Fingerprinting Protein Adsorption Classes to Identify 'Ideal' Plant Protein-Based Emulsifiers

Simha Sridharan [1], *Rammile Ettelaie* [1], *Rik Sarkar* [2], *Maryam Afzali-Dela* [1,3], *Brent Murray* [1,3], *Nicholas Watson* [1,3], *Anwesha Sarkar* [1,3]

[1] *Food Colloids and Bioprocessing Group, School of Food Science and Nutrition, University of Leeds, Leeds, United Kingdom*
[2] *School of Informatics, University of Edinburgh, Edinburgh, United Kingdom*
[3] *National Alternative Protein Innovation Centre (NAPIC), United Kingdom*

L.Sridharan@leeds.ac.uk

Surfactants derived from petrochemicals are well classified and specifically designed for colloidal stabilization applications in industrial, food and bio colloids. However, due to their origins, these surfactants are not sustainable and can cause bio-toxicity. Therefore, proteins, especially derived from plants, can be considered as green surfactant to replace synthetic surfactants in food and soft matter formulations. However, proteins are poorly classified rendering their application for surfactant applications more arbitrary experimentation than targeted design. To address this gap, we present a systematic data driven approach that integrates data science with colloidal physics using Self-Consistent Field Theory (SCFT) modelling. This approach predicts adsorbed shapes of proteins from their primary amino acid sequences. We applied this approach for the first time to a large database of plant-derived proteins, simulating their adsorption and identifying emergent adsorbed shape (e.g.: loop-like, train-like) at interfaces. These adsorbed shapes were used to functionally cluster proteins using unsupervised machine learning (ML)-based clustering. By comparing the proteins in these clusters with surface-active agent, we uncover sequence features that influence adsorption properties. We benchmarked our functional clustering with state-of-the-art Protein Language Models (ProteinLM) such as ESM-2 and ProtBERT, to further understand the important sequence features that play a role in adsorption and adsorbed shape. For instance, our results show that the number of hydrophobic amino acids of a protein only weakly correlates ($R^2 = 0.3$) with its propensity to adsorb, challenging state-of-the-art in protein surfactant science. Finally, we validated experimentally the results using a combination of interfacial tension, Laser diffraction and confocal laser scanning microscopy in few plant proteins (identified from the clusters) in their disorder architecture showing emulsion formation and emulsion stability of those disordered plant proteins resembling those of dairy proteins. Overall, this study demonstrates the practical relevance of this new data science-driven approach, paving the way for rational protein design.

Keywords:
Plant Proteins, Machine Learning, Emulsions, Colloids, Surface Science

Acknowledgements:
Acknowledgement: Simha Sridharan acknowledges the funding from UKRI Guarantee fund for Marie Curie Postdoctoral Fellowship (EP/Z000785/1)