Developing an Active Learning Framework for Transmembrane Helix Prediction
Wehner, Jessica¹; Ganapathiraju, Madhavi*¹,²

¹Bioengineering and Bioinformatics Summer Institute, Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA 15261
²Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15261

Over 30% of an organism’s proteins are membrane proteins¹ (MPs), which play key roles in important functional pathways such as signal transduction. However, experimental determination of MP structure is difficult, making computational prediction desirable. State-of-the-art prediction methods achieve acceptable accuracies for many proteins but perform poorly on uniquely structured proteins like aquaporin. Here, we developed an active learning framework that explores the space of unlabeled sequences and asks for experimentally determined labels for as few data points as possible. The purpose is to develop a strategy of identifying the next protein, which when solved experimentally, would disambiguate the labels of the remaining sequences. We computed features from primary sequence¹ and performed cluster-density driven selection of labels, yielding Q² of 74.0% at 280 (1%) selected data points. Further computations are expected to be completed by the end of the project.


* Correspondence should be addressed to madhavi+@pitt.edu