RESEARCH HIGHLIGHT

Beyond peptide targeting sequences: machine learning of cellular condensate localization

Jonathon A. Ditlev[1,2,3] and Julie D. Forman-Kay[1,2]

Cell Research (2025) 0:1–2; https://doi.org/10.1038/s41422-025-01115-6

Proteins within cells must navigate complex intracellular environments to co-localize with partners and regulate functional cellular organization. In a recent *Science* paper, Kilgore et al. report the development of ProtGPS, a machine learning-trained predictor of protein localization within biomolecular condensates in cells that can be used to predict the ability of disease-linked mutations to dysregulate protein localization to biomolecular condensates.

Cellular function depends on exquisite spatiotemporal control of proteins and other biomolecules to ensure that activities underlying cell fitness occur at the right time and in the right place. Previous work identified peptide targeting sequences in proteins (e.g., nuclear localization or export signals) that enable protein localization,[1] often to specific membrane-bound organelles, allowing researchers to predict and define targeting of proteins.

The concept of functional biological organization has been revolutionized by the identification of biomolecular condensates, including non-membrane-bound organelles and other non-membrane-bound intracellular and extracellular compartments, which form with significant contributions from phase separation.[2] Much like membrane-bound compartments, condensate function emerges from their biomolecular composition. Unlike peptide targeting sequences for specific membrane-bound organelles, the targeting of proteins to specific biomolecular condensates is not well understood. Recent computational analyses of protein regions that do not assume stable conformations on their own, termed intrinsically disordered regions (IDRs), have shown that they contain evolutionarily conserved "bulk" sequence features that are linked with their function and ability to localize with other proteins that participate in similar functions, including within condensates.[3,4] IDRs can contribute to condensate formation via particular molecular features, such as the pattern of aromatic residues,[5] that generate unique protein-rich solvent environments within condensates. Knowledge of these features has been used to develop machine learning computational models to predict composition of six condensates.[6]

Kilgore et al. report the development of ProtGPS, based on a protein language model, that predicts localization of human proteins to any of twelve unique condensates found in cells.[7] The authors used ~5500 protein sequences annotated for localization to nuclear speckles, PML bodies, postsynaptic densities, stress granules, Cajal bodies, P-bodies, nucleoli, transcription condensates, RNA transport granules, chromatin, cell junctions, and nuclear pore complexes. The ProtGPS model including multiple members of protein families in the training sets accurately predicts localization to the twelve types of condensates (AUC–ROC values from 0.83 to 0.95).

A strong test of the ability of a model to learn important features for a function, such as localization, is using the model for design. The authors built on ProtGPS to create a design algorithm, which additionally restrained sequences to those sharing known properties of IDRs (considered important for condensate localization) and sampling from natural proteins, in order to generate "new" proteins that could target desired condensates. Similar approaches have previously been used to design synthetic intrinsically disordered proteins with specific features that drive their localization to organelles and condensates.[8] The authors generated ten 100-residue-long sequences that were predicted to incorporate into the nucleolus. Co-expression of the designed sequences and a nucleolar marker, NPM1, resulted in high enrichment of four of the ten in the nucleolus while the remaining six were enriched to a lesser degree. Using ten designed sequences predicted to localize to nuclear speckles, however, did not lead to observable enrichment in speckles. While this illustrates the challenge of designing condensate-localizing sequences, for two of the ten, cytoplasmic condensates containing the designed protein and mislocalized nuclear speckle marker protein were seen, showing that the model has learned about speckle protein features.

A long-standing question is how disease-linked mutations found in IDRs drive functional changes in cells. There is growing appreciation for the potential effects of disease-linked variants in disordered regions on localization to condensates and condensate properties.[9] Previous work from the Young lab identified thousands of disease-linked mutations in condensate-associated proteins that are predicted to affect condensate formation or behavior.[10] Thus, a possible explanation for how mutations in IDRs promote disease is that they lead to mislocalization of condensate-targeted proteins. To test this exciting possibility, the authors used ProtGPS to predict significant changes in localization of disease-linked missense or truncation variants. They identified ten missense and ten truncation variants of proteins predicted by ProtGPS to perturb well-described normal cellular localization. Expression of fluorescently tagged variants in cells confirmed marked changes in condensate localization

[1]Molecular Medicine Program, Hospital for Sick Children, Toronto, ON, Canada. [2]Department of Biochemistry, University of Toronto, Toronto, ON, Canada. [3]Cell and Systems Biology Program, Hospital for Sick Children, Toronto, ON, Canada. email: jonathon.ditlev@sickkids.ca; forman@sickkids.ca

Published online: 14 April 2025

compared to wild-type protein, pointing to the ability of ProtGPS to predict the effect of mutations on targeting condensates.

Being able to predict protein localization in the complex cellular environment based solely on known sequences is an important step that will demonstrate a deep understanding of functional cellular organization, as well as provide hypotheses to guide experiments. To date, success in predicting condensate localization has been limited, due to our knowledge base of only a small subset of functional protein condensates, as well as the confounding impact of post-translational modifications, cellular conditions, and expression levels, including for other protein and nucleic acid binding partners. ProtGPS offers a potentially expandable machine learning model that can, in principle, be used to probe protein localization to any of the 100s of condensates identified in recent years[11] with appropriate training data. This will be incredibly useful for studying diseases linked with condensates, including neurodegenerative disorders and cancer. Building on the ability of ProtGPS to predict localization of disease-linked mutants, further development is needed to understand how mutations in condensate-localizing proteins alter the ability of other proteins to co-localize in condensates, as recently observed with autism spectrum disorder-linked Shank2 variants and FMRP.[12]

The inclusion of training set family members in the ProtGPS model exploits the evolutionary conservation of disordered region sequence features that specify condensate localization and other functional attributes,[3] parallel to the conservation of sequences that specify protein folds. Successes of tools such as ProtGPS that use such disordered region sequence features for prediction and design highlight the importance of these features for encoding functional protein localization within the dense cellular milieu.

REFERENCES

1. Owji, H. et al. *Eur. J. Cell Biol.* **97**, 422–441 (2018).
2. Spannl, S. et al. *Traffic* **20**, 890–911 (2019).
3. Zarin, T. et al. *Elife* **10**, e60220 (2021).
4. Vashishtha, S. et al. *J. Mol. Biol.* **437**, 168953 (2025).
5. Martin, E. W. et al. *Science* **367**, 694–699 (2020).
6. Saar, K. L. et al. *Nat. Commun.* **15**, 5418 (2024).
7. Kilgore, H. R. et al. *Science* **387**, 1095–1101 (2025).
8. Strome, B., Elemam, K., Pritisanac, I., Forman-Kay, J. D. & Moses, A. M. *bioRxiv* https://doi.org/10.1101/2023.04.28.538739 (2023).
9. Tsang, B. et al. *Cell* **183**, 1742–1756 (2020).
10. Banani, S. F. et al. *Dev. Cell* **57**, 1776–1788.e8 (2022).
11. Rostam, N. et al. *Nat. Methods* **20**, 673–676 (2023).
12. Liyanage, L. A. *bioRxiv* https://doi.org/10.1101/2024.11.01.621623 (2024).

COMPETING INTERESTS

J.A.D. is an advisor for Dewpoint Tx and Neurophase, A BioInnovation Institute Novo Nordisk Initiative.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Jonathon A. Ditlev or Julie D. Forman-Kay.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.