Comparative analysis of KAN with traditional neural networks like MLP on predicting Protein Folding Structure

Kshitij Tyagi
Mentor: Taniya Sarkar

November 26, 2024

Abstract

The recently introduced KAN(Kolmogorov-Arnold Network) models have demonstrated significant advantages, including scalability and learnability of the activation functions on edges. In this paper, we explore the potential improvements offered by the KAN architecture over traditional MLPs (Multi-Layer Perceptrons) in predicting the secondary structures of proteins, as well as comparing the performance of KANs with various activation functions. Our findings reveal that the traditional Spline-KANs exhibit slightly faster convergence with similar accuracy while some refinements (RBF-KANs and Chebyshev-KANs) show much faster convergences with great stability. Furthermore, we assessed the robustness of KAN with different activation functions, concluding that all KAN variants are robust with varying levels or robustness at varying levels of contamination.

1 Introduction and Motivation

Protein structure prediction is a fundamental challenge in bioinformatics and computational biology, with profound implications for understanding protein function, drug discovery, and disease mechanisms. Accurate prediction of protein structures can provide valuable insights into the molecular mechanisms of biological processes and aid in the development of therapeutic interventions. The advent of deep learning models like AlphaFold has marked a significant milestone in this field, demonstrating remarkable accuracy in predicting protein structures. However, the complexity and high computational demands of such models pose barriers to their widespread adoption, particularly for smaller-scale research and educational projects.

This study aims to investigate simpler neural network architectures, specifically Multi-layer Perceptrons (MLPs) with the recently introduced Kolmogorov-Arnold Networks (KANs), for the task of protein secondary structure prediction. By comparing the performance of MLPs and KANs on a benchmark dataset,

we aim to assess the feasibility of using these simpler models for protein structure prediction tasks. Our investigation focuses on evaluating the accuracy, convergence rate, and robustness of these models, providing insights into their practical applicability for protein secondary structure prediction.

2 Literature Review

Protein folding prediction, a cornerstone of computational biology, aims to determine a protein's three-dimensional structure from its amino acid sequence. This understanding is crucial for elucidating protein function and developing new drugs. MLPs are a class of feedforward artificial neural networks that have been the foundational blocks for protein folding prediction tasks due to their simplicity and ease of implementation. However, recent advancements in neural network architectures offer promising alternatives.

According to the authors of Liu et al. (2024), KANs have enhanced accuracy, interpretability, and neural scaling compared to MLPs, making them potentially well-suited for scientific tasks like protein folding. However, critical reviews highlight some limitations as well. Vikas Dhiman raises questions about their ability to overcome the curse of dimensionality Dhiman (2024). Additionally, how to enhance the models incorporating structural knowledge of the data, vital for protein folding tasks, remains an open challenge Samadi et al. (2024).

Research is actively addressing many limitations. Wav-KANs integrate wavelet functions, potentially leading to improved accuracy, faster training, and increased robustness in capturing the intricacies of protein sequences Bozorgasl and Chen (2024a). Furthermore, studies exploring KANs as surrogate models in evolutionary algorithms demonstrate their effectiveness in protein folding optimization tasks Hao et al. (2024).

A crucial aspect is comparing the performance of KANs with established methods like MLPs in various tasks. Research suggests that while original B-spline based KANs might not outperform MLPs, modified KANs with low-order orthogonal polynomials show promise Shukla et al. (2024). Abueidda et al. (2024) further demonstrate the effectiveness of RBF-based KANs (DeepOKAN) in mechanics problems, suggesting their potential for protein folding prediction as well. Recent studies have explored various KAN configurations, demonstrating their adaptability to different problem domains. Beyond the original B-spline based KANs, several variants have emerged:

- RBF-KANs: Employing Gaussian radial basis functions (RBFs) as activation functions, RBF-KANs have shown promising results in fields like mechanics Abueidda et al. (2024). Their ability to capture complex nonlinear realtionships might prove beneficial in protein folding prediction.

- Chebyshev Polynomial KANs Utilizing Chebyshev polynomials as activation functions, Chebyshev Polynomial KANs have demonstrated improved performance compared to B-spline based KANs in certain tasks Shukla

et al. (2024). Their orthogonal properties and efficiency might be advantageous for protein folding.

These diverse KAN architectures offer a rich landscape for exploration, with potential to outperform traditional MLPs in protein folding prediction.

2.1 Research Gap and Objectives

While the reviewed literature suggests promise for KANs in protein folding, a direct comparison with MLPs using a protein sequence dataset and established evaluation metrics is missing. This study aims to address this gap by evaluating the performance of KANs with different basis functions (B-splines, wavelets, and potentially RBFs) against MLPs in predicting protein structures.

3 Data Description

This study utilizes the 'Protein Secondary Structure Prediction' dataset from Kaggle (details provided in the references), collected from the RSCB Protein Data Bank on 6th July 2018, containing protein sequences and their corresponding secondary structure annotations. The dataset contains the eight-state secondary structures (SST-8) which are merged to create the final feature, the three-state tertiary structures (SST-3):

1. **C**: denotes loops, turns, or bends,
2. **H**: denotes helices and
3. **E**: denotes bridges or strands in the protein structure.

The dataset also contains other features like the pdb_id to locate the entry from RSCB PDB; the chain_code to locate the specific chain as a protein may consist of multiple; the length of the sequence and whether the peptide contains nonstandard amino acids (B, O, U, X, or Z).

The dataset includes sequences of varying lengths, representing a diverse set of proteins. The protein chain was then converted into a list of trigrams to allow for spatial structure since the tertiary structure of protein depends on the order of the amino acids as well as the amino acids present. The input features for the models are the trigrams (example shown in the figure below) of amino acids, and the target labels are the SST-3 codes of each residue in the sequence. This data was then encoded to make each input a vector and the output was one-hot encoded so that it could be used by the neural networks. The dataset was carefully filtered and 100 random data points were randomly sampled. This smaller sample size is analogus to a scenario where limited data is available for model training.

Figure 1: Illustration of N-grams

4 Models and Comparative Analysis

The models were investigated each with a single hidden layer of 25 neurons and trained with the RMSprop optimizer:

- **Multilayer Perceptron** (MLP): MLPs are based on the Universal Approximation Theorem. These consist of interconnected layers of neurons, where each neuron applies a fixed, non-linear activation function (e.g., ReLU) to a weighted sum of its inputs from the previous layer and some bias terms. These weights are learned during training to map the input data to the desired output.
- **Kolmogorov-Arnold Network** (KAN): Unlike MLPs with fixed activation functions on neurons, KANs employ learnable activation functions directly on the edges connecting neurons. The Kolmogorov-Arnold Representation Theorem suggests that any complex, multivariate function can be reformulated as an aggregate of a finite $(2n + 1)$ number of univariate functions. In the traditional KAN, these activation functions are represented by spline functions, allowing for greater flexibility compared to the linear weights in MLPs. This enables KANs to potentially capture more complex relationships within the data, leading to advantages in tasks like protein folding prediction where interpretability and accurate representation of underlying physical principles are crucial.

5 Results and Analysis

Both models, KAN and MLP, were trained for 50 epochs. Initially, the KAN model exhibited a lower accuracy compared to the MLP. However, KAN demonstrated rapid improvement, catching up within the first 4 epochs and then stabilizing. In contrast, the MLP model showed slow but steady progress, eventually reaching an accuracy comparable to KAN (approximately 44%) by the final 10 epochs. Notably, KAN achieved this performance with significantly fewer epochs, highlighting its efficiency and faster convergence in training [refer to Figure 2].

Next, we decided to investigate how KANs with different activation functions would perform. So, we repeated the same experiment with Chebyshev

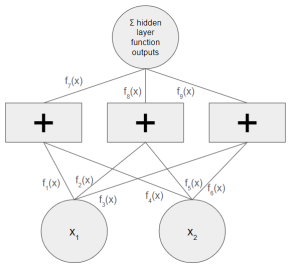
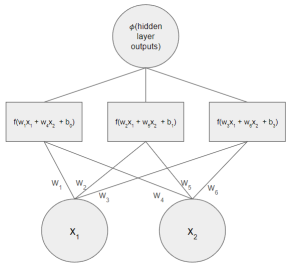
Model	KAN	MLP
Structure		
Formula	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \Phi_{q,p}(x_p) \right)$	$f(\mathbf{x}) \approx \sum_{i=1}^{N} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$

Table 1: Comparison of MLP vs KAN model structure

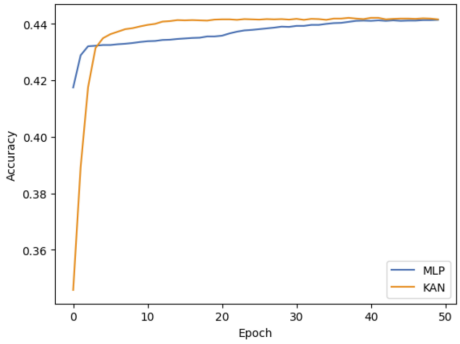


Figure 2: Accuracies of KAN and MLP on the dataset over 50 epochs

Polynomial KANs and RBF-KANs in addition to the Basis Spline (traditional) KANs. The comparison shows that even though the traditional Spline KAN performed better than MLP with lesser number of epochs, it is actually the lowest starting accuracy among the KAN variants. On the other hand, the RBF-KAN began with a modest accuracy but experienced a sharp improvement after the first epoch, continuing to increase gradually. The Chebyshev Polynomial KAN had the highest accuracy among the three from the very start and showed the most stability, demonstrating minimal deviations within the first 13 epochs and stabilizing early in the training process. Despite these differences in early performance, all three KAN models converged to similar accuracies by the end of the training, with RBF-KAN continuing to show improvements until the final epochs. [refer to Figure 3]

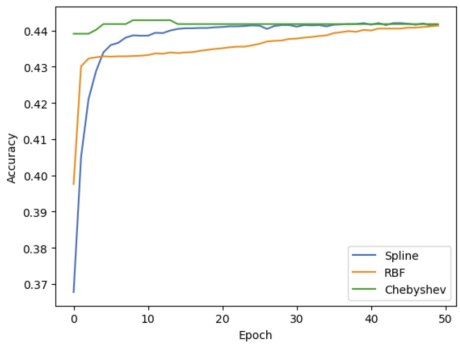
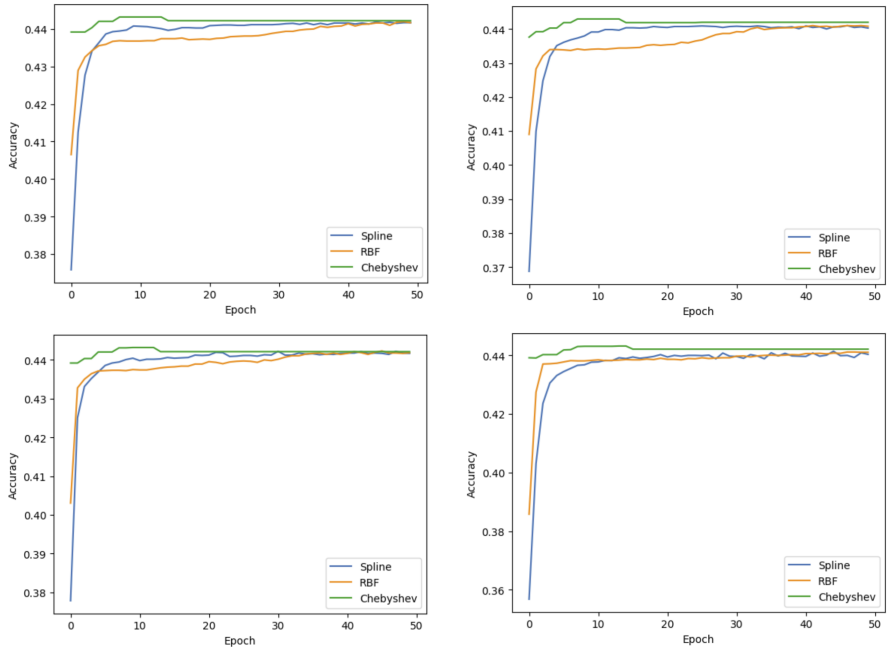


Figure 3: Comparison of KAN model performances with different activation functions on the dataset over 50 epochs

In the second part of the comparative study, we investigate the robustness of these different KAN models. Model robustness is crucial as it ensures reliability and accuracy in handling real-world data, which is often noisy and variable. It improves generalization to unseen datasets, enhances stability, and increases resilience against adversarial attacks, particularly in critical applications like healthcare, which is a direct application of this protein structure prediction task. To check for model robustness, we contaminate the original data at different contamination proportion (multiples of 10% upto 40%). We do this by randomly replacing the feature variable with any of the other 2 choices from the SST-3 values. The results are shown in the figures below.



In the four graphs presented, the differences from the original comparison without any contamination are negligible, with most models converging to similar accuracies at 50 epochs. However, at 40% contamination, the Spline KAN begins to exhibit significant oscillations, indicating potential instability under high levels of data contamination. While this may require some further research, overall, the refinements of the traditional KAN models (RBF KAN and Chebyshev polynomial KAN) seem to be quite robust to the effect of data contamination.

6 Additional Metrics

Model	Precision			Recall		
	C	H	E	C	H	E
MLP	0.63	0	0	0.46	0	0
Spline-KAN	0.48	0.23	0.08	0.64	0.30	0.09
Chebyshev-KAN	0.41	0.08	0.01	0.65	0.06	0.04
RBF-KAN	0.68	0	0	0.43	0	0

The table showcases the precision and recall metrics for the different models across each of the three secondary structure types: Coil (C), Helix (H), and Strand (E). Based on the results, the MLP model demonstrates high precision (0.63) and moderate recall (0.46) for Coil (C) structures but fails to predict Helix (H) and Strand (E) structures accurately, both having precision and recall values of 0. The Spline-KAN model provides a more balanced performance, with precision values of 0.48 for Coil, 0.23 for Helix, and 0.08 for Strand, along with the highest recall for Coil (0.64) and Helix (0.30), indicating a better detection capability for these structures. Chebyshev-KAN shows a high recall for Coil (0.65) but lower precision (0.41) and poor performance for Helix and Strand structures. RBF-KAN achieves the highest precision for Coil (0.68) but, like the MLP model, fails to detect Helix and Strand structures, with precision and recall values of 0. Overall, Spline-KAN appears to be the most balanced model, while Chebyshev-KAN and RBF-KAN demonstrate strengths primarily in detecting Coil structures.

7 Limitation and Future Directions

This study has several limitations. This paper does not introduce any improvements over KAN. Instead, we tried to provide a comprehensive study of the performances of KAN models for the protein structure prediction task. On the application side, the main limitation was that the dataset used was relatively small, and the model architectures employed were simple, limiting the generalizability of our findings. Future research should utilize larger datasets and more complex models with additional hidden layers to validate these results and explore the full potential of KANs in protein structure prediction. Additionally, our models were not well-equipped to handle spatial structures; although tri-grams were employed to partially address this, integrating KAN

with convolution networks or LSTM can significantly enhance KAN's performance in protein structure prediction. Numerous activation functions are also being introduced for KAN models apart from the 3 functions we studied here, like wavelet functions Bozorgasl and Chen (2024b), Jacobi polynomials, orthogonal polynomials, etc. Future work should focus on these advanced techniques to better capture the spatial dependencies in protein structures and enhance prediction performance.

7.1 References

Abueidda, D. W., Pantidis, P., and Mobasher, M. E. (2024). DeepOKAN: Deep operator network based on Kolmogorov Arnold Networks for Mechanics problems. *arXiv (Cornell University)*.

Bozorgasl, Z. and Chen, H. (2024a). WAV-KAN: Wavelet Kolmogorov-Arnold Networks. *Social Science Research Network*.

Bozorgasl, Z. and Chen, H. (2024b). Wav-kan: Wavelet kolmogorov-arnold networks.

Dhiman, V. (2024). KAN: Kolmogorov–Arnold Networks: A review. Technical report.

Hao, H., Zhang, X., Li, B., and Zhou, A. (2024). A first look at Kolmogorov-Arnold networks in surrogate-assisted evolutionary algorithms. *arXiv (Cornell University)*.

KaggleDataset (2018). Protein secondary structure. https://www.kaggle.com/datasets/alfrandom/protein-secondary-structure.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. (2024). KAN: Kolmogorov-Arnold Networks. *arXiv (Cornell University)*.

Samadi, M. E., Müller, Y., and Schuppert, A. (2024). Smooth Kolmogorov Arnold networks enabling structural knowledge representation. *arXiv (Cornell University)*.

Shukla, K., Toscano, J. D., Wang, Z., Zou, Z., and Karniadakis, G. E. (2024). A comprehensive and FAIR comparison between MLP and KAN representations for differential equations and operator networks. *arXiv (Cornell University)*.