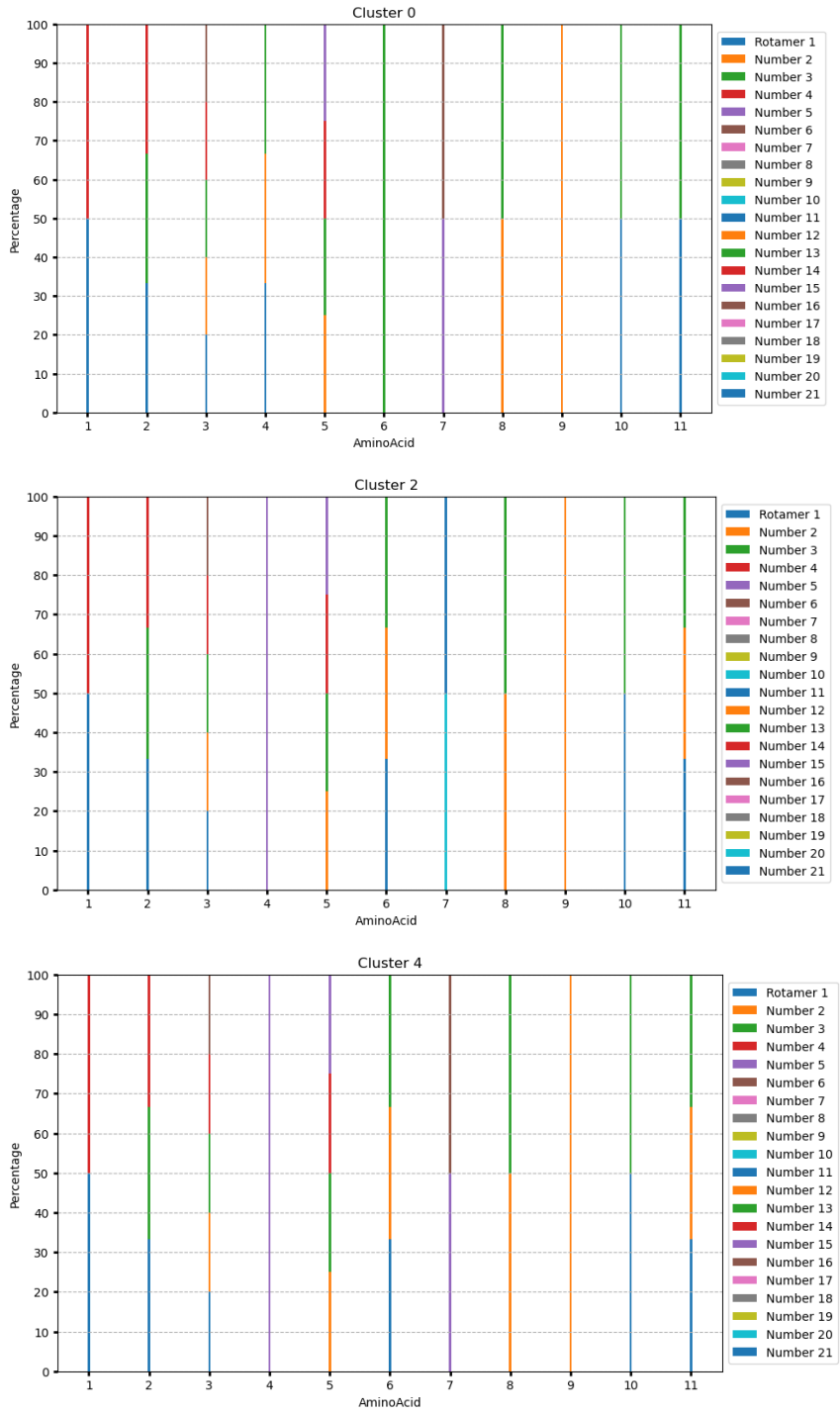


Fig18. Clusters 0,2 and 4 for unweighted kMeans on angles dataset

We also observe that in both cases the rotamers of amino acid ARG292 appear in specific couples, either 15-16 or 20-21 together in each cluster.

An interesting observation is that the rotamer 5 of ASP283 dominates both cluster 2 and 4 in the unweighted k-means clustering, despite its low frequency of appearance at only 6.82%. This result highlights the potential impact of not weighting the rotamers based on their frequency of occurrence. In the weighted approach, such dominance might have been minimized or accounted for, providing a different perspective on the clustering results.

We can again visualize our results using UMAP and t-SNE.

Fig19. Weighted clustering results visualized on t-sne reduced data (angles)

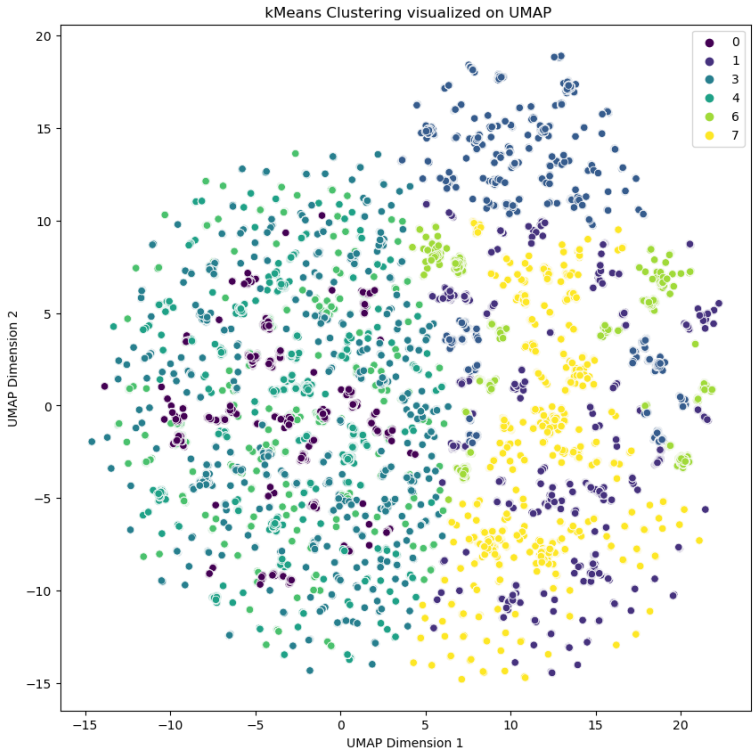


Fig20. Weighted clustering results visualized on umap reduced data (angles)

Upon analyzing both the UMAP and t-SNE diagrams of the coordinates dataset, we observed that the clusters appeared more distorted compared to the weighted clustering case. This observation indicates that the spatial arrangement of the data points based on the 26 coordinate angles does not exhibit clear and well-defined clusters. The distortion in the clusters suggests that relying solely on the individual angles may not be sufficient to accurately distinguish and separate the data points into distinct groups. To achieve more reliable and meaningful cluster formation in the dataset, it would be beneficial to consider additional factors or features in conjunction with the coordinate angles.

Coordinates Dataset

The coordinates dataset consists of 69,119 rotamers, each characterized by 163 coordinate angles. Similar to the angle dataset, we conducted a comprehensive analysis to gain insights into the structural properties of the rotamers. The results obtained from the coordinate dataset are presented below, providing valuable information on the distribution, clustering, and potential functional implications of the rotamers based on their spatial coordinates.

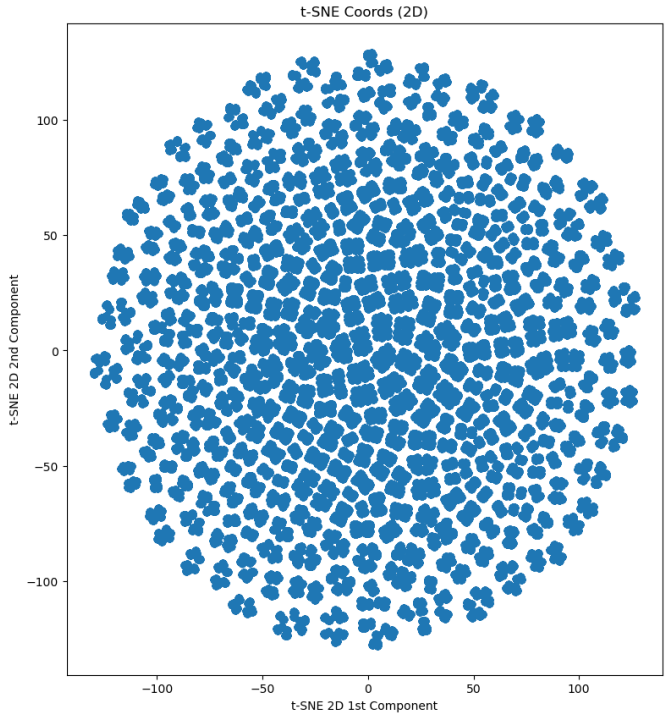


Fig21. t-SNE on coords dataset

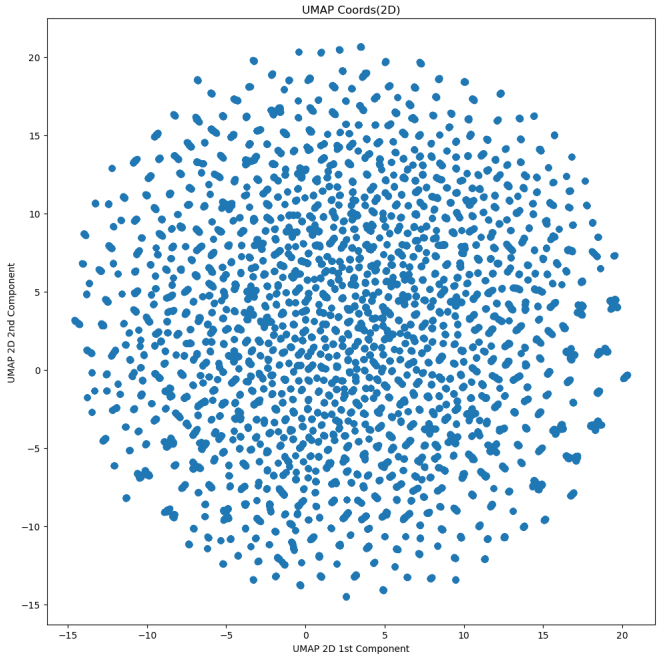


Fig22. UMAP on coords dataset

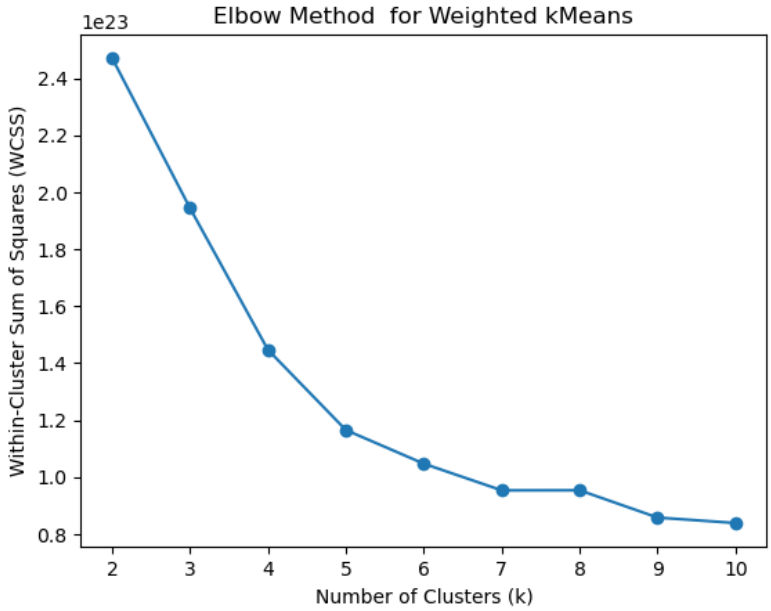


Fig23. Elbow method for weighted kMeans on coords dataset

The resulting sizes of the clusters were 46079, 11520, 5760, 4680, 1080 which suggest that we can treat the last one to three clusters as outliers. By visualizing the cluster centers we observe that the three of them are similar to each other while the rest two are dissimilar. This could suggest that the last two are outliers. But it can also suggest that our clusters are not well distinguished in terms of their contents.

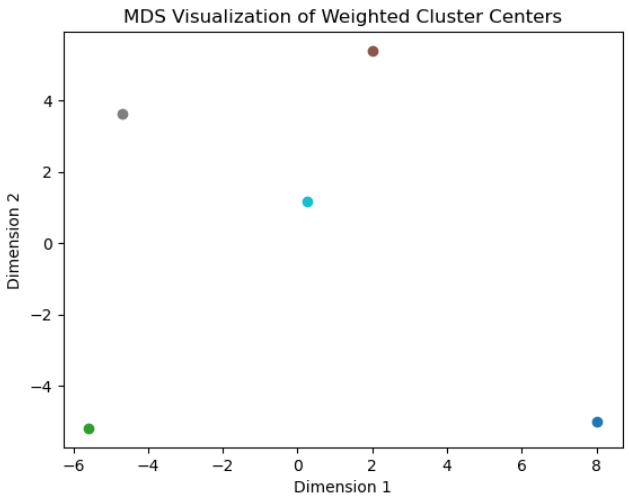


Fig24.Cluster centers visualization for weighted kMeans on coords dataset

Cluster #	Amino Acid: Rotamer
1	PHE285: 3
2	ASN284: 2
3	PHE285: 3
4	PHE285: 3 GlU88: 1

Table3. Dominant amino acid rotamers in clusters of weighted kMeans (coords)

We can visualize the rotamers of PHE, ASN and GLU in the following pictures

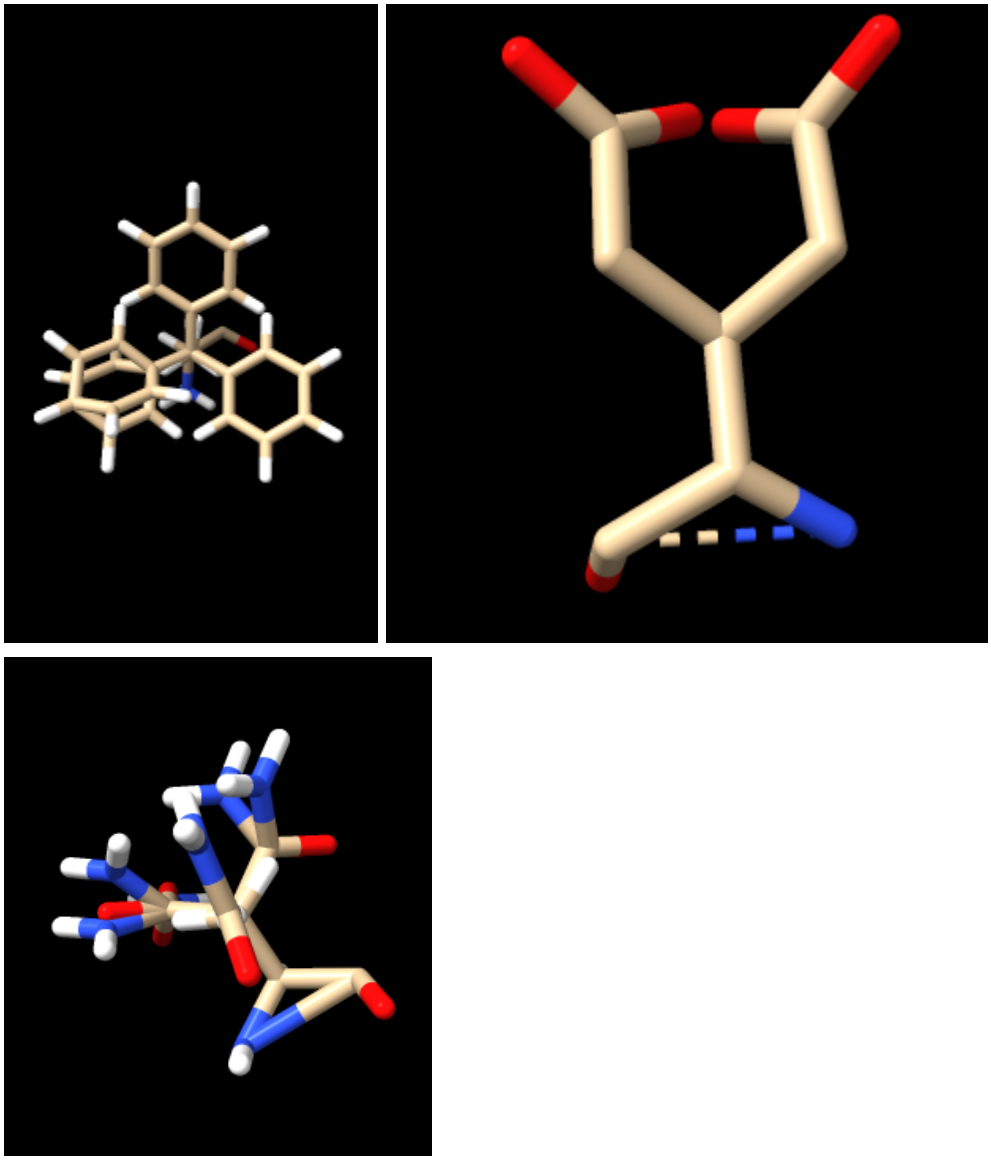


Fig25. Rotamers of PHE, ASN and GLU

We can also visualize the consistency of the above mentioned clusters in the following pictures.

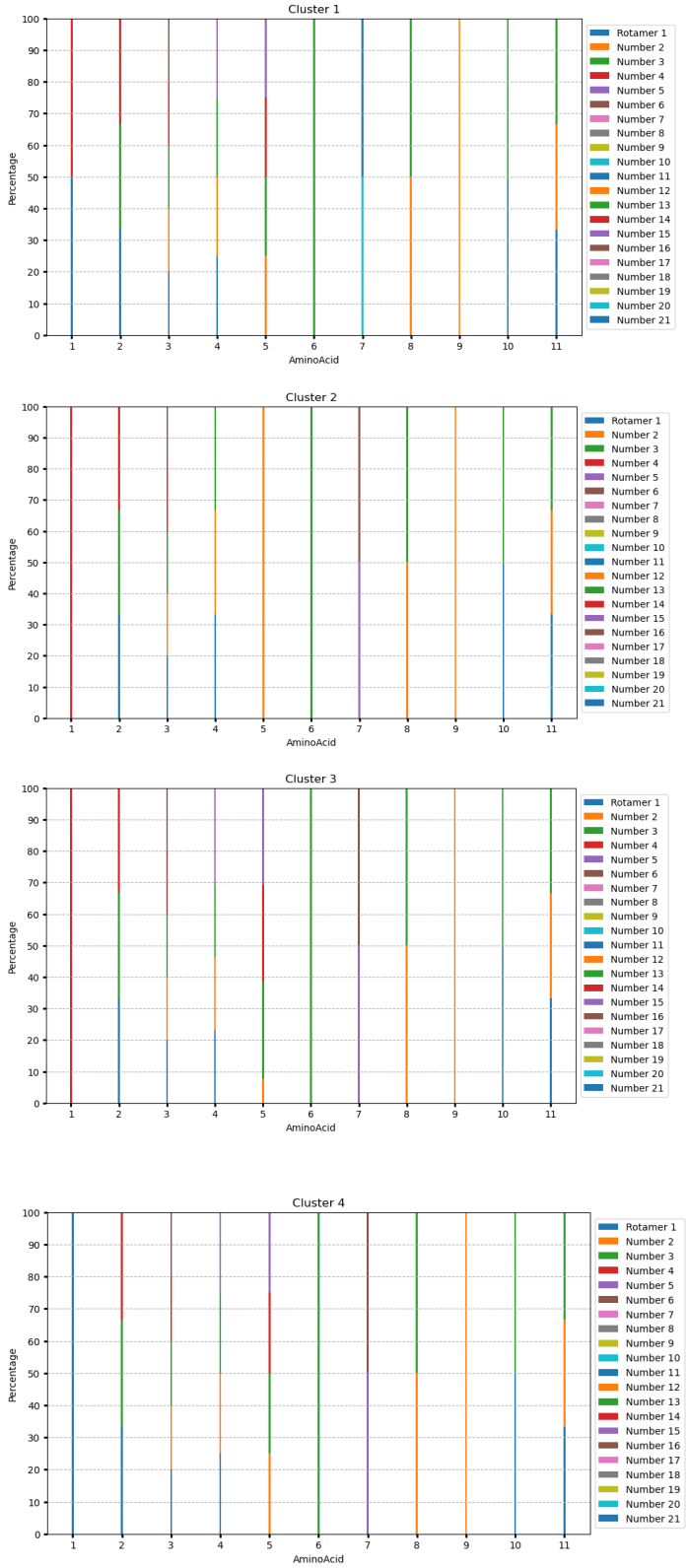


Fig26. Clusters 1,2, 3 and 4 for weighted kMeans on angles dataset

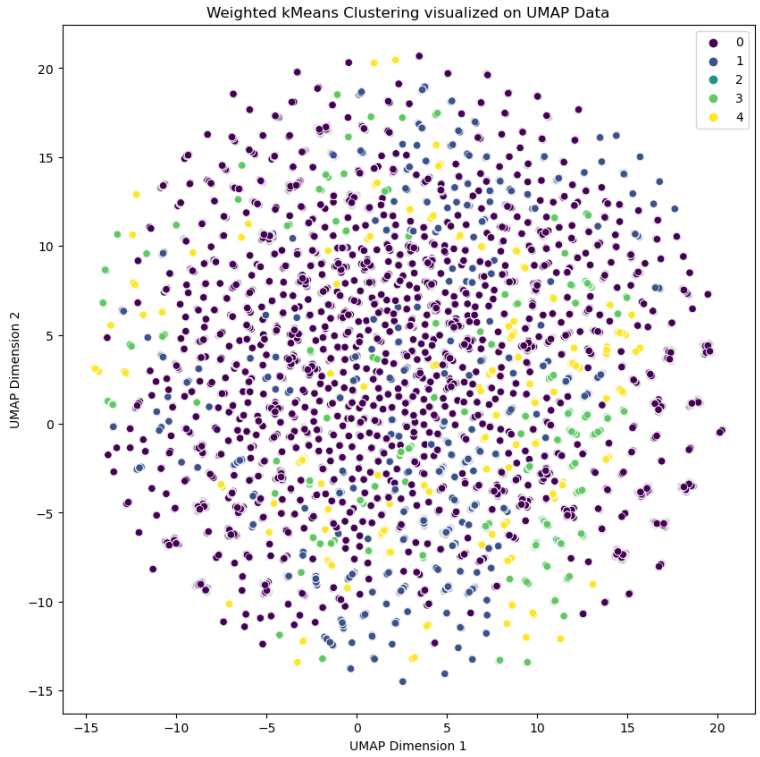


Fig27. Weighted clustering results visualized on umap reduced data (coords)

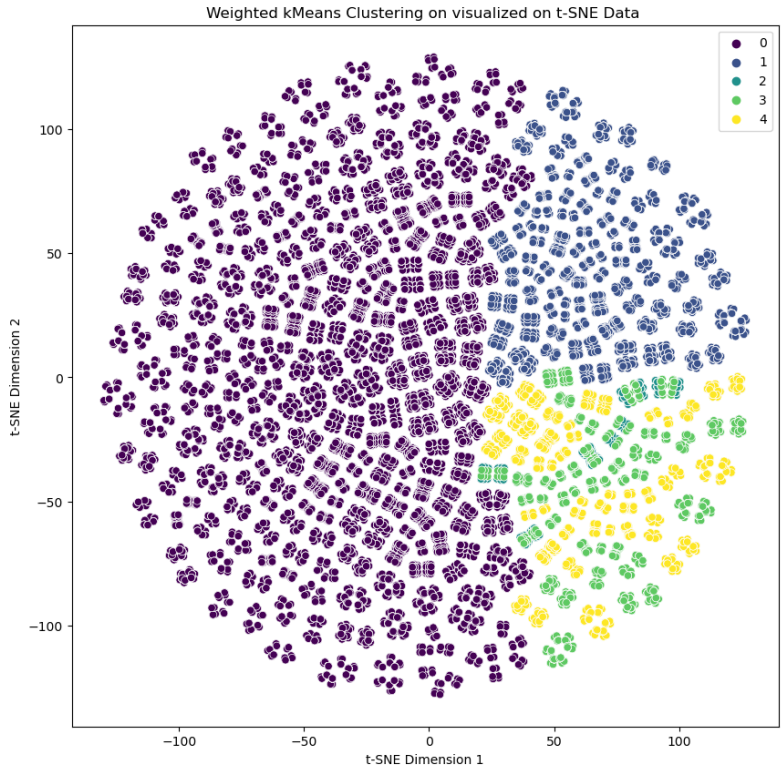


Fig28. Weighted clustering results visualized on t-sne reduced data (coords)

Unweighted Clustering

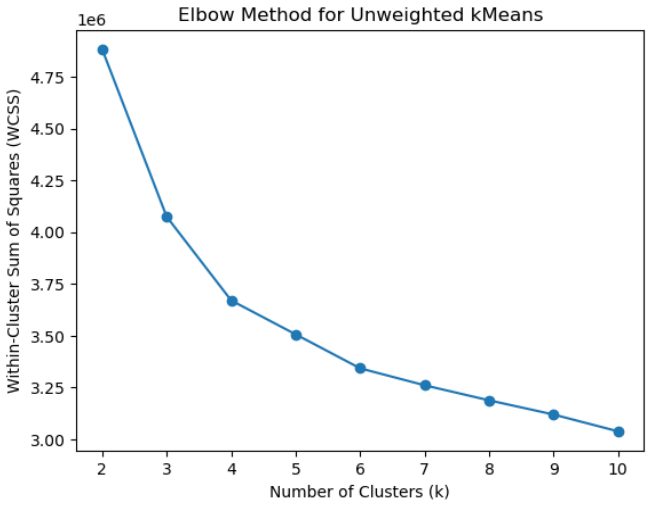


Fig29. Elbow method for unweighted kMeans on coords dataset

The resulting sizes of the clusters were 23040, 23039, 11520, 11520 which suggest that we cannot treat any of them as an outlier. By visualizing the cluster centers we observe four distinct clusters.

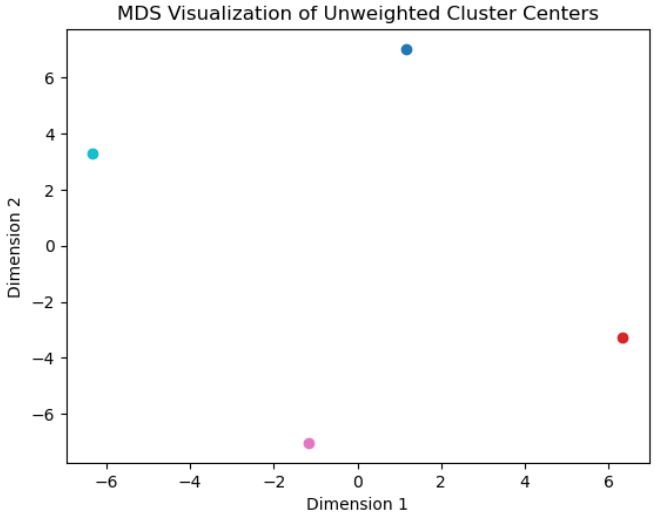


Fig30.Cluster centers visualization for weighted kMeans on coords dataset

Cluster #	Amino Acid: Rotamer
1	PHE285: 3
2	PHE285: 3

Table4. Dominant amino acid rotamers in clusters of unweighted kMeans (coords)

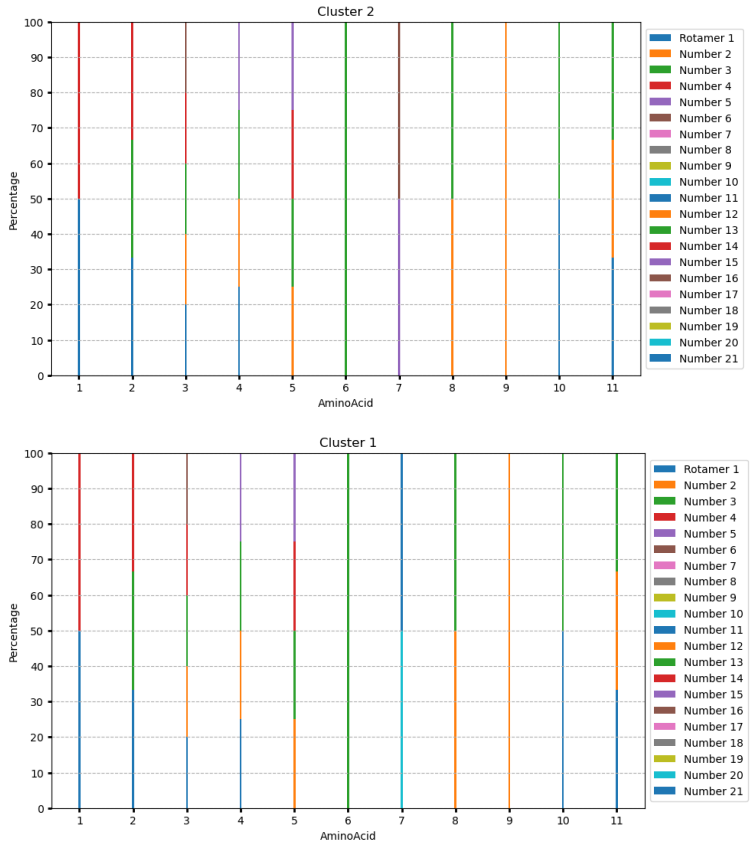


Fig31. Clusters 1,2 for weighted kMeans on angles dataset

One interesting finding in our analysis is the dominance of rotamer 3 for PHE285 across almost all clusters. This observation can be explained by the high percentage of occurrence of this particular rotamer, which is approximately 92%. The prevalence of rotamer 3 suggests that it plays a crucial role in the conformational preferences of PHE285 and may have functional significance in the binding site. Further investigation is warranted to explore the specific interactions and structural implications associated with this prevalent rotamer in different clusters.

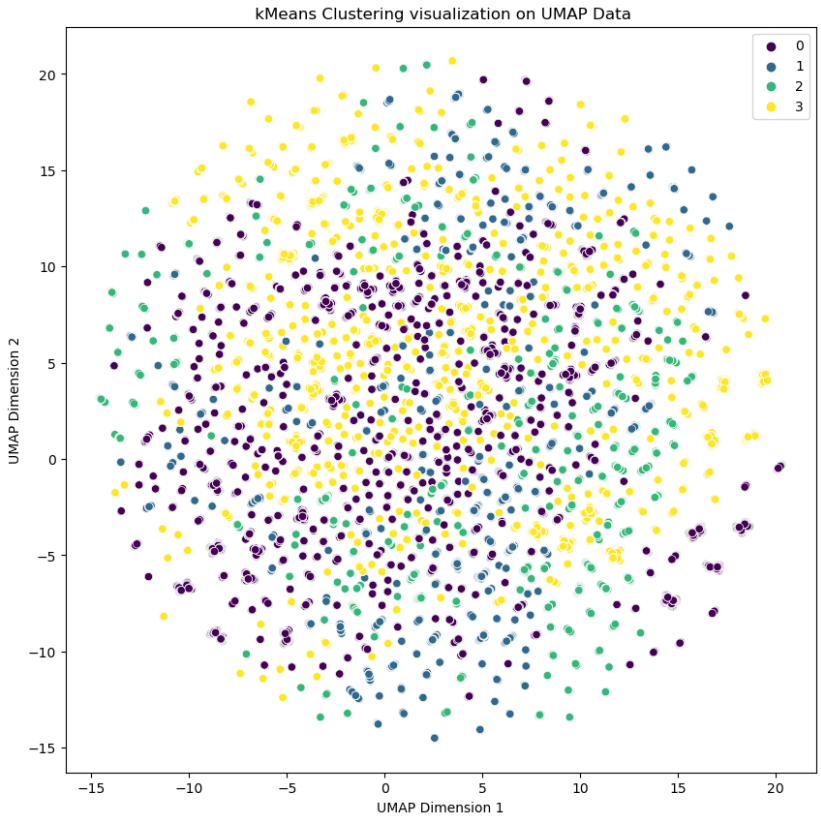


Fig32. Unweighted clustering results visualized on umap reduced data (coords)

After applying UMAP embedding to our data and utilizing the cluster labels obtained from the full-dimensional data, we find that the assigned points within each cluster appear to be intermixed, lacking a clear and distinct cluster formation. However, surprisingly, when employing t-SNE visualization, we observe a remarkable symmetry and the presence of four completely separate clusters. This discrepancy suggests that different dimensionality reduction techniques may yield varying results in terms of cluster separability and the ability to capture underlying patterns in the data.

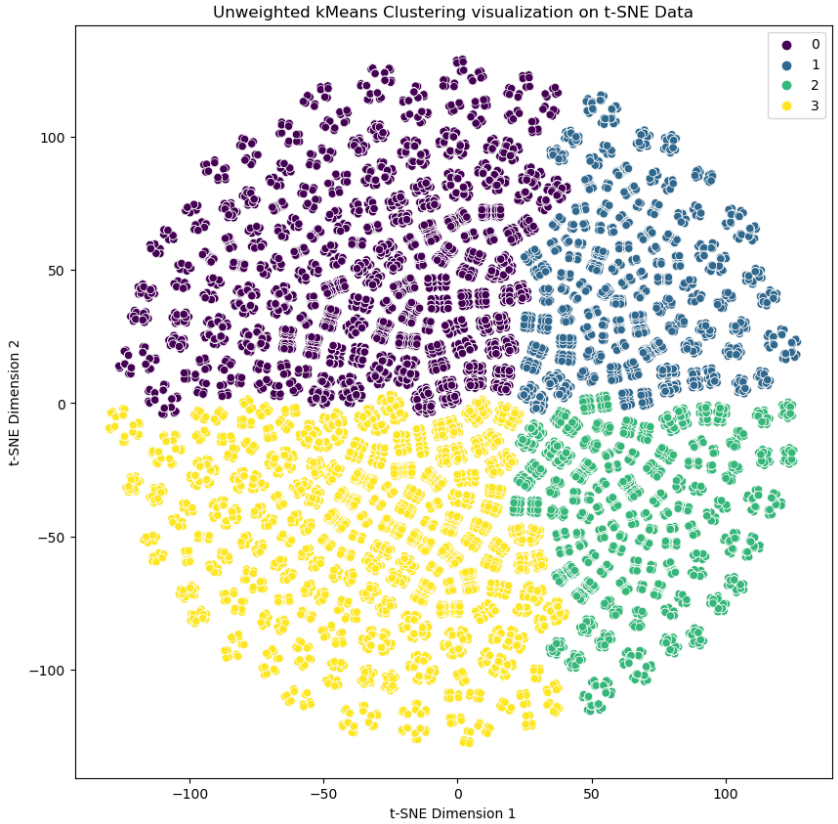


Fig33. Unweighted clustering results visualized on t-SNE reduced data (coords)

Conclusions

In our comparative analysis between the two active site prediction tools, P2Rank and DeepPocket, we found that P2Rank emerged as a better choice. P2Rank exhibited several advantages over DeepPocket, including its accessibility and superior performance in terms of result quality. P2Rank was more easily accessible, making it more convenient for researchers to use. Additionally, the results obtained from P2Rank demonstrated higher accuracy and provided valuable insights into the structural characteristics of the binding sites, including the identification of dominant rotamers within the clusters. These findings suggest that P2Rank is a reliable and effective tool for predicting protein binding sites and can serve as a valuable resource for further research and drug development efforts.

The analysis revealed several key findings. First, when attempting to analyze an active site with a large number of amino acids, computational limitations were encountered due to the high computational power required. This highlights the need for efficient resource management or alternative approaches for studying complex active sites.

The comparison of weighted and unweighted clustering approaches using k-means showcased differences in cluster formation and composition. Weighted clustering, which considered the frequency of appearance of rotamers, provided more meaningful insights into the dominant rotamers within each cluster. This emphasizes the importance of considering rotamer frequencies when analyzing clusters.

The visualization of the clustering results using t-SNE and UMAP, two dimensionality reduction techniques, allowed for a better understanding of the structural characteristics of the data. Both methods captured distinct patterns and clusters, with t-SNE revealing an underlying symmetry. These visualizations provided valuable insights into the organization of the data.

Additionally, our analysis of dominant rotamers within each cluster revealed interesting findings. Notably, certain rotamers exhibited a significant presence across multiple clusters, indicating their high frequency of occurrence. For example, Rotamer 3 of PHE285 emerged as a dominant rotamer in several clusters, suggesting its prevalence in the dataset.

Furthermore, we observed specific patterns in the amino acid ARG292, which appeared in distinct pairs (15-16 or 20-21) within each cluster. This observation implies the existence of preferential associations between specific rotamers of ARG292, potentially indicating structural motifs or functional relationships within the protein.

It is worth noting that the rotamers of GLU, ASN, and PHE also played an important role in the weighted clustering analysis of both the angle and coordinate datasets. This finding is particularly noteworthy as it suggests that the influence of these specific rotamers is not limited to a particular dataset, indicating their broader significance and potential functional relevance in the context of the protein's active site.

Overall, the analysis of the coordinates and angles of the conformations of the active site, along with the application of clustering, dimensionality reduction, and examination of dominant rotamers, contributed to a better understanding of the structural characteristics and clustering patterns within the dataset. These findings have implications for enzyme active site research, including potential applications in drug design and therapeutic interventions.

Thank you

I would like to express my sincere gratitude to my professors, Ioanis Emiris and Evangelia Chrysina, for their invaluable insights, guidance, and advice throughout the research and execution of this analysis. Their expertise and support greatly contributed to the success of this project. I would also like to extend my appreciation to my colleague, Andreas Zamanos, for his assistance in modifying the code of Deep Pocket. His collaboration was instrumental in overcoming the challenges we encountered. Furthermore, I am deeply thankful to Kostantina Roka for her contribution on the creation of the code and her dedicated supervision. Her guidance and valuable advice played a crucial role in the development of this study. Their collective efforts and support have been instrumental in the completion of this research. Thank you.

References

https://www.sciencedirect.com/science/article/pii/S0969212601006797?via%3Dihub
https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P49841/annotations
https://github.com/devalab/DeepPocket
https://prankweb.cz/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4340754/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2697722/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4775866/
https://towardsdatascience.com/using-weighted-k-means-clustering-to-determine-distribution-centres-locations-2567646fc31d
https://medium.com/@dey.mallika/unsupervised-learning-with-weighted-k-means-3828b708d75d
https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
https://uu.diva-portal.org/smash/get/diva2:1503863/FULLTEXT01.pdf
https://umap-learn.readthedocs.io/en/latest/