Enhancing Interpretability of Protein Sequence Embeddings through Unsupervised Clustering

Farzaneh Saadati
farzaneh.saadati@uga.edu
University of Georgia
Athens, Georgia, USA

Noyon Dey
noyon.dey@uga.edu
University of Georgia
Athens, Georgia, USA

Korede Bish
Korede.Bish@uga.edu
University of Georgia
Athens, Georgia, USA

ABSTRACT

In this project, we aim to enhance the interpretability of protein sequence embeddings through unsupervised clustering methods, focusing specifically on kinase protein sequences. We embedded kinase protein sequences and utilized dimensionality reduction via t-SNE to visualize the local data structure effectively. Clustering was performed using K-means and Spectral Clustering, and the results were evaluated through metrics such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). We applied saliency maps to enhance the interpretability of clustering results, aiming to provide a clear understanding of which parts of the sequences are most influential. Our findings indicate that Spectral Clustering achieved better alignment with biological subfamily structures compared to K-means, suggesting a marginal improvement in capturing meaningful relationships within protein sequence embeddings. The use of saliency maps further demonstrated the interpretability of the model by highlighting critical features that align well with biological domain knowledge. Despite the challenges in data preprocessing and computational limitations, this work shows that adapting clustering techniques for better interpretability contributes valuable insights for biological data analysis, aiding in more informed interpretations of protein sequence studies.

ACM Reference Format:
Farzaneh Saadati, Noyon Dey, and Korede Bish. 2025. Enhancing Interpretability of Protein Sequence Embeddings through Unsupervised Clustering. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 BACKGROUND

Proteins are fundamental to biological systems, serving a wide array of functions that include catalyzing metabolic reactions, DNA replication, and signaling. Understanding the relationships between different protein sequences is essential for uncovering their biological roles and potential implications in disease mechanisms [6]. Kinase proteins, in particular, play a crucial role in cellular signaling and are involved in many regulatory processes, making them a significant focus of study in areas such as cancer research and drug development [1].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Traditional clustering methods, such as K-means and Spectral Clustering, have often been employed to group protein sequences based on their embeddings. However, these methods struggle to provide meaningful biological interpretability, especially when applied to advanced embedding models like ESM2 [5], which capture complex sequence-level features. Without adequate interpretability, it becomes challenging to align clustering results with biological knowledge, limiting the practical insights that can be drawn from the data. To address these issues, techniques that enhance interpretability are crucial for advancing our understanding of protein function and aiding in the development of targeted therapeutics. Hence, in this work, we aim to enhance the overall interpretability of protein sequence embeddings; by applying saliency maps to clustering results to find influential sequences.

2 RELATED WORK

Recent language models for protein sequences, such as ESM2 [5] and ESM3 [4], capture crucial sequence-level information that reflects protein structure and function. These embeddings are biologically informative and useful for tasks like clustering, simulating evolutionary information to group proteins with shared functionalities or evolutionary backgrounds.

Dimension reduction techniques like PCA and UMAP are frequently employed to visualize high-dimensional embeddings of protein sequences. These methods enhance interpretability by creating low-dimensional representations, revealing cluster boundaries and patterns within data [9].

Traditional clustering algorithms are widely used in bioinformatics to categorize protein sequences. However, they often lack biological interpretability, as they primarily focus on statistical data distribution [2].

Recent advancements in explainable deep learning, such as Integrated Gradients, provide insights into complex model outputs. [7] introduced CFA, an explainable deep learning model for annotating cis-regulatory modules, showing that highlighting significant features can enhance interpretability, which can be similarly applied to protein embeddings.

Proposed Phosformer [10], a transformer-based model for phosphorylation prediction in proteins, incorporating explainable features to align with biological functions. This work demonstrates the importance of integrating explainability into embedding-based methods for biologically relevant interpretations, which aligns with the goals of this project.

3 METHODOLOGY

In Figure 1, we have the pipeline of the project. The primary problem addressed in this project is enhancing the interpretability of

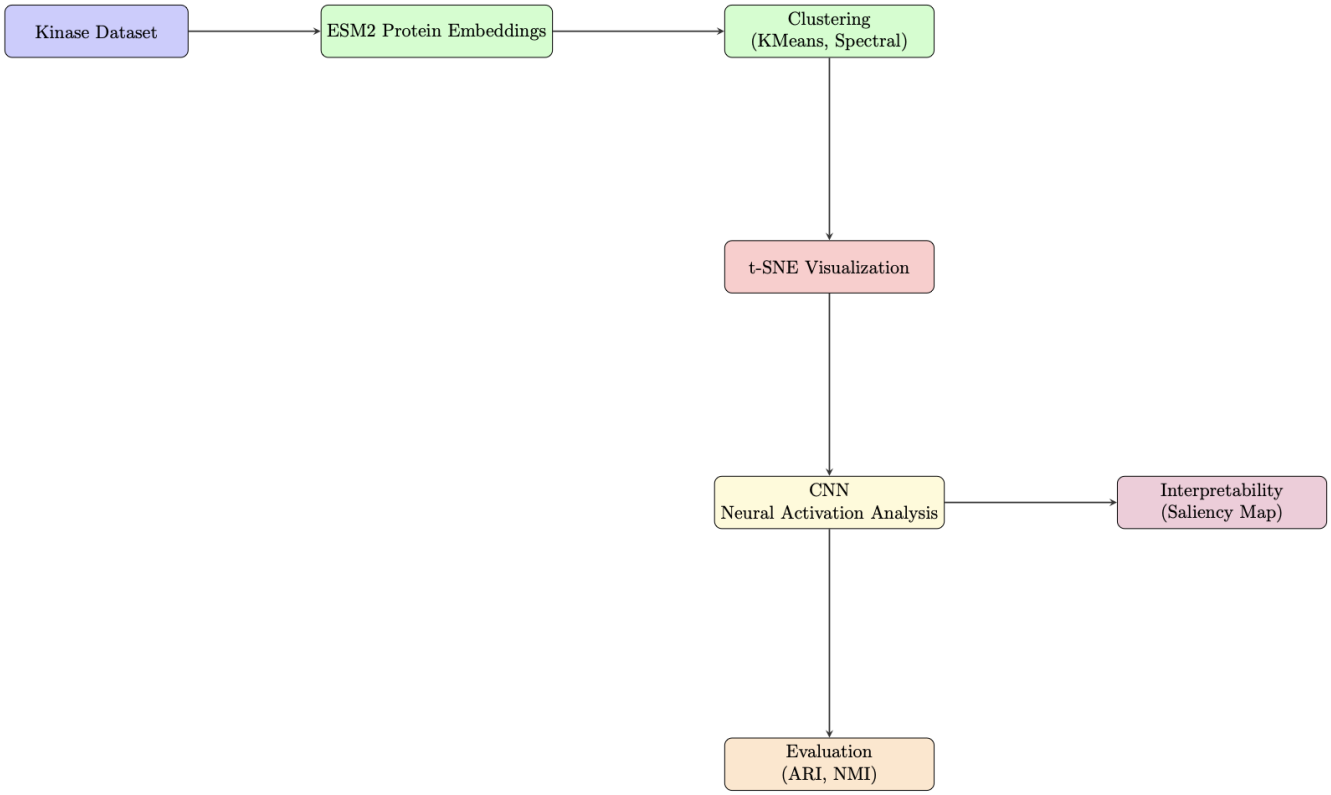


Figure 1: The pipeline illustrates the workflow of the project, starting with the Kinase dataset processed through ESM2 protein embeddings. These embeddings are clustered using KMeans and Spectral clustering methods, followed by t-SNE visualization to reveal the structure of the data. A CNN-based neural activation analysis is then applied for deeper interpretability, utilizing saliency maps to highlight key activation regions. Finally, the results are evaluated using metrics such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) to assess clustering quality and model performance.

protein sequence embeddings through unsupervised clustering. The input to our model consists of kinase protein sequences (e.g., P35916, VGFR3), represented by their amino acid sequences. These sequences are embedded using a pre-trained embedding model (ESM2) to generate high-dimensional vectors. In the first step, we visualized the embeddings by considering their subfamilies, which we used as a ground truth.

Next, we performed clustering of these protein sequences into biologically meaningful groups based on the embeddings, without explicit information (like labels representing subfamilies). The problem is defined as an unsupervised clustering task, with the goal of identifying natural groupings within the protein data and providing explanations for these groupings based on biological features.

3.1 Model and Method Details

Embedding Model: We utilized the ESM2 model, a state-of-the-art protein language model, to generate embeddings for the protein sequences. The ESM2 model employs a transformer architecture that captures sequence-level dependencies, generating embeddings that are informative of the protein's structural and functional properties.

In the clustering step, we used two unsupervised clustering algorithms: K-means and Spectral Clustering. K-means was used for its simplicity and efficiency in partitioning the data, while Spectral Clustering was used for its ability to capture more complex relationships within the data by considering local neighborhood structures.

Then, to facilitate visualization and clustering, we applied t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce the dimensionality of the protein embeddings from high-dimensional space to a 2D space shown in Figure 2. The t-SNE helps preserve local relationships between data points, making it easier to interpret the clusters visually.

3.2 Interpretability Techniques

Protein sequence embeddings are fed into a CNN (Convolutional Neural Network)-based model [8]. We wanted to incorporate a CNN model that allows for a deeper exploration of interpretability by examining which neurons in the network are activated by the input embeddings or their derived visual representations. By combining this with saliency maps, the contribution of individual input features (such as amino acids in protein sequences) to the activation of

specific neurons can be visualized. This can provide insights into the relationship between sequence features and clustering decisions.

3.3 Training Objective

Since this is an unsupervised learning task, there is no direct training process involved. Instead, the objective was to find meaningful clusters of protein sequences that align with known biological subfamilies. The embeddings were generated using a pre-trained model, and the clustering methods like K-means and Spectral clustering were performed on these embeddings to identify groups with similar biological properties.

4 EXPERIMENT

One interesting experiment we conducted was to compare the interpretability of clustering results obtained through K-means versus Spectral Clustering. Specifically, we used Saliency map to both clustering results to identify the key sequence features that influenced the cluster assignments. For each protein sequence, we generated local explanations using saliency map to highlight the important amino acids that contributed to its clustering.

You can find the code for the project on GitHub: **TML Final Project Repository**.

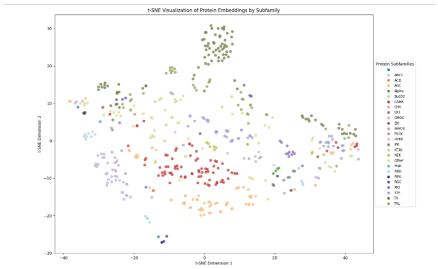
4.1 Experiment Settings

In this experiment, we focused on embedding protein sequences to enhance the interpretability of their relationships with respective subfamilies. Protein embeddings were generated to capture meaningful biological information and facilitate subsequent analysis. To analyze the embeddings, we applied t-SNE for dimensionality reduction.

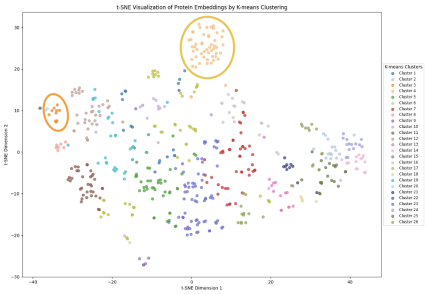
For clustering, we employed K-means and Spectral Clustering techniques. These methods were chosen for their complementary strengths in grouping data. K-means offers simplicity and efficiency, while Spectral Clustering effectively captures non-linear relationships in the embedding space.

By visualizing the clustered embeddings Figure 2, we were able to identify clear and interpretable groupings of protein subfamilies. Moreover, our findings indicate that Spectral Clustering achieved better alignment with biological subfamily structures compared to K-means, suggesting a marginal improvement in capturing meaningful relationships within protein sequence embeddings. The use of saliency maps further demonstrated the interpretability of the model by highlighting critical features that align well with biological domain knowledge. Despite the challenges in data preprocessing and computational limitations, this work shows that adapting clustering techniques for better interpretability contributes valuable insights for biological data analysis, aiding in more informed interpretations of protein sequence studies.

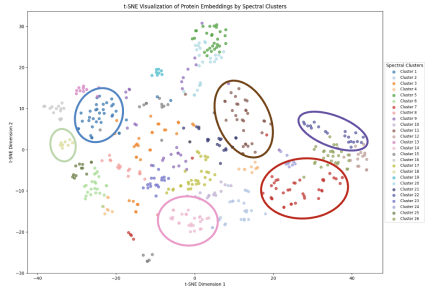
Figure 3 illustrates the saliency map generated for protein embeddings, highlighting the importance of sequence positions and embedding dimensions in contributing to clustering and interpretability. Each cell in the heatmap corresponds to a saliency value, where higher values (depicted in yellow) represent regions of the protein sequence and embedding space that are most influential in the model's predictions.



(a) Ground truth of protein subfamilies.



(b) K-means clustering of protein embeddings.



(c) Spectral clustering of protein embeddings.

Figure 2: (a) ground truth of 26 protein subfamilies, (b) displays the K-means clustering, and (c) represents the Spectral Clustering.

This visualization demonstrates how the model identifies critical regions in the protein sequence, aligning with domain knowledge by emphasizing biologically meaningful sequence positions. The saliency map serves as an interpretability tool, providing insights into the embedding space by correlating high-importance regions with biologically relevant patterns.

4.2 Experiment Results and Analysis

We evaluated the quality of clustering using Normalized Mutual Information (NMI) [3], can measure the mutual information between predicted labels and real labels Eqs. (2). Also, Adjusted Rand Index (ARI) [3] analyzes the degree of coincidence between predicted labels and real labels Eqs. (1).

ARI measures the similarity between the predicted clusters and the true biological groupings. Values closer to 1 indicate a high level of agreement between the predicted clusters and ground truth.

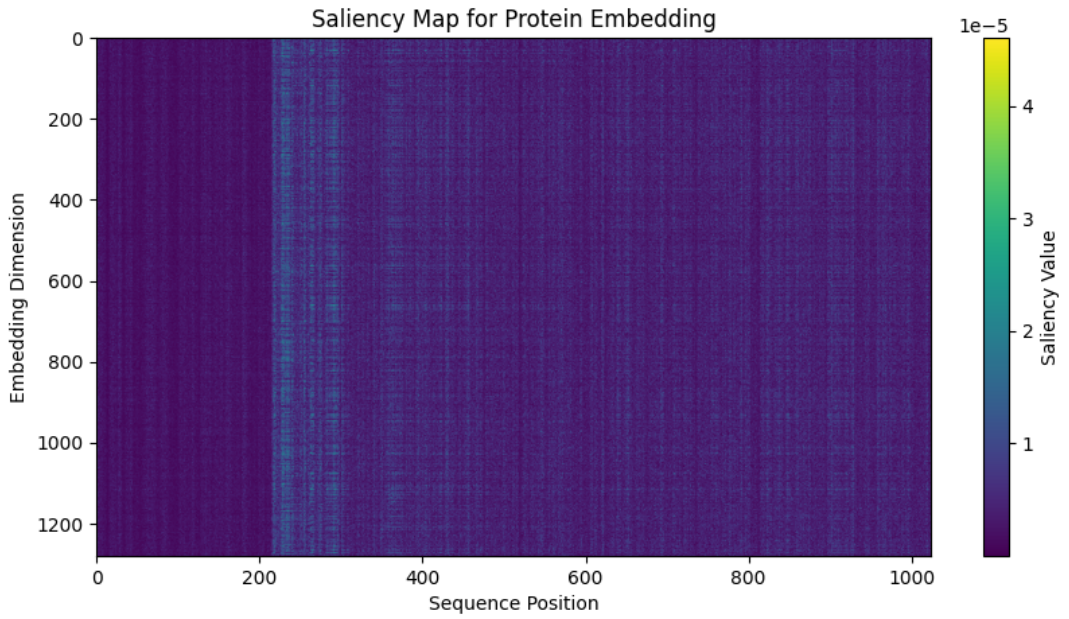


Figure 3: Saliency map highlighting the most influential sequence positions and embedding dimensions in protein embeddings.

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \quad (1)$$

NMI assesses the quality of clustering by evaluating how well the predicted clusters capture the underlying biological relationships. The score ranges from 0 (poor clustering) to 1 (perfect clustering).

$$\text{NMI} = \frac{2 \cdot I(U,V)}{H(U) + H(V)} \quad (2)$$

where $I(U,V)$ is the mutual information between clusters, and $H(U), H(V)$ are the entropies.

Clustering Method	ARI	NMI
K-means	0.2352	0.4554
Spectral Clustering	**0.2374**	**0.5505**

Table 1: Comparison of ARI and NMI scores for K-means and Spectral Clustering.

Table 1 compares the performance of K-means and Spectral Clustering using ARI and NMI metrics. Spectral Clustering achieves a slightly higher ARI (0.2374) than K-means (0.2352), showing comparable agreement with ground truth. In contrast, the NMI score for Spectral Clustering (0.5505) significantly outperforms K-means (0.4554), indicating better capture of meaningful relationships. These results suggest Spectral Clustering is more effective for biologically interpretable clustering of protein embeddings.

Key Takeaway: Adapting clustering techniques to enhance interpretability of protein embeddings contributes valuable insights for biological data analysis, supporting more informed interpretations in protein sequence studies.

REFERENCES

[1] Khushwant S Bhullar, Naiara Orrego Lagarón, Eileen M McGowan, Indu Parmar, Amitabh Jha, Basil P Hubbard, and HP Vasantha Rupasinghe. 2018. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer* 17 (2018), 1–20.

[2] Mahnoor Chaudhry, Imran Shafi, Mahnoor Mahnoor, Debora Libertad Ramírez Vargas, Ernesto Bautista Thompson, and Imran Ashraf. 2023. A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry* 15, 9 (2023), 1679.

[3] Jianhua Jia, Xuan Xiao, Bingxiang Liu, and Licheng Jiao. 2011. Bagging-based spectral clustering ensemble selection. *Pattern Recognition Letters* 32, 10 (2011), 1456–1467.

[4] John C Lin, Anagha Lokhande, Curtis E Margo, and Paul B Greenberg. 2022. Best practices for interviewing applicants for medical school admissions: a systematic review. *Perspectives on Medical Education* 11, 5 (2022), 239–246.

[5] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 6637 (2023), 1123–1130.

[6] Nature n.d.. *Protein Function.* https://www.nature.com/scitable/topicpage/protein-function-14123348

[7] Fritz Forbang Peleke, Simon Maria Zumkeller, Mehmet Gültas, Armin Schmitt, and Jędrzej Szymański. 2024. Deep learning the cis-regulatory code for gene expression in selected model plants. *Nature Communications* 15, 1 (2024), 3488.

[8] Ruijie Quan, Wenguan Wang, Fan Ma, Hehe Fan, and Yi Yang. 2024. Clustering for protein representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 319–329.

[9] Stefan Schoenfelder and Peter Fraser. 2019. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* 20, 8 (2019), 437–455.

[10] Zhongliang Zhou, Wayland Yeung, Saber Soleymani, Nathan Gravel, Mariah Salcedo, Sheng Li, and Natarajan Kannan. 2024. Using explainable machine learning to uncover the kinase–substrate interaction landscape. *Bioinformatics* 40, 2 (2024), btae033.