



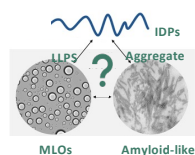
Protein Language Model Identifies Disordered, Conserved Motifs Driving Phase Separation

Yumeng Zhang¹, Jared Zheng², Bin Zhang^{1*}

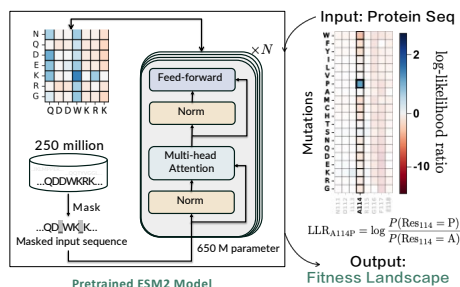
¹MIT Department of Chemistry, ²MIT Department of Electrical Engineering and Computer Science

Motivation & Questions

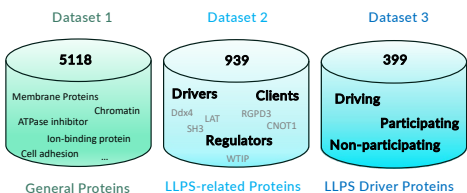
Intrinsically disordered proteins (IDPs) are key players in liquid-liquid phase separation (LLPS) and frequently regulate the formation of membraneless organelles (MLOs). Mutations in IDPs can disrupt their multivalent interaction network, altering the phase behavior, and leading to various diseases.



Language Model: Predict Protein Fitness Landscape



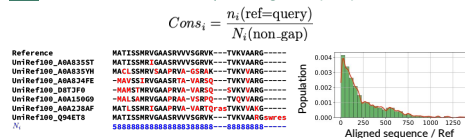
Dataset Preparation



Results: Biophysical Roles behind ESM2

❖ ESM2 scores reflect evolutionary conservation.

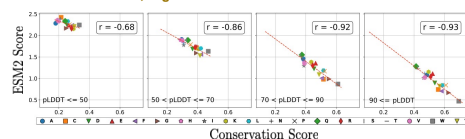
Conservation Score: from multi-sequence alignment (MSA)



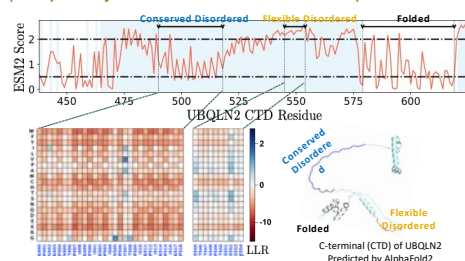
ESM2 Score: LLR value in PlogP scale

$$\text{ESM2 Score} = -\sum_{i=1}^{20} \left(\frac{e^{\text{LLR}_{\text{aa}_i}}}{\sum_{j=1}^{20} e^{\text{LLR}_{\text{aa}_j}}} \log \left(\frac{e^{\text{LLR}_{\text{aa}_i}}}{\sum_{j=1}^{20} e^{\text{LLR}_{\text{aa}_j}}} \right) \right)$$

1. ESM2 predicts Amino Acid Fitness landscape that correlates with conservation score, regardless of structural disorder.

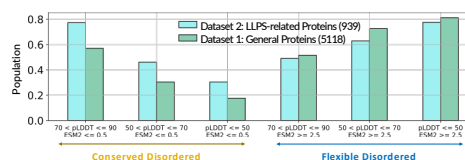


2. ESM2 predicts conserved residues on disordered regions (exemplified by human UBQLN2 C-terminal Domain).



Results: Conservation in LLPS-related IDPs

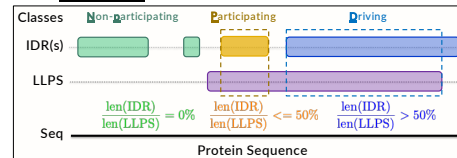
❖ Higher population of conserved disorder residues appear in LLPS-related proteins.



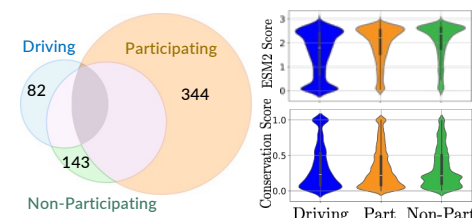
❖ Conservation in LLPS driving IDPs (Dataset 3).

Dataset 3: 399 proteins that have disordered region(s) overlap with experimental characterized LLPS region

Further Categorization: based on IDP involvement in LLPS



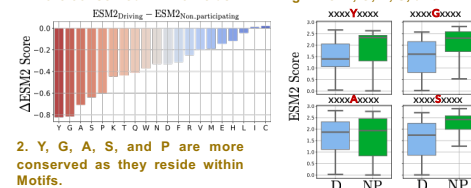
1. LLPS-driving IDPs have more conserved residues.



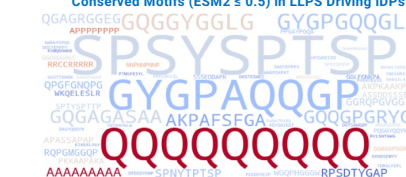
Results: Conserved Motifs in LLPS Driving IDPs

❖ Conserved Amino Acids overlap with Stickers.

1. More conserved Amino Acids in Driving IDPs: Y, G, A, S, and P.



2. Y, G, A, S, and P are more conserved as they reside within Motifs.



Conclusions and Future Plan

Conclusions:

- ESM2 Score is highly correlated with sequence conservation.
 - ESM2 identifies disordered, conserved regions.
 - LLPS-driving IDPs contain more conserved residues.
- Future Plan:
- Investigate the function of motifs during LLPS predicted by ESM2 predictions via simulations.

References

- Lee et al, *Exp Mol Med* (2022)
- Lin et al, *Science* (2023)
- Jumper et al, *Nature* (2021)
- Sood et al, *Biophysics J.* (2024)

Acknowledgement

