PatchProt: Hydrophobic patch prediction using protein foundation models

Dea Gogishvili[1,2,*], Emmanuel Minois-Genin[1], Jan van Eck[2], and Sanne Abeln[1,2]

1 Bioinformatics, Computer Science Department, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
2 AI Technology for Life, Department of Computing and Information Sciences, Department of Biology, Utrecht University, Utrecht, Netherlands

*Corresponding author. E-mail: d.gogishvili@vu.nl
Data and code implemented in this study are available at: https://github.com/Deagogishvili/chapter-multi-task.

Abstract

Predicting large hydrophobic patches on the protein surfaces is a complex learning task [1]. Proteins typically hide hydrophobic residues within their core to avoid interaction with water, a phenomenon known as the hydrophobic effect [2, 3]. When such *sticky* residues appear on the surface, they can play key roles in functional protein-protein, -ligand, or -membrane interactions [4–6], as well as induce amyloid fibril formation in the context of aggregation diseases [7–9]. Keeping these residues internal is thought to be a key strategy to avert protein aggregation [10–12]. Hydrophobic areas on the surface of the protein can influence experimental processes, such as gel formation, protein crystallisation [13], and separation techniques [14]. Previously we developed a method to define the largest hydrophobic patch (LHP) - the largest connected hydrophobic area on the protein surface [1]. Additionally, we demonstrated the significance of exposed hydrophobic surfaces in terms of human disease [1]. LHPs can be used to identify aggregation-prone regions [15] which pose significant hurdles for the development of therapeutic proteins, such as monoclonal antibodies [15, 16]. Importantly, predicting the exposure of hydrophobic residues on the protein surface is not a trivial problem. Traditional methods predict the majority of hydrophobic residues to be fully buried [1, 17]. The continued evolution of the tools and methodologies is needed to deepen our understanding of protein hydrophobicity, especially in the context of neurodegenerative diseases.

This study builds upon protein foundation models and draws inspiration from recent advancements in deep learning architectures [18]. Current methodologies in protein property predictions focus on either global or local predictions. Here, we aimed to bridge this gap in current technology. The novelty of this framework lies in several key aspects. First, we introduce a multi-task learning approach that simultaneously predicts both global and local (L)HP values, a feature that has not been previously explored at the residue level. This dual-focus methodology

enables the model to learn commonalities and differences across tasks to improve generalisation, allowing us to explore other (un)related global tasks with limited data availability. Hence, we extended our train and test datasets with normalised expression annotations. This addition was inspired by our previous study, where we showed that highly hydrophobic proteins are generally expressed at lower levels in the human proteome [1]. Second, our parameter-efficient fine-tuning methodology enabled us to effectively train large transformer models, overcoming one of the major bottlenecks of large language models. Our framework allowed us to *(i)* outperform the state-of-the-art methods in primary tasks; *(ii)* improve the global LHP predictions; *(iii)* obtain the first model that can predict (L)HPs on a residue level. Moreover, PatchProt demonstrated the possibility of foundation models and multi-task strategies to improve the accuracy of protein property predictions even with sparse datasets.

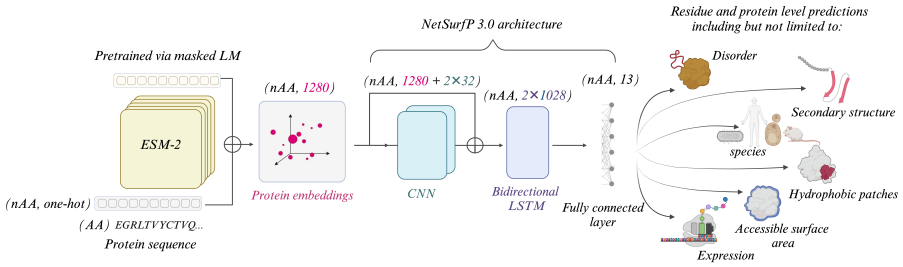

Fig. 1. Model architecture. The model takes protein sequence as input and predicts both global and local protein properties. The model consists of an embedding output from ESM-2 protein language model [19] and the downstream architecture similar to NetSurfP-3 [18]. Additionally, a parameter-efficient fine-tuning strategy was implemented [20, 21]. The decoding head consists of a residual block with two convolutional neural network (CNN) layers and a two-layer bidirectional long short-term memory (BiLSTM) network. The output is fed into a fully connected layer to provide predictions for all residues- and protein-level tasks.

References

1. J. H. M. van Gils, D. Gogishvili, J. van Eck, R. Bouwmeester, E. van Dijk, and S. Abeln, "How sticky are our proteins? quantifying hydrophobicity of the human proteome," *Bioinformatics advances*, vol. 2, no. 1, p. vbac002, 2022.
2. K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, pp. 1501–1509, 1985.
3. K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, pp. 7133–7155, 1990.
4. C. Chothia and J. Janin, "Principles of protein–protein recognition," *Nature*, vol. 256, pp. 705–708, 1975.

5. L. Young, R. Jernigan, and D. Covell, "A role for surface hydrophobicity in protein-protein recognition," *Protein Science*, vol. 3, no. 5, pp. 717–729, 1994.
6. S. M. Gowder, J. Chatterjee, T. Chaudhuri, and K. Paul, "Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins," *The Scientific World Journal*, vol. 2014, 2014.
7. M. G. Iadanza, R. Silvers, J. Boardman, H. I. Smith, T. K. Karamanos, G. T. Debelouchina, Y. Su, R. G. Griffin, N. A. Ranson, and S. E. Radford, "The structure of a β2-microglobulin fibril suggests a molecular basis for its amyloid polymorphism," *Nature Communications*, vol. 9, p. 4517, Dec. 2018.
8. M. D. Tuttle, G. Comellas, A. J. Nieuwkoop, D. J. Covell, D. A. Berthold, K. D. Kloepper, J. M. Courtney, J. K. Kim, A. M. Barclay, A. Kendall, W. Wan, G. Stubbs, C. D. Schwieters, V. M. Y. Lee, J. M. George, and C. M. Rienstra, "Solid-state NMR structure of a pathogenic fibril of full-length human [alpha]-synuclein," *Nat Struct Mol Biol*, vol. 23, pp. 409–415, May 2016.
9. F. Chiti and C. M. Dobson, "Protein misfolding, functional amyloid, and human disease," *Annu. Rev. Biochem.*, vol. 75, pp. 333–366, 2006.
10. C. M. Dobson, "Protein folding and disease: a view from the first horizon symposium," *Nature Reviews Drug Discovery*, vol. 2, pp. 154–160, 2003.
11. S. Abeln and D. Frenkel, "Disordered Flanks Prevent Peptide Aggregation," *PLoS Comput Biol*, vol. 4, no. 12, p. e1000241, 2008.
12. S. Abeln and D. Frenkel, "Accounting for protein-solvent contacts facilitates design of nonaggregating lattice proteins," *Biophysical journal*, vol. 100, no. 3, pp. 693–700, 2011.
13. P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *Journal of molecular biology*, vol. 293, pp. 321–331, 1999.
14. L. Moruz and L. Käll, "Peptide retention time prediction," *Mass spectrometry reviews*, vol. 36, no. 5, pp. 615–623, 2017.
15. K. Sankar, S. R. Krystek Jr, S. M. Carl, T. Day, and J. K. Maier, "Aggscore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 11, pp. 1147–1156, 2018.
16. J. M Redington, L. Breydo, and V. N Uversky, "When good goes awry: the aggregation of protein therapeutics," *Protein and Peptide Letters*, vol. 24, no. 4, pp. 340–347, 2017.
17. J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105 – 132, 1982.
18. M. H. Høie, E. N. Kiehl, B. Petersen, M. Nielsen, O. Winther, H. Nielsen, J. Hallgren, and P. Marcatili, "Netsurfp-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning," *Nucleic acids research*, vol. 50, no. W1, pp. W510–W515, 2022.
19. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.
20. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.
21. J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning," 2021.