Task 12:
Principal Component Analysis on a Torus
Classical Complex Systems WiSe 2023/24

Karthik Jayadevan
(Matriculation Number: 5582876)

February 23, 2024

1 Introduction

Proteins are fundamentally composed of monomeric units called amino acids. Each of these amino acids possesses a standard *backbone* structure that facilitates the linkage of one amino acid to another in the sequence. These linkages are referred to as *peptide bonds*.

Backbone dihedral angles, (ϕ and ψ), play a key role in the structural configuration of proteins (where the word *di-hedral* accounts for the fact that the angles are measured between two faces/planes). ϕ is defined as the angle in the chain C'-N-C^{α}-C' (where C' denotes the carbon atom in the carboxyl group, N is Nitrogen and C^{α} denotes the alpha carbon). Similarly, ψ is the angle in the chain C'-N-C^{α}-C' [1]. The definitions of ϕ and ψ is depicted in the figure 1 [2]. The molecule that is analyzed in

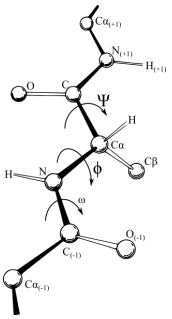


Figure 1: Depiction of dihedral angles [2]

this project is alanine dipeptide (Ac-Ala-$NHCH_3$), a peptide consisting of two alanine molecules linked by a peptide bond (see figure 2).

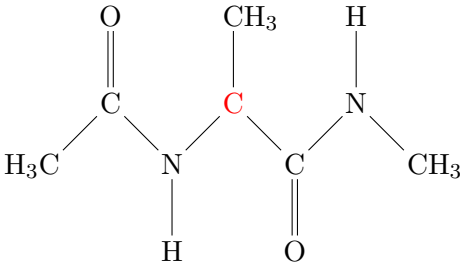


Figure 2: Line diagram of alanine dipeptide. The alpha Carbon (marked in red) is connected to two peptide bonds.

1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful technique for reducing the complexity of a high-dimensional system[3]. It involves expressing the system in a coordinate space defined by *principal components*. The original data projected along these components are linear combinations of the original variables, organized so that the projection along the first principal component (PC1) accounts for the maximum variance within the system. Each subsequent principal component captures progressively less variance than its predecessor. Hence this method is commonly used to reduce the dimensionality of a high-dimensional system like that in macromolecules.

1.2 Dihedral Principal Component Analysis (dPCA)

While PCA works well for linear data, the periodic nature of the dihedral angle data gives rise to ar-

tifacts in the calculation of PCA. The method can misinterpret the proximity of angles (for example $-180°$ and $179°$) and give misleading results. In the context of such periodic data, hence the calculation of mean, variance and covariance can also be affected (see equations 8, 9), which are central to the PCA method.

A solution to this issue was proposed in the dihedral-PCA (dPCA) method [4]. The angles are transformed using sine and cosine functions, effectively linearizing the circular data.

$$\begin{pmatrix}\phi\\ \psi\end{pmatrix} = \begin{pmatrix}\sin(\phi)\\ \cos(\phi)\\ \sin(\psi)\\ \cos(\psi)\end{pmatrix} \tag{1}$$

The data gets remapped as follows:

$$[-\pi,\pi)\times[-\pi,\pi) \longmapsto \mathbb{R}^4 \tag{2}$$

This sort of 'unwrapping' of the circular data circumvents the issues in the PCA mentioned above [5]. The detailed method is outlined in section 2.3.

1.3 Dihedral Principal Component Analysis on a Torus (dPCA+)

While the dPCA method is designed to address the problems arising from periodic data, it gives rise to complexities that make the results difficult to interpret. Since the dihedral angle space form a torus, a method preserving such a topology was introduced in [6](named dPCA+, with the '+' to indicate its superiority over dPCA). The basic idea here is to minimize the projection error caused by the periodicity of the dihedral angles. This is implemented by identifying a maximal gap in the sampling and shifting the data such that the maximal gap lays at the periodic boundary. Hence the data gets remapped as:

$$(\phi,\psi)^T \mapsto (\phi+\phi_{\text{offset}}, \psi+\psi_{\text{offset}}) \tag{3}$$

This transformed data can then be analyzed by a standard PCA.

2 Methods

2.1 Task I: Basic Data Consideration

The data given in the ASCII file contained 2.5×10^6 points in the trajectory of alanine dipeptide. The trajectory was given as two columns containing the ϕ and ψ angles respectively:

$$\phi,\psi \in [-180°, 180°[\tag{4}$$

In the task instructions, it was given that the points were separated by Δt =200 fs. Since the number of data points, say N, was 2.5×10^6, the timescale of the entire trajectory could be calculated as

$$\begin{aligned} t_{\text{final}} &= N\times\Delta t\\ &= 2.5\times10^6\times200\,\text{fs}\\ &= 2.5\times10^6\times200\times10^{-15}\text{s}\\ &= 5\times10^{-7}\text{s}\\ &= 500\,\text{ns}\end{aligned} \tag{5}$$

After importing the data using the `read_csv` function in the `pandas` package and setting the above timescale, the time evolution of the dihedral angles were plotted as shown in figure 3. Here one can see how the angles evolve on several timescales. Fast oscillatory motions were observed on the picosecond scale 3(A,B), as well as shifts between different conformational states at nanosecond scale (C,D,E and F).

The circular (and not linear) nature of the given data (equation 4) gives rise to some problems for analysis. At the stage of basic data visualization, one could already see that a quick viewing of the plot in figure 3 can be misinterpreted. For example, consider the jump from $\phi_1\approx-100°$ to $\phi_2\approx+140°$ in the panel E of the figure. For linear data, this jump would mean that the value changes by $\Delta\phi = 140-(-100) = 240$ units. However, since these are angles, the actual jump[1] is only by $110°$ (see figure 4).

A common way of visualizing dihedral angles is to make a Ramachandran plot[7], which is a two-dimensional heatmap of the dihedral angles. After converting the angles to radians, this plot was made here using the built-in `matplotlib` package `hist2d`, with 200 bins. The bins were chosen through trial to avoid noisy spikes arising from a higher bin count. A three-dimensional plot of the same was made in figure 6.

[1] A function called `angle_difference` was defined in the Jupyter notebook to calculate the actual difference between two angles.

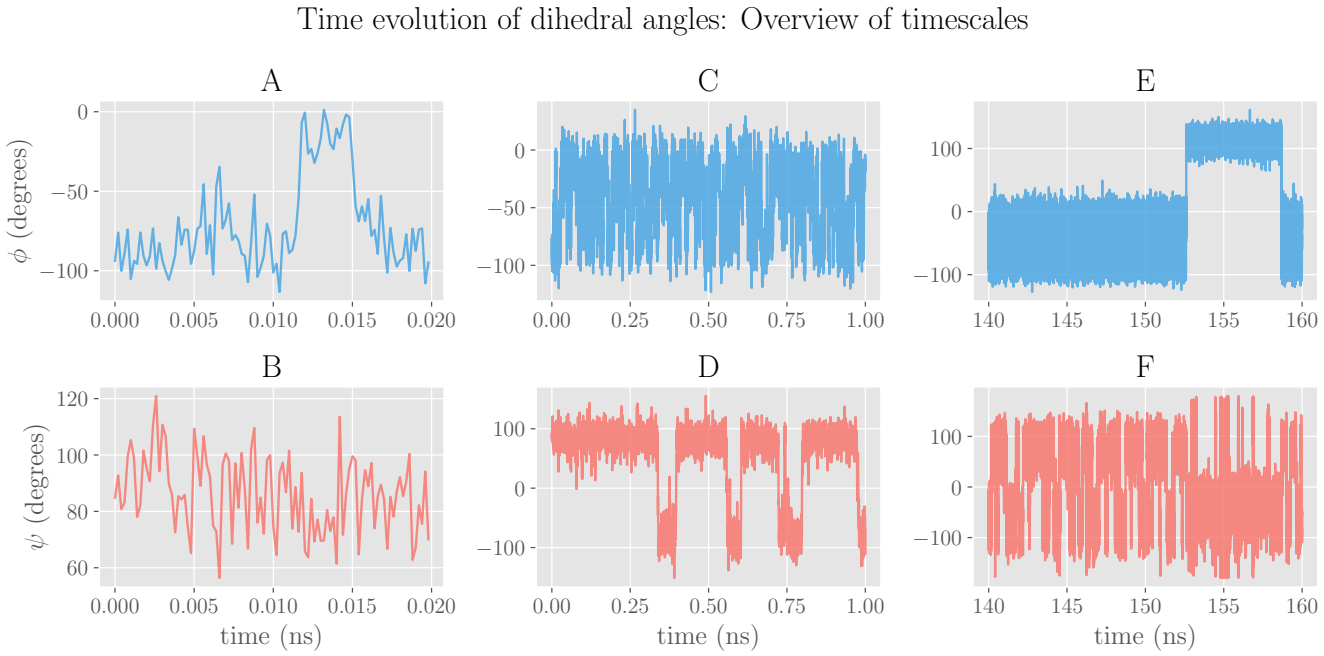


Figure 3: Time evolution of dihedral angles ϕ and ψ at different time scales: (A) and (B) in picoseconds, (C) and (D) in fractions of a nanosecond, (E) and (F) in tens of nanoseconds.

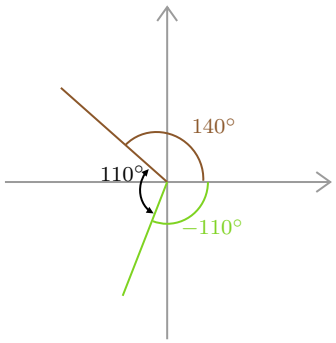


Figure 4: Example to show possible misinterpretation of circular data: If $\phi_1 = 140°$ and $\phi_2 = -110°$, it is easy to see geometrically that $\Delta\phi = 110°$.

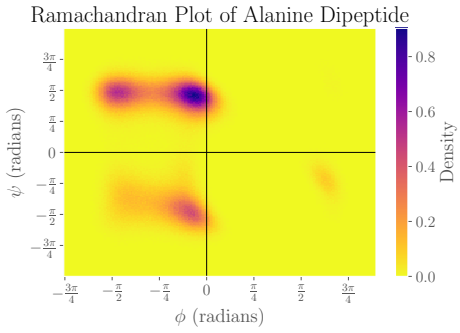


Figure 5: The Ramachandran plot of alanine dipeptide, a two-dimensional probability distribution as a function of ϕ and ψ.

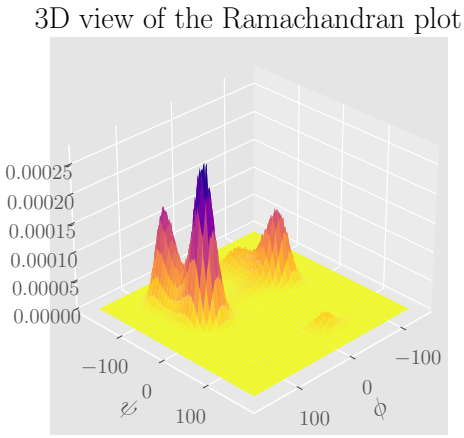


Figure 6: 3D-projection of the Ramachandran plot made in figure 5.

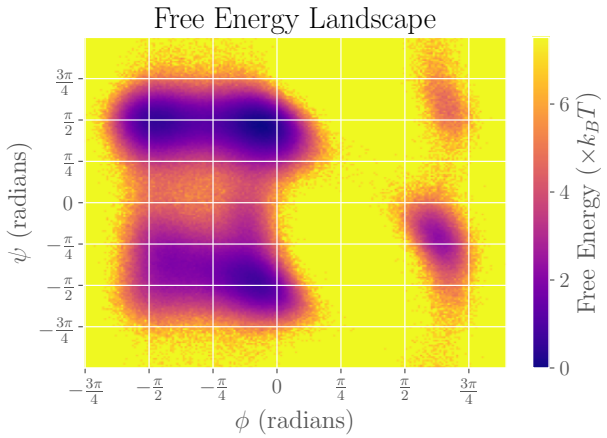


Figure 7: The plot of two-dimensional free energy given by equation 6. Setting the logarithmic scale enhances the visibility of the various regions around the minima, as compared to figure 5

Transforming the distribution to logarithmic scale improved the visibility of the different states observed in the plot (see figure 7). Hence the two-dimensional free energy $\Delta G(\phi, \psi)$ was determined as follows:

$$\Delta G(\phi, \psi) = -\ln\left(\frac{P(\phi, \psi) + \rho_{\min}}{\rho_{\max}}\right), \qquad (6)$$

where $\rho_{\min}$ and $\rho_{\max}$ correspond to the minimum and maximum values in the distribution $P(\phi, \psi)$. The free energy was hence obtained in units of $k_B T$. The function `calculate_free_energy` was defined in the notebook to compute the two-dimensional free energy given the probability distribution as a function of two coordinates. This function was used in all four tasks to make the free energy plot. A simple histogram of the dihedral angles (figure 8) indicate multiple minima positions. The `bins='auto'` parameter of `hist2d` can be used to choose the bin count automatically, which gave about 350 bins for ϕ and 150 for ψ. The minima from the 1d distributions were roughly identified on the free energy plot by annotating the plot with vertical and horizontal lines (corresponding to constant ϕ and ψ values respectively). The points of intersection rightly matched with the observed minima in the Ramachandran plot. The plot also showed that the landscape has a nature of continuity at the boundary (for example, look along the constant ϕ line near $\phi = \frac{3\pi}{4}$ in figure 9). This is another reason to consider a periodic treatment of dihedral angles.

The points of minima in the heatmap (figure 9) correspond to different stable conformational states of the dipeptide, since it spends more steps in the trajectory in these states. The distribution of the trajectory indicated that the most stable conformations were the two states in the second quadrant (ψ^+, ϕ^-), which correspond to the β sheet conformation of polypeptides [8]. The sharpest peak appeared around the point $\phi = -10°$, $\psi = 85°$.

2.2 Task II: Principal Component Analysis (PCA)

The trajectory of the dipeptide molecule were imported as two arrays:

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{pmatrix} \qquad \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{pmatrix} \qquad (7)$$

The first task in the project was to perform a regular PCA of the variables given. The covariance between the variables is defined as

$$\mathrm{Cov}(\phi, \psi) = \frac{1}{N}\sum_{i=1}^{N} (\phi - \langle\phi\rangle)(\psi - \langle\psi\rangle), \qquad (8)$$

and the variance of each variable

$$\mathrm{Var}(\xi) = \mathrm{Cov}(\xi, \xi) = \frac{1}{N}\sum_{i=1}^{N} (\xi - \langle\xi\rangle)^2 \qquad (9)$$

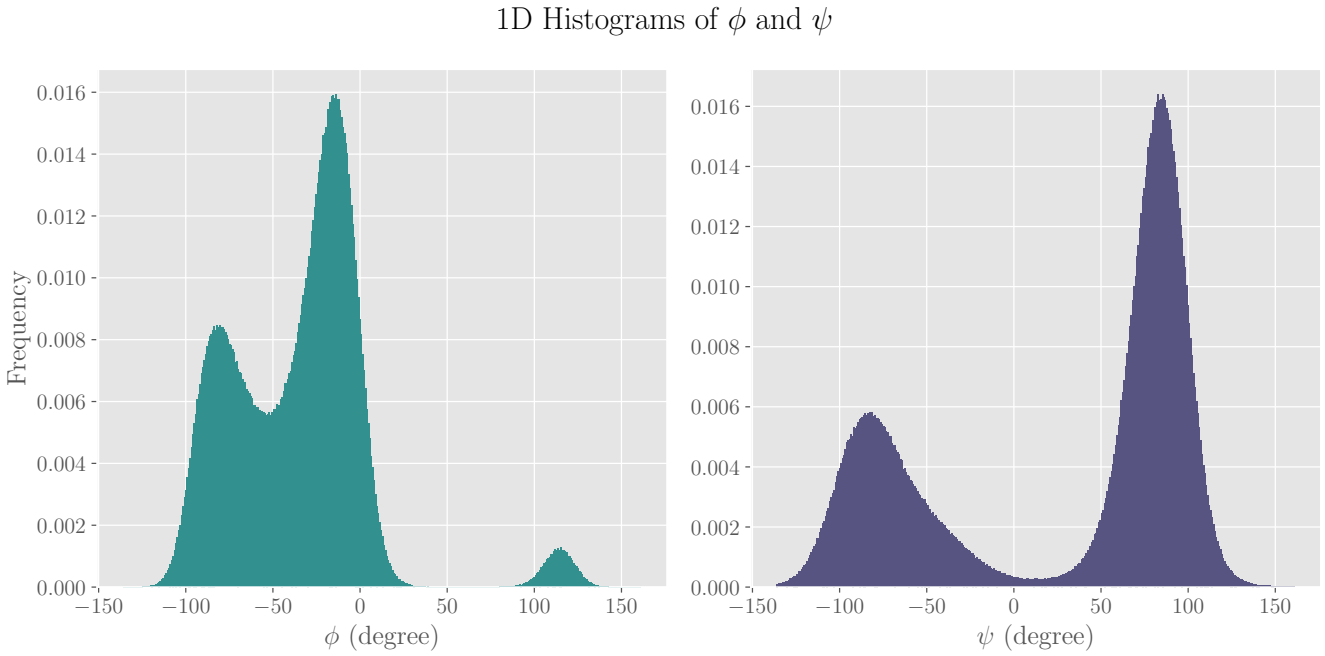


Figure 8: One-dimensional distributions of ϕ (left) and ψ (right). The maximas correspond to stable conformational states. ϕ has two major peaks around $-10°$ and $-80°$, and one minor peak at about $120°$, while the distribution of ψ has two major peaks, one around $-80°$ and another at about $85°$. The distribution is shown here in degrees to easily identify angles.

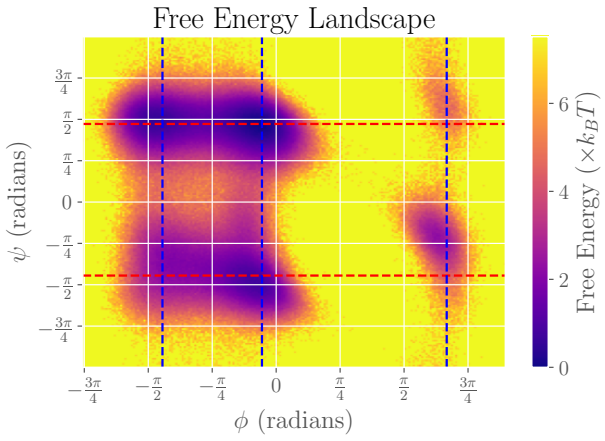


Figure 9: Free energy landscape annotated with rough estimates of minima obtained from figure 8.

with $\xi \in \{\phi, \psi\}$. The covariance matrix is defined with these two quantities as elements:

$$C = \begin{pmatrix} \mathrm{Var}(\phi) & \mathrm{Cov}(\phi,\psi) \\ \mathrm{Cov}(\phi,\psi) & \mathrm{Var}(\psi) \end{pmatrix} \tag{10}$$

The covariance matrix was computed using the `numpy` method `cov(arr1,arr2)`, where `arr1` and `arr2` are the arrays whose covariance is studied. Diagonalizing the matrix, the eigenvalues (λ_1, λ_2) and eigenvectors $\vec{e}_1$, $\vec{e}_2$ were computed. The eigenvector corresponding to the higher eigenvalue of the two is the principal component 1 (PC 1), a unit vector. After centering the data by subtracting the mean value, the data points were projected along the principal components, by taking a dot-product.

$$(\vec{e}_0, \vec{e}_1)^T \cdot (\phi(t), \psi(t))^T = (V_1(t), V_2(t))^T \tag{11}$$

Finally, the free energy and the one-dimensional distributions of the data projected (labeled V_1 and V_2) were plotted (figures 15 and 16 respectively).

2.3 Task III: dPCA

To linearize the (periodic) data, the sine and cosine transformations of the dihedral angles were computed, thereby getting four arrays. The PCA (or more precisely, dPCA) was performed on the four new variables $\sin\phi, \cos\phi, \sin\psi$ and $\cos\psi$. The arrays of these four variables were stacked into a single variable `stacked_arr` (of order $4 \times N$) to compute the covariance matrix. The covariance matrix (equation 10) in this case was a 4×4 matrix, compared to the PCA, where the matrix had order 2×2. The elements of this covariance matrix can be explicitly written down as in equation 12.

The four obtained eigenvectors were arranged in decreasing order, and the principal components were identified. Following the guidelines of the given task, the data was projected only to the first two principal components. A scree plot, the free energy and one-dimensional distributions were plotted for the resulting components.

2.4 Task IV: dPCA+

As given in the task description, the maximal gap in the histogram were defined by minimizing the point density of a corridor as follows. For a bin size of 150, the two-dimensional histogram of the free energy along both dihedral angles is plotted. This plot was chosen over the regular Ramachandran plot because the latter only showed points of high concentration, so the choice of a maximal gap would be too broad to choose from. The maximal gap was identified in three stages:

1. As a starting point, a visual estimate of a potential cut point for ϕ was made. A line at the estimated point was plotted in the 2D free energy plot (vertical line corresponding to constant ϕ value, and horizontal line corresponding to constant ψ).

2. This cut point was fine-tuned by following a kind of bisection method successively. For example, if one low point-density corridor was identified between ϕ-values $\pi/4$ and $\pi/2$, a constant ϕ line was made at their mean position, *i.e.,* at $\phi = 3\pi/8$. The initial guessed line was then shifted to this position to see if it could be further improved. After two iterations, a ϕ value of $11\pi/32$ was finalized to be the first guess (figure 13). The estimated positions were also visualized in the one-dimensional projections (Note that at this point, this value is still a *guess*).

3. To fine-tune the above choice, a function `find_best_maximal_gap` was defined to perform a search for the bins with the least number of counts in the neighbourhood of the guessed value. On inputting the histogram data, this function identifies the best bin to shift the data. The numpy function `searchsorted` is utilized for this purpose[2]. The values returned by this search function is finalized for further analysis.

The initial guess and optimized values (called ϕ_0 and ψ_0 from here onward) are shown in the free energy plot in figure 11. The angles were shifted by defining the function `shift_angles`. Hence the range of the data was transformed as equation 13. Using the new shifted distributions of ϕ and ψ, a PCA is performed with these variables, and associated plots made, just as in the previous tasks.

3 Results

3.1 PCA

A scree-plot of the eigenvalues shows the percentage of total variance explained by each PC in the direct PCA analysis of the dihedral angles. The plot obtained (figure 14) indicates that the first PC explains about 80% of the total variance in the system. The two-dimensional free energy and one-dimensional distributions of V_1 and V_2 were also plotted (figures 15 and 16). The distribution along V_1 shows two peaks, suggesting that projecting onto V_1 captured two major conformations or states of the dipeptide. However, the broader and multi-peaked distribution along V_2 indicated additional dynamics that V_1 alone does not signify.

By projecting only onto the first principal component (V_1), possibly the transitional states between the major conformations that are captured by V_2 are

[2]The `searchsorted` function here takes a sorted array (`bin_edges` in this case) and a value (`guess_value`) and returns the index at which this value should be inserted to maintain the order of the array.

$$C_{\text{dPCA}} = \begin{pmatrix} \text{Var}(\sin(\phi)) & \text{Cov}(\sin(\phi), \cos(\phi)) & \text{Cov}(\sin(\phi), \sin(\psi)) & \text{Cov}(\sin(\phi), \cos(\psi)) \\ \text{Cov}(\cos(\phi), \sin(\phi)) & \text{Var}(\cos(\phi)) & \text{Cov}(\cos(\phi), \sin(\psi)) & \text{Cov}(\cos(\phi), \cos(\psi)) \\ \text{Cov}(\sin(\psi), \sin(\phi)) & \text{Cov}(\sin(\psi), \cos(\phi)) & \text{Var}(\sin(\psi)) & \text{Cov}(\sin(\psi), \cos(\psi)) \\ \text{Cov}(\cos(\psi), \sin(\phi)) & \text{Cov}(\cos(\psi), \cos(\phi)) & \text{Cov}(\cos(\psi), \sin(\psi)) & \text{Var}(\cos(\psi)) \end{pmatrix} \tag{12}$$

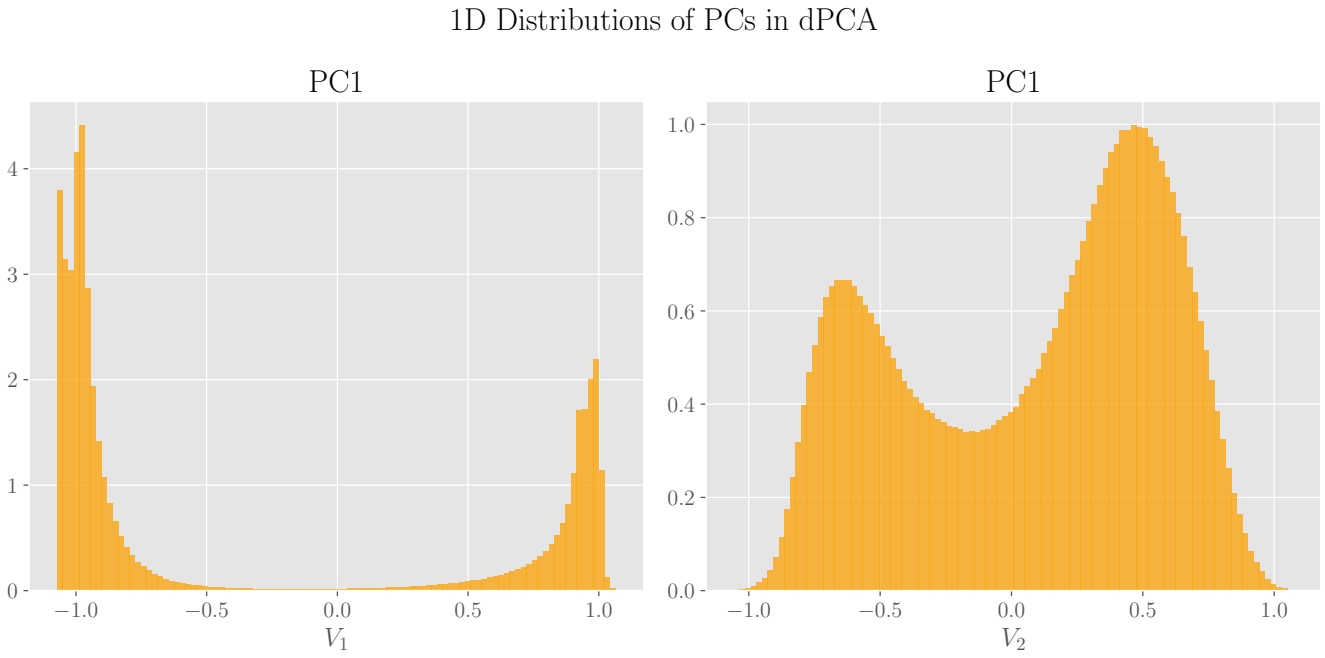


Figure 10: The one-dimensional projections of the first two PCs obtained as a result of dPCA.

$$[-\pi, \pi) \times [-\pi, \pi) \mapsto [-\pi + \phi_{\text{offset}}, \pi + \phi_{\text{offset}}) \times [-\pi + \psi_{\text{offset}}, \pi + \psi_{\text{offset}}) \tag{13}$$

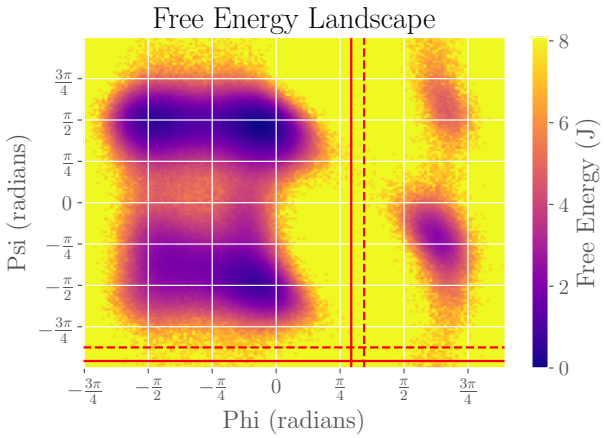


Figure 11: Identifying the maximal gap window. After an initial visual guess (dotted lines), the `find_best_maximal_gap` function optimizes this choice by a search (optimized position in continuous red line). The final cut points are $\phi_0 \approx 52.8°$ and $\psi_0 \approx -172.3°$.

neglected. These could represent less stable, but significant, intermediate states of the dipeptide's motion or conformational changes that contribute to the overall flexibility and function of the molecule.

3.2 dPCA

The scree plot of the dPCA looked more distributed (figure 17). V_1 explained a lesser percentage of variance compared to that in PCA. From the plots of the free energy (18) and the one-dimensional projections of the data (figure 19), it can be inferred that the dPCA method seeks to maintain the periodicity inherent in the data. However, this came at a cost of a lack of interpretability of the results. The plot showed a certain level of sharpness in the minima. The dPCA is better suited for capturing the periodic nature of the data, which is inherent in dihedral angles, while PCA might introduce miscalculations since it assumes linearity in the data. Although dPCA attempts to capture the cyclic nature of the dihedral angle data, it introduced new complexity in interpreting the landscape [4].

3.3 dPCA+

Figure 20 shows the variation of the two-dimensional free energy resulting from the dPCA+ analysis. In the original data plot 7, it can be seen that the heatmap gets cut off abruptly at the periodic boundary, for example in the first quadrant. These kinds of abrupt cuts were improved after shifting the data along the maximal gap.

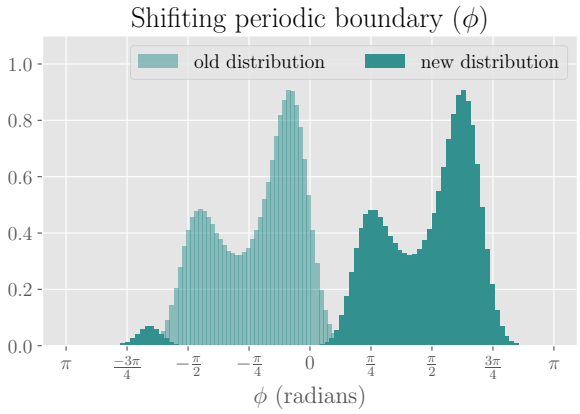
4 Summary and Discussion

This project focused on the analysis of an alanine dipeptide simulation data using the methods of PCA, dPCA, and dPCA+. After an overview on the nature of the data points, its periodicity and time evolution, three analysis techniques were employed to analyze the dihedral angle data.

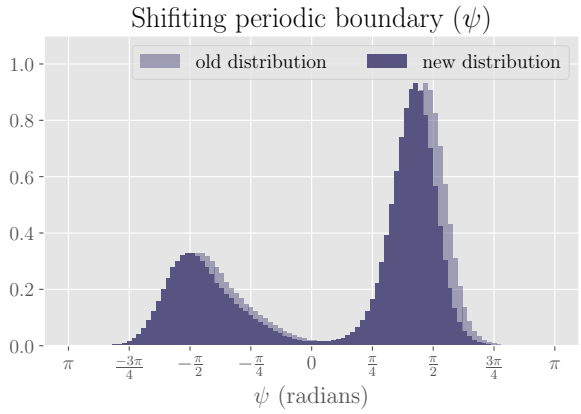
The PCA technique highlights the overall conformations present in the molecule. Although it was straightforward to implement, in the context of analysis of backbone dihedral angles, traditional PCA struggles with periodic metrics, as it's primarily designed for linear data. Both covariance and variance calculations, which are central to PCA, become flawed.

In contrast, dPCA transforms these angles using sine and cosine functions, effectively linearizing this circular data, preserving the true relationships and providing a more accurate analysis. This transformation addresses the core issue of the undefined mean in a circular context, ensuring that the resulting analysis represents the actual dynamics of the molecule in study. Although the minima of the free energy landscape looked more sharper to distinguish, they may give rise to complicated pattern, which are not so easy to interpret into the actual conformational states of the molecule.

Finally, the dPCA+ technique is designed to preserve the torus topology, and uses a linear transformation that minimizes periodicity-induced errors in covariance estimation and avoids artificial extra dimensions or distortions in probability distribution [6]. This was a comparatively simple algorithm to implement, with the only restriction that the data under consideration should have regions of maximal gap present in them.



(a) Original and shifted distributions for ϕ



(b) Original and shifted distributions for ψ

Figure 12: Shifted angles for ϕ and ψ

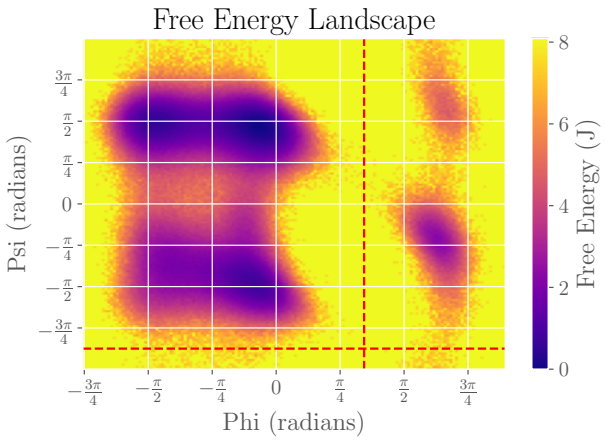


Figure 13: Initial (visual) guess for the maximal gap

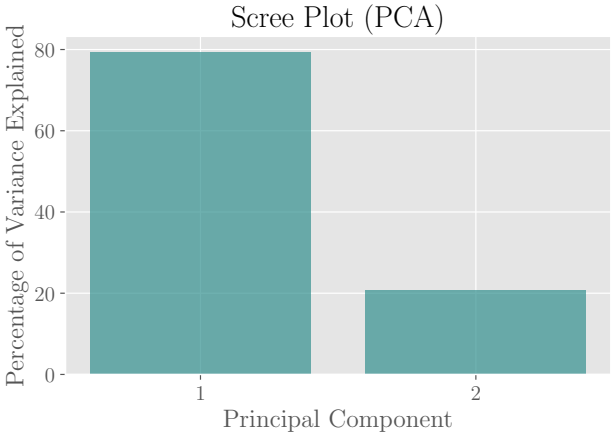


Figure 14: A scree plot of the two principal components obtained from the PCA of ϕ and ψ. PC 1 explains 79.29% of the total variance, while PC 2 explains 20.71% of the total variance.

A Structure of document and notebook

The former part of this document is structured in the conventional format of a research article. All code was performed in a Jupyter notebook, which is exported as a PDF and appended after this report. Some of the functions which are not directly significant to the main tasks are saved as a separate Python file `ccs_project_helpers.py`. The styling of the plots were done using the `ggplot` template with additional customizations made in a separate style file `ccs_project.mplstyle`.

Due to unforeseen complications in the LaTeXformatting, the figures in this document

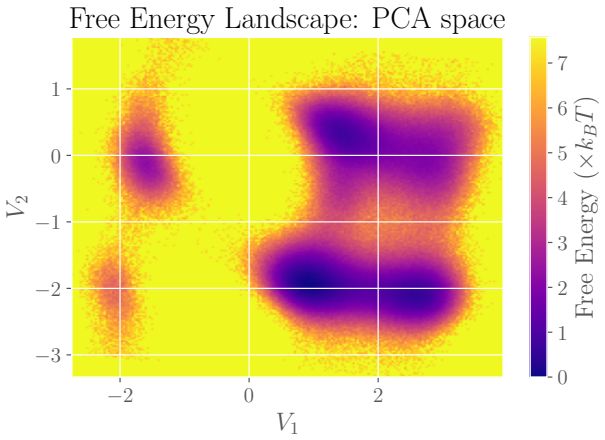


Figure 15

may not appear in their intended sequential order. Hence it is advised to follow the hyperlinks of the images and citations for better readability.

B Use of LLMs

The use of Large Language Models (LLMs) has become indispensable in assisting various sectors including research, education, and technology. In this project, the LLM ChatGPT by OpenAI was used occasionally, mainly to find word synonyms, clarify specific command syntax and to ease repetitive tasks (like creating labels for subplots), which helped in making the process more efficient. All code developed in the solution was authored independently, and ChatGPT was used responsibly and fairly as a support tool.

Some specific uses:

- Generated a blank LaTeXtemplate for the report with placeholders to add different sections and bibliography entries.
- Generated a function to label ticks in multiples of $\pi/4$ for the plots.
- It was also used in the learning stage to clarify meanings of definitions (like peptides and residues), although it did not prove to be much effective.

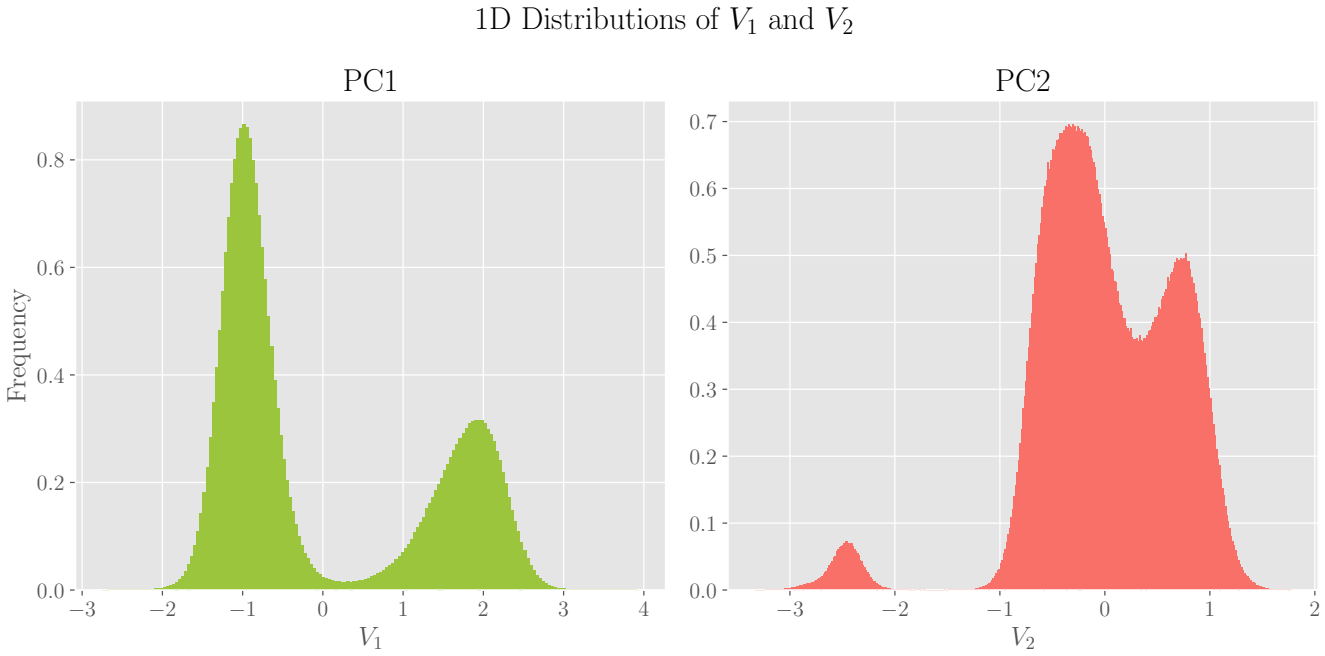


Figure 16

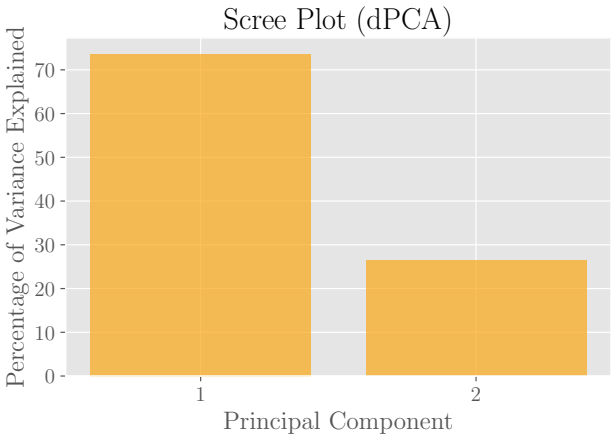


Figure 17: Scree plot from the dPCA: From the code, it was found that the four PCs explain 62.82%, 20.38%, 9.94% and 6.86% of the total variance respectively. Here the data was projected only to the first two PCs, according to the task guidelines.

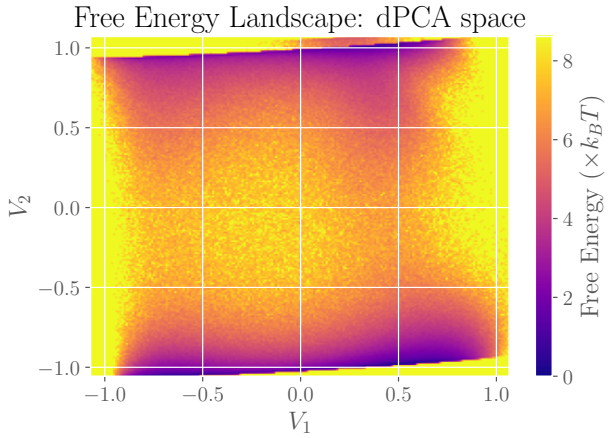


Figure 18: The free energy plot resulting from the dPCA.

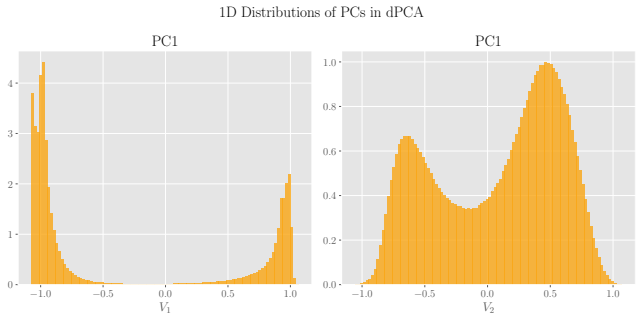


Figure 19: One-dimensional distributions of the PCs from dPCA.

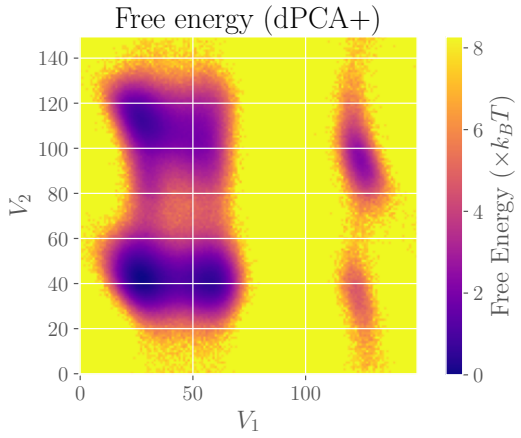


Figure 20

References

[1] *Dihedral angle - Wikipedia*. URL: https://en.wikipedia.org/wiki/Dihedral_angle?oldformat=true#Proteins.

[2] 'Dcrjsr' and Adam Rędzikowski. *Protein_backbone_PhiPsiOmega_drawing.svg.png (PNG Image, 315 × 599 pixels)*. URL: https://upload.wikimedia.org/wikipedia/commons/thumb/9/97/Protein_backbone_PhiPsiOmega_drawing.svg/315px-Protein_backbone_PhiPsiOmega_drawing.svg.png.

[3] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[4] Yuguang Mu, Phuong H. Nguyen, and Gerhard Stock. "Energy landscape of a small peptide revealed by dihedral angle principal component analysis". In: *Proteins: Structure, Function and Genetics* 58 (1 Jan. 2005), pp. 45–52. ISSN: 08873585. DOI: 10.1002/prot.20310.

[5] Alexandros Altis et al. "Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis". In: *Journal of Chemical Physics* 128 (24 2008). ISSN: 00219606. DOI: 10.1063/1.2945165.

[6] Florian Sittel, Thomas Filk, and Gerhard Stock. *Principal component analysis on a torus: Theory and application to protein dynamics*.

[7] G.N. Ramachandran and V. Sasisekharan. "Stereochemistry of polypeptide chain configurations." In: *J. mol. Biol* 7 (1963), pp. 95–99.

[8] Jane S. Richardson and David C. Richardson. "Principles and Patterns of Protein Conformation". In: Springer US, 1989, pp. 1–98. DOI: 10.1007/978-1-4613-1571-1_1.