Isolating a hand-crafted explanation for improved interpretability of biological foundation models

Jan van Eck[1] and Sanne Abeln[1]

1. AI Technology for Life, Department of Computing and Information Sciences, Department of Biology, Utrecht University, Utrecht, Netherlands

1 Introduction

The explainability of foundation models in the life science domain remains a challenge. While biological knowledge is based on well-established biochemical principles, foundation models often rely on unknown and implicit rules. To bridge the gap between these complex machine-learned representations and interpretable biological properties, isolating predefined explanations are essential. These explanations help align the learning processes of models with concepts that are understandable to humans. In this study, we focus on isolating a predefined biological explanation within the latent space of a foundation model. By doing so, we aim to enhance the interpretability of the predictions.

2 Biological Background

As a case study, we focus on the role of hydrophobicity in predicting peptide aggregation. Hydrophobicity refers to the tendency of certain amino acids to avoid water. This often results in residues being buried within the core of proteins [1]. Hydrophobic residues that are exposed to the surface of proteins can contribute to protein aggregation. This can lead to misfolding which plays a key role in many diseases, including neurodegenerative disorders such as Alzheimer's and Parkinson's [2].

3 Methods

We used a dataset of hexapeptides from WALTZ-DB 2.0 [3] to predict peptide aggregation. The model was based on an autoencoder architecture, with the ESM2 protein language foundation model [4] as the encoder for peptide sequences. Hydrophobicity was isolated as a crafted explanation during the prediction process. To achieve this, we implemented a combination of multiple loss functions. Orthogonality loss was used to ensure that the latent space features remained independent of each other, while the crafted explanations loss was designed to specifically target the isolation of the hydrophobicity signal. Additionally, aggregation loss and reconstruction loss were implemented to effectively predict aggregation. The prediction scores were then compared with AggBERT, the current state-of-the-art method for peptide aggregation prediction [5].

4 Results

By combining the multiple loss functions, the model successfully isolated the hydrophobicity signal from the remaining subspace. This separation allowed the model to improve interpretability without sacrificing performance. SHAP (SHapley Additive exPlanations) analysis further highlighted the significant role of hydrophobicity in peptide aggregation predictions, demonstrating that it was the most influential factor. Besides the improved interoperability, the model achieved an AUROC score of 0.89 surpassing the current state-of-the-art method.

5 Conclusion

This research not only establishes a state-of-the-art model for peptide aggregation prediction but also lays the groundwork for a framework to interpret deep learning models in biological applications. It offers new insights into the contribution of crafted known explanations like hydrophobicity within complex representations derived from deep learning methods. While this study successfully demonstrates the ability to isolate and estimate contribution of a single crafted explanation (hydrophobicity), determining the influence of multiple crafted explanations on the model remains a subject for future work.

References

1. K. A. Dill, "Theory for the folding and stability of globular proteins," Biochemistry, vol. 24, pp. 1501–1509, 1985.
2. F. Chiti and C. M. Dobson, "Protein misfolding, functional amyloid, and human disease," Annu. Rev. Biochem., vol. 75, pp. 333–366, 2006.
3. Louros, N., Konstantoulea, K., De Vleeschouwer, M., Ramakers, M., Schymkowitz, J., & Rousseau, F. (2020). WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. Nucleic acids research, 48(D1), D389-D393.
4. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130.
5. Perez, R., Li, X., Giannakoulias, S., & Petersson, E. J. (2023). AggBERT: Best in Class Prediction of Hexapeptide Amyloidogenesis with a Semi-Supervised ProtBERT Model. Journal of Chemical Information and Modeling, 63(18), 5727-5733.