Chemical codes promote selective compartmentalization of proteins

Henry R. Kilgore[1,*,†], Itamar Chinn[2*], Peter G. Mikhael[2*], Ilan Mitnikov[2*], Catherine Van Dongen[1], Guy Zylberberg[2], Lena Afeyan [1,3], Salman Banani[1,4], Susana Wilson-Hawken[1,5], Tong Ihn Lee[1], Regina Barzilay[2,†], Richard A. Young[1,3,†]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA.

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[4]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

[5]Program of Computational & Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*These authors contributed equally to this work.
† Corresponding author. Email: hkilgore@wi.mit.edu (H.R.K), regina@csail.mit.edu (R.B.), young@wi.mit.edu (R.A.Y).

Summary

Cells have evolved mechanisms to distribute ~10 billion protein molecules to subcellular compartments where diverse proteins involved in shared functions must efficiently assemble. Such assembly is presumed to unfold as a result of specific interactions between biomolecules; however, recent evidence suggests that distinctive chemical environments within subcellular compartments may also play an important role. Here, we test the hypothesis that protein groups with shared functions also share codes that guide them to compartment destinations. To test our hypothesis, we developed a transformer large language model, called ProtGPS, that predicts with high performance the compartment localization of human proteins excluded from the training set. We then demonstrate ProtGPS can be used for guided generation of novel protein sequences that selectively assemble into specific compartments in cells. Furthermore, ProtGPS predictions were sensitive to disease-associated mutations that produce changes in protein compartmentalization, suggesting that this type of pathogenic dysfunction can be discovered in silico. Our results indicate that protein sequences contain not only a folding code, but also a previously unrecognized chemical code governing their distribution in specific cellular compartments.

Introduction

Groups of proteins involved in shared functions must efficiently assemble to fulfill their physiological functions[1]. For example, the fidelity of gene transcription hinges on the assembly of over a hundred different proteins at promoters, where some bind DNA sequences directly and others interact with DNA-bound proteins instead[2,3]. Selective protein-protein and protein-nucleic acid interactions are thought to be the predominant driving force leading to the assembly of specific proteins at locations where they carry out diverse functions[4-7]. Shape complementarity among structurally stable portions of proteins have dominated models of protein assembly, but there is now considerable evidence that large assemblies of proteins with shared functions also occur through weak multivalent electrostatic interactions[8-15]. Nearly all cellular functions involve formation of such assemblies, which have been described as condensates, aggregates, puncta, hubs and non-membrane bound compartments (Fig. 1a). In a recent study, we used small chemical probes to demonstrate that different condensates can harbor distinct internal chemical environments, suggesting that such assemblies have different solvent properties[16]. It is thus possible that protein molecules that assemble selectively with others in a condensate do so, in part, as a consequence of their compatibility with the internal solvating environment of that compartment[17-20]. This suggests the hypothesis that

proteins with shared physiological functions tend to assemble due to a shared chemical solvation code, which if verified, would have important implications for our understanding of cellular function and dysfunction.

Evidence for shared chemical codes in condensate compartments

To learn whether collections of proteins that assemble into specific condensate compartments have shared chemical codes, we adapted an evolutionary scale protein transformer language model (ESM2) to predict protein assembly into distinct compartments[21,22]. The transformer architecture of ESM2 allows for simultaneous relationships between all amino acids in an input sequence to be learned, providing a general strategy to detect chemical codes embedded in the amino acid sequence of a protein. We focused our studies on a set of 5,541 human protein sequences that have been annotated for twelve condensate compartments using the UNIPROT[23] and CD-Code[24] databases (Fig. 1b). The compartment identities of the proteins in these databases were determined with various experimental techniques and curated by experts in compartment annotation and whole sequences were used as input[44]. Compartment annotated whole protein sequences were used as input. A neural network classifier was jointly trained with ESM2 to develop a model, termed ProtGPS, which computes the independent probability of a protein being found within each of the twelve different condensate compartments (Fig. 1c). The area under the receiver operator curve (AUC-ROC) showed that protein compartments could be predicted with remarkable accuracy (0.83-0.95) across the 12 different compartments (Fig. 1d). The performance of the ProtGPS model indicates it detects patterns in the protein primary structure that segregate with presence in different condensate compartments.

Guided generation of novel protein sequences for compartment selectivity

In order to validate that ProtGPS has learned the chemical codes associated with condensate localization, we sought to design novel protein sequences that, when produced in cells, would selectively assemble into a compartment of interest. To test this idea, we initially designed protein sequences using an autoregressive greedy search algorithm (GS)[25]. We repeatedly feed a growing sequence into ProtGPS and extend the sequence with additional residues at the N-terminus up to a desired length, choosing an amino acid at each step that causes a protein to be classified as a match for the desired compartment (Fig. S1). For each protein, a plasmid was constructed that encoded a generated polypeptide of up to 150 amino acids with an N-terminal nuclear localization

sequence to ensure transport into the nucleus and a C-terminal mCherry protein to ascertain the location of the protein by microscopy. In all, we generated eight novel protein sequences were designed to assemble selectively into nucleoli (Table S1). However, although these proteins entered the nucleus, they failed to assemble selectively into nucleoli (Fig. S1).

The failure of our initial efforts to generate proteins that selectively compartmentalize in nucleoli led us to consider how the GS algorithm might be imperfect for this task and to design another approach that might be more successful. With GS and ProtGPS, protein sequences are generated without consideration of the chemical space of proteins found in nature, but are over optimized towards prediction of subcellular compartments. We sought to create an approach that could overcome this limitation by applying a concept borrowed from medicinal chemistry, where it is common to consider whether a molecule shares desirable physicochemical properties with others[26,27]. We also chose to favor the use of "intrinsically disordered" domains because they have been implicated in protein association with condensates and because they are less likely to introduce competing folded states[28-31]. To apply these concepts toward protein generation, we sought to constrain generation to (1) proteins in the chemical space[32] learned by ESM2, (2) domains that are intrinsically disordered, and (3) sequences that should localize to the intended compartment. Thus, we used additional features of protein chemical space and intrinsic disorder for our Markov Chain Monte Carlo algorithm (MCMC) (Fig. 2a).

We then used MCMC to perform guided generation of proteins, using the additional features described above, that would selectively assemble into two condensate assemblies, nucleoli[9] and nuclear speckles[33], that were selected because they are well-studied, have distinctive functions and morphologies, and possess unambiguous marker proteins (Fig. 2a). A total of twenty 100 amino acid long protein sequences, ten targeted to nucleoli and ten to nuclear speckles, were generated using the MCMC (Table S2, Fig. 2a). Specifically, we use blocked Gibbs sampling with MCMC where we start from a random subsequence and iteratively select residues to mutate such that the final sequence follows the data distribution defined by the proteins that ESM2 was trained on and with ProtGPS and DRBERT[34] to obtain the localization and disordered properties desired, respectively. For each protein, a plasmid was constructed that encoded the generated protein attached to an N-terminal nuclear localization sequence and a C-terminal mCherry protein. Each of the proteins were expressed in human cells together with the nucleolus marker NPM1-GFP or the nuclear speckle marker SRSF2-GFP and cells expressing both a test protein (mCherry) and the condensate marker (GFP) were isolated using flow cytometry.

Imaging of cells revealed that all ten proteins designed to assemble into nucleoli (NUC1-10) did indeed concentrate in nucleoli (Fig. 2b-d, S2). Imaging of cells expressing the ten proteins designed by MCMC to assemble into speckles revealed that six of ten (SPL1, 5, 7, 8,10) were enriched in nuclear speckles and two of ten were enriched in cytoplasmic bodies (Fig. 2d). The partition ratios of the SPL proteins in speckles were smaller than those observed for NUC proteins in nucleoli, which may be a consequence of the known distribution of speckle proteins into both nuclear speckles and much smaller RNA splicing condensates at nascent transcripts[35]. Two of the SPL proteins (SPL2 and SPL3) formed cytoplasmic foci that recruited at least one key speckle protein to the body (Fig. 2d, S4), suggesting that SPL2 and SPL3 have chemical features that assemble speckle proteins outside of the nucleus. We conclude that most proteins designed by MCMC to assemble into the two condensate compartments studied here - nucleoli and nuclear speckles – have indeed gained features that promote their concentration into these compartments.

Pathogenic mutations can alter chemical codes

Mutations can create pathogenic effects by altering a protein's function or altering a protein's subcellular distribution. Because ProtGPS can accurately predict the subcellular localization of normal proteins, it might be able to identify pathogenic mutations that cause a change in the subcellular location of a mutant protein. To test this possibility, we turned to the ClinVar[36] database, a public archive of a vast number of human variations classified for diseases. In all, we collected data for 121,971 single point mutations and 83,211 truncation mutations, and then used ProtGPS to determine if the changes in amino acid sequences were predicted to alter the subcellular distribution of the mutant proteins (Fig. 3a). We began our analysis by considering how mutations might influence the information content (of condensate) compartments. We computed the change in Shannon entropy[37] of the twelve condensate compartments to learn whether the predicted information content of these compartments changed. We find that for the mutated sequences, the compartment entropy is consistently higher when compared to the normal wild-type sequences across all compartments (Fig. 3b). Furthermore, we find that pathogenic single point and truncation mutations, when compared to normal proteins, tend to increase the Wasserstein distance[38], a metric of dissimilarity between two probability distributions (Fig. 3b). As expected, comparisons of truncation mutant proteins to normal proteins result in a larger Shannon entropy change and Wasserstein distance than those seen for single point mutations and normal proteins (Fig. 3b). These observations indicate that within this collection of pathogenic

proteins, sequence variation may alter the predicted compartments of proteins in ProtGPS, suggesting that some mutant proteins may no longer partition selectively into compartments in the same manner as their normal counterparts.

To experimentally test the prediction that some pathogenic variants cause a change in subcellular localization, we selected for study 20 pathogenic mutations (10 single point and 10 truncation mutations) in proteins involved in a broad range of biological functions and diseases and that caused a change in the chemical compartment code based on Wasserstein distance (Table S4). We then generated a panel of mouse embryonic stem cell (V6.5) lines stably expressing each protein from a doxycycline-inducible expression cassette, treated cells with doxycycline and conducted live cell confocal microscopy analysis. Differences in the subcellular localization between normal and mutant proteins would appear as changes in the fluorescence patterns displayed in micrographs. When comparing images of normal proteins with their mutant counterparts, we noted large differences in compartment appearance for 14 mutant proteins (e.g., BRD3) and smaller differences for the others (e.g., ESRP1) (Fig. 3c, d, Fig. S4, S5, Table S4). As an example of a large difference, BRD3 is a transcriptional coactivator that should normally associate with transcriptional condensates in the nucleus. However, the point mutation F334S causes a loss of signal in nuclear condensates and increased signal in the cytoplasm (Fig. 3c). The spliceosomal factor SRSF2, which normally associates with condensates engaged in RNA splicing in the nucleus, provided another example of a large difference. Specifically, SRSF2 carrying the S54H mutation lost the characteristic punctate signal in the nucleus concomitant with substantially elevated signal in the cytoplasm (Fig. 3c). The effects of the ESRP L259V mutation were, by contrast, more subtle, with a small reduction in the size of bodies in the nucleus (Fig. 3c). These results indicate that there is not a direct correlation between the Wasserstein Distance and the magnitude of the effect of compartment distribution, but support the notion that ProtGPS can detect changes in compartment codes due to pathogenic mutations that are demonstrable in an experimental setting.

Discussion

Our studies suggest that proteins have evolved to harbor at least two types of codes, one for folding and another for intracellular compartmentalization. Deep-learning algorithms such as AlphaFold2, RoseTTAFold, Chroma, EvoDiff, ESM2, and others can predict a protein's 3D structure from its linear amino acid sequence[22,39-44]. We here describe ProtGPS, which can predict a protein's selective assembly into specific condensate compartments

in cells and be used to guide generation of novel protein sequences whose cellular compartmentalization could largely be experimentally validated. The complexity of the underlying physiochemical rules for both protein folding and protein localization have proven difficult to parse using human interpretable approaches, and these deep-learning approaches therefore provide valuable predictive and analytical tools for the study of protein structure and function.

Previous studies of protein compartmentalization have suggested amino acid codes exist for some compartments. For the membrane-bound nucleus, for example, there are well-known nuclear localization sequences that facilitate the transport of protein from the cytoplasm to the nucleus[45-47]. More recently, models were used to identify patterns in protein sequences associated with specific compartments, especially those bounded by a membrane, but these did not sample a broad range of compartments and lacked generative experiments[48-50]. For nonmembrane compartments, here called condensates, there is recent evidence of patterned amino acid sequence features that can engender selective assembly of certain proteins into transcriptional and nucleolar condensates[51-53]. These observations are consistent with the concept of a chemical code in amino acid sequences that promotes the selective distribution of proteins into specific compartments. Furthermore, there is recent evidence of distinctive chemical environments within condensates, suggesting that these compartments have different solvent properties [16,53,54]. Thus, the patterns of amino acid sequences in proteins would be expected to both promote specific folding behaviors and to favor residence in compartments compatible with their solvent properties.

The patterns of amino acid sequences that occur in proteins appear overall to be highly constrained in biology[55-57], and we suggest that this is due, in part, to the requirements for both proper folding and subcellular distribution. In our efforts to develop ProtGPS as a guide for generating novel protein sequences that promote selective subcellular distribution, we found that protein sequences sampled from collections of natural proteins were far more successful at concentrating in the desired compartment than those generated without this consideration. Analogous to the medicinal chemist's aspiration to increase drug-like attributes such as on-target specificity and low off-target effects when generating variations in the chemistry of small molecule therapeutics, designing proteins to preferentially distribute in biochemically relevant regions of the targeted cell population might improve upon their therapeutic properties[16,54,58]. In addition, exploring the chemical space of proteins naturally present in specific biological compartments may provide an especially valuable guide to the generation of optimal chemical matter directed to target proteins in specific compartments. Indeed, there are widely used and efficacious anti-cancer

therapeutics that concentrate in transcriptional condensates at oncogenes[58] due to the chemical environment of those compartments[16,54]. It is evident that similar considerations will apply to the design of protein therapeutics. We suggest that further understanding of the chemical environment established by amino acid patterns in proteins will lead to more efficacious disease therapeutics.

We conclude that ProtGPS can predict a protein's selective assembly into specific condensates and guide generation of novel protein sequences whose cellular compartmentalization can be experimentally validated. We anticipate that future studies will advance this field by improving compartment annotation, conducting additional tests of generated proteins, deploying alternative ML approaches, and further exploring the effects of pathogenic mutations.

Acknowledgments:

Support was provided by NIH grant nos. GM144283 (R.A.Y.) and CA155258 (R.A.Y.), NSF grant no. PHY2044895 (R.A.Y.), Damon Runyon Cancer Research Foundation Fellowship grant no. 2458-22 (H.R.K.), and the MIT Jameel Clinic for Machine Learning in Health (R.B., P.G.M., I.C., I.M.) and Eric and Wendy Schmidt Center, Broad Institute (P.G.M., I.C.). We thank Christina Lilliehook, Alessandra Dall'Agnese, Mike Gallagher, Yana Petri, Jinyi Yang, Shannon Moreno, and Jeremy Wohlwend for helpful comments and thank Caitlin Rausch and Warbler Creative for graphical artwork.

Competing Interests:

R.A.Y. is a founder and shareholder of Syros Pharmaceuticals, Camp4 Therapeutics, Omega Therapeutics, Dewpoint Therapeutics and Paratus Sciences, and has consulting or advisory roles at Precede Biosciences and Novo Nordisk. R.B. has consulting or advisory roles at Dewpoint Therapeutics, J&J, Amgen, Outcomes4Me, Immunai and Firmenich. H.R.K. is a consultant of Dewpoint Therapeutics. Itamar Chinn is a consultant of Paratus Sciences. The remaining authors declare no competing interests.

References

1 Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular and Cell Biology* **18**, 285-285, doi:10.1038/NRM.2017.7 (2017).

2 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).

3 Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45-54, doi:10.1038/s41586-019-1517-4 (2019).

4 Jena, S. *et al.* Noncovalent interactions in proteins and nucleic acids: beyond hydrogen bonding and π-stacking. *Chemical Society Reviews* **51**, 4261-4286, doi:10.1039/D2CS00133K (2022).

5 Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509, doi:10.1038/nature22366 (2017).

6 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408, doi:10.1038/s41586-020-2188-x (2020).

7 Walport, L. J., Low, J. K. K., Matthews, J. M. & Mackay, J. P. The characterization of protein interactions – what, how and how much? *Chemical Society Reviews* **50**, 12292-12307, doi:10.1039/D1CS00548K (2021).

8 Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, doi:10.1126/SCIENCE.AAF4382 (2017).

9 Feric, M. *et al.* Coexisting liquid phases underlie nucleolar subcompartments. *Cell* **165**, 1686-1697 (2016).

10 Alberti, S. & Hyman, A. A. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nature Reviews Molecular Cell Biology* **22**, 196-213, doi:10.1038/s41580-020-00326-6 (2021).

11 Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annual Review of Biophysics* **49**, 107-133, doi:10.1146/annurev-biophys-121219-081629 (2020).

12 Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M. & Forman-Kay, J. D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell* **183**, 1742-1756, doi:10.1016/j.cell.2020.11.050 (2020).

13 Cho, W.-K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412-415, doi:10.1126/science.aar4199 (2018).

14 Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).

15 Sheinerman, F. B., Norel, R. & Honig, B. Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology* **10**, 153-159, doi:https://doi.org/10.1016/S0959-440X(00)00065-8 (2000).

16 Kilgore, H. R. *et al.* Distinct chemical environments in biomolecular condensates. *Nature Chemical Biology*, doi:10.1038/s41589-023-01432-0 (2023).

17 Yu, Y., Wang, J., Shao, Q., Shi, J. & Zhu, W. The effects of organic solvents on the folding pathway and associated thermodynamics of proteins: a microscopic view. *Scientific Reports* **6**, 19500, doi:10.1038/srep19500 (2016).

18 Ben-Naim, A. Solvent effects on protein association and protein folding. *Biopolymers* **29**, 567-596, doi:https://doi.org/10.1002/bip.360290312 (1990).

19 Klibanov, A. M. Improving enzymes by using them in organic solvents. *Nature* **409**, 241-246, doi:10.1038/35051719 (2001).

20 Prabhu, N. & Sharp, K. Protein−Solvent Interactions. *Chemical Reviews* **106**, 1616-1623, doi:10.1021/cr040437f (2006).

21 Chandra, A., Tünnermann, L., Löfstedt, T. & Gratz, R. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife* **12**, e82819, doi:10.7554/eLife.82819 (2023).

22 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130, doi:10.1126/science.ade2574 (2023).

23 The UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2021).

24 Rostam, N. *et al.* CD-CODE: crowdsourcing condensate database and encyclopedia. *Nature Methods* **20**, 673-676, doi:10.1038/s41592-023-01831-0 (2023).

25 Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nature Communications* **12**, 2403, doi:10.1038/s41467-021-22732-w (2021).

26 Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855-861, doi:10.1038/nature03193 (2004).
27 Beckers, M., Fechner, N. & Stiefl, N. 25 Years of Small-Molecule Optimization at Novartis: A Retrospective Analysis of Chemical Series Evolution. *Journal of Chemical Information and Modeling* **62**, 6002-6021, doi:10.1021/acs.jcim.2c00785 (2022).
28 Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* **25**, 187-211, doi:10.1038/s41580-023-00673-0 (2024).
29 van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **114**, 6589-6631, doi:10.1021/cr400525m (2014).
30 Zarin, T. *et al.* Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *eLife* **10**, e60220, doi:10.7554/eLife.60220 (2021).
31 Bob, S., Khaled, E., Iva, P., Julie, D. F.-K. & Alan, M. M. Computational design of intrinsically disordered protein regions by matching bulk molecular properties. *bioRxiv*, 2023.2004.2028.538739, doi:10.1101/2023.04.28.538739 (2023).
32 Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823-823, doi:10.1038/432823a (2004).
33 Ilik, İ. A. *et al.* SON and SRRM2 are essential for nuclear speckle formation. *eLife* **9**, e60579, doi:10.7554/eLife.60579 (2020).
34 Ananthan, N., John Malcolm, F., Simon, L. & Sergei, M. DR-BERT: A Protein Language Model to Annotate Disordered Regions. *bioRxiv*, 2023.2002.2022.529574, doi:10.1101/2023.02.22.529574 (2023).
35 Guo, Y. E. *et al.* Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572**, 543-548, doi:10.1038/s41586-019-1464-0 (2019).
36 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).
37 Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379-423, doi:10.1002/j.1538-7305.1948.tb01338.x (1948).
38 Villani, C. in *Optimal Transport: Old and New* (ed Cédric Villani) 93-111 (Springer Berlin Heidelberg, 2009).
39 Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100, doi:10.1038/s41586-023-06415-8 (2023).
40 Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070-1078, doi:10.1038/s41586-023-06728-8 (2023).
41 Sarah, A. *et al.* Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.2009.2011.556673, doi:10.1101/2023.09.11.556673 (2023).
42 Sidney Lyayuga, L. *et al.* Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion. *bioRxiv*, 2023.2005.2008.539766, doi:10.1101/2023.05.08.539766 (2023).
43 Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **0**, eadl2528, doi:10.1126/science.adl2528.
44 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).
45 De Robertis, E. M., Longthorne, R. F. & Gurdon, J. B. Intracellular migration of nuclear proteins in Xenopus oocytes. *Nature* **272**, 254-256, doi:10.1038/272254a0 (1978).
46 Dingwall, C., Sharnick, S. V. & Laskey, R. A. A polypeptide domain that specifies migration of nucleoplasmin into the nucleus. *Cell* **30**, 449-458, doi:10.1016/0092-8674(82)90242-2 (1982).
47 Lu, J. *et al.* Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell Communication and Signaling* **19**, 60, doi:10.1186/s12964-021-00741-y (2021).
48 Kobayashi, H., Cheveralls, K. C., Leonetti, M. D. & Royer, L. A. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature Methods* **19**, 995-1003, doi:10.1038/s41592-022-01541-z (2022).
49 Jiang, Y. *et al.* MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Computational and Structural Biotechnology Journal* **19**, 4825-4839, doi:https://doi.org/10.1016/j.csbj.2021.08.027 (2021).
50 Thumuluri, V., Almagro Armenteros, J. J., Johansen, Alexander R., Nielsen, H. & Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research* **50**, W228-W234, doi:10.1093/nar/gkac278 (2022).

51 Patil, A. *et al.* A disordered region controls cBAF activity via condensation and partner recruitment. *Cell* **186**, 4936-4955.e4926, doi:https://doi.org/10.1016/j.cell.2023.08.032 (2023).
52 Lyons, H. *et al.* Functional partitioning of transcriptional regulators by patterned charge blocks. *Cell* **186**, 327-345.e328, doi:https://doi.org/10.1016/j.cell.2022.12.013 (2023).
53 King, M. R. *et al.* Macromolecular condensation organizes nucleolar sub-phases to set up a pH gradient. *Cell*, doi:10.1016/j.cell.2024.02.029.
54 Kilgore, H. R. & Young, R. A. Learning the chemical grammar of biomolecular condensates. *Nat. Chem. Biol.*, 1298-1306, doi:10.1038/s41589-022-01046-y (2022).
55 Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673-677, doi:10.1126/science.1257360 (2015).
56 Repecka, D. *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence* **3**, 324-333, doi:10.1038/s42256-021-00310-5 (2021).
57 Andre, J. F., Aina, M.-A., Cristina, H.-C., Jörn, M. S. & Ben, L. The genetic architecture of protein stability. *bioRxiv*, 2023.2010.2027.564339, doi:10.1101/2023.10.27.564339 (2023).
58 Klein, I. A. *et al.* Partitioning of cancer therapeutics in nuclear condensates. *Science* **368**, 1386, doi:10.1126/science.aaz4427 (2020).

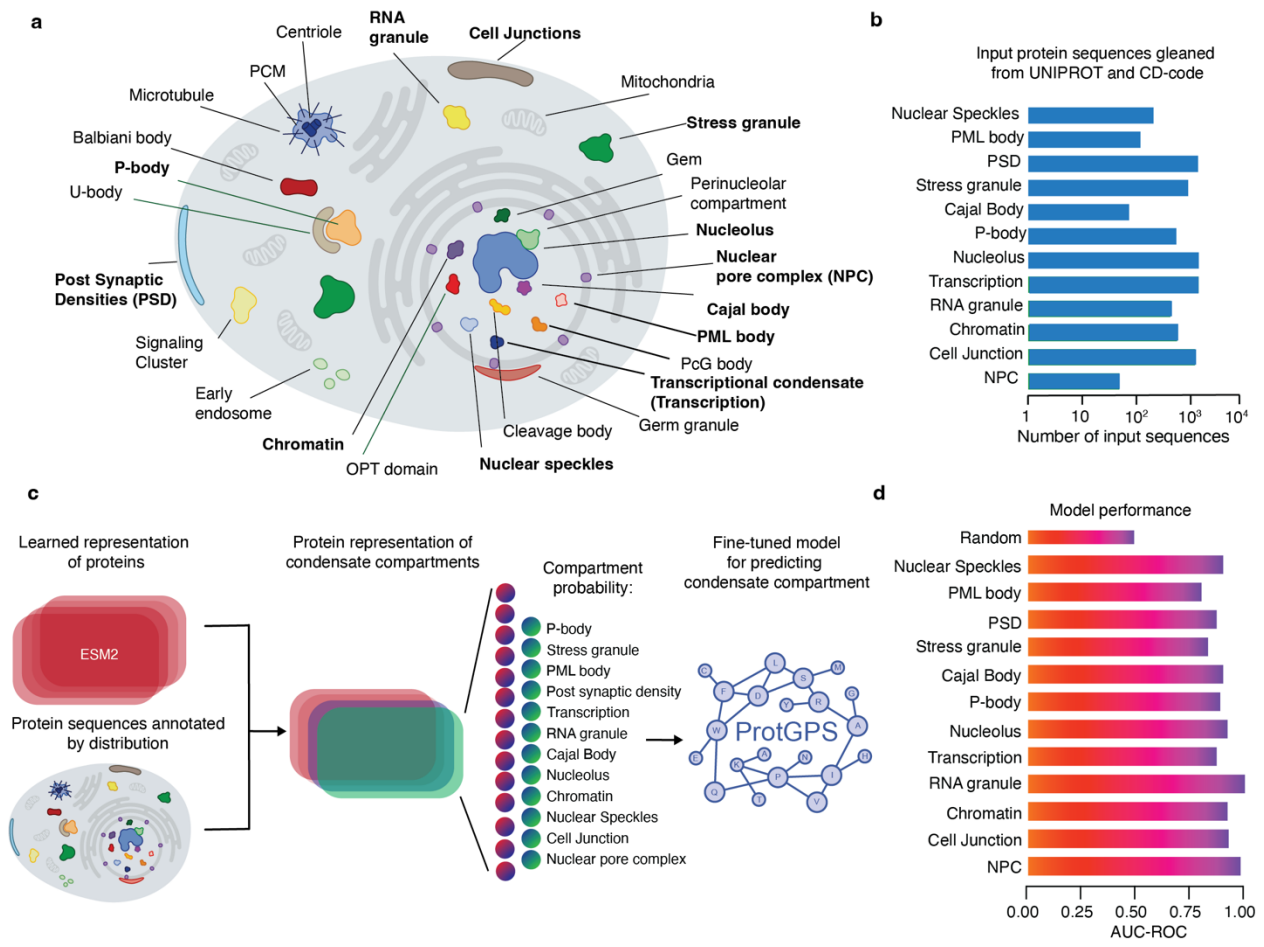


Figure 1. ProtGPS classifies protein compartment with high performance. a. Graphical depiction of some cellular compartments found in eukaryotic cells, compartments in bold were studied in this work. **b**. Bar graph showing the number of protein sequences gathered from UNIPROT and the CD-code database used in the development of ProtGPS. **c**. Schematic showing the approach toward developing ProtGPS. **d**. Bar graph showing the area under the receiver-operator curve for classification of withheld test data (15 % of total) with ProtGPS.

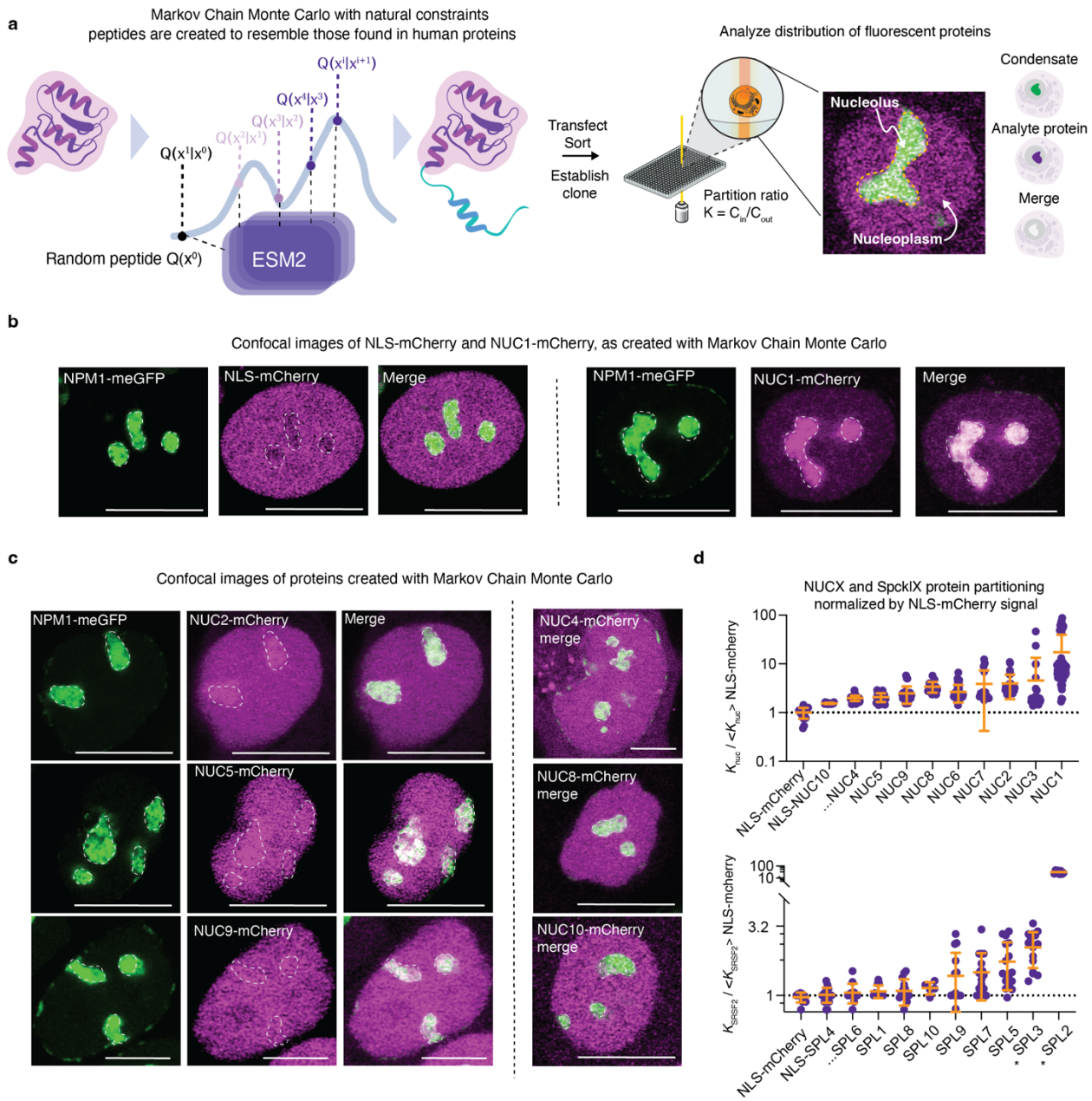
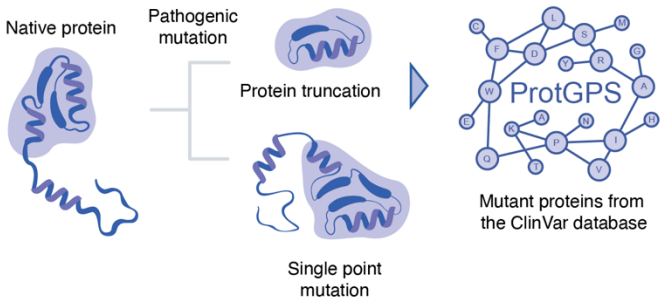
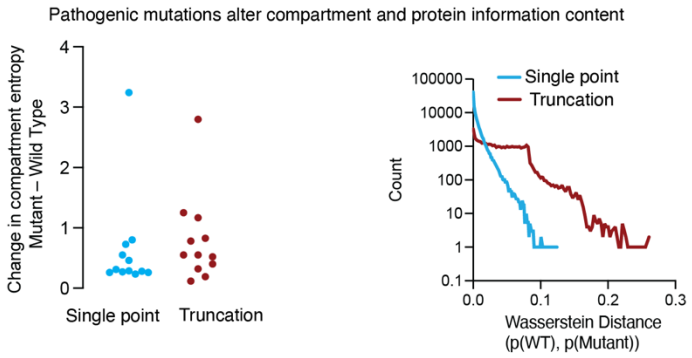


Fig. 2. Generative modeling creates novel proteins that concentrate in a desired condensate. a. Schematic showing the use of Naturally constrained Markov Chain Monte Carlo to generate proteins and assay them in live cells (MCMC) (*see supporting information for more details*). **b.** Live cell image of a colon cancer cell (HCT-116) tagged at the endogenous NPM1 locus with GFP and expressing nucleolus targeted protein NUC1-mCherry, scale: 10 microns**. c.** Live cell confocal micrographs of NUCX-mCherry proteins in HCT-116 cells expressing NPM1-GFP from the endogenous locus cells, scale: 10 microns. **d**. Dot plots showing the computed partition ratios of NUCX ($K = I_{nucleolus} / I_{nucleoplasm}$) and SPLX-mCherry ($K = I_{SRSF2} / I_{nucleoplasm}$ or $= I_{SRSF2} / I_{cytoplasm}$, as indicated by *) proteins relative to the NLS-mCherry control protein, dotted line is the average value of NLS-mCherry protein. See Table S3 for more information.

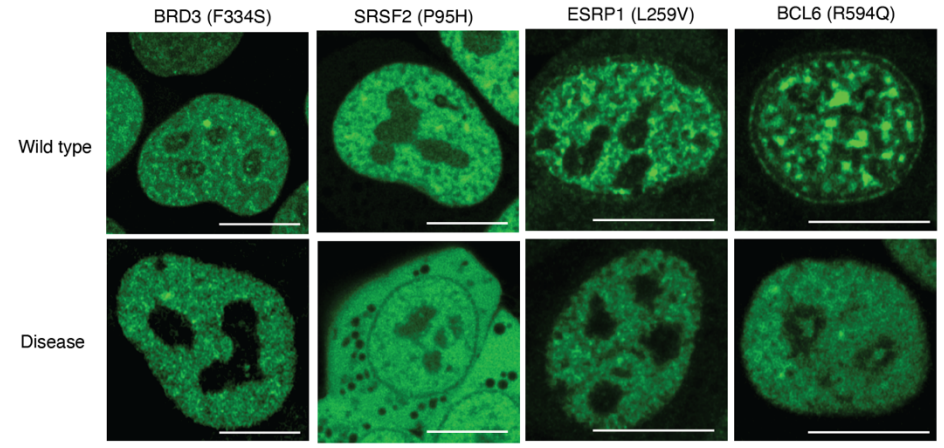
a


Predicting changes in subcellular location caused by pathogenic mutations
Native protein
Pathogenic mutation
Protein truncation
Single point mutation
ProtGPS
Mutant proteins from the ClinVar database

b


Pathogenic mutations alter compartment and protein information content
Change in compartment entropy
Mutant – Wild Type
Single point
Truncation
Count
Single point
Truncation
Wasserstein Distance
(p(WT), p(Mutant))

c


BRD3 (F334S)
SRSF2 (P95H)
ESRP1 (L259V)
BCL6 (R594Q)
Wild type
Disease

d

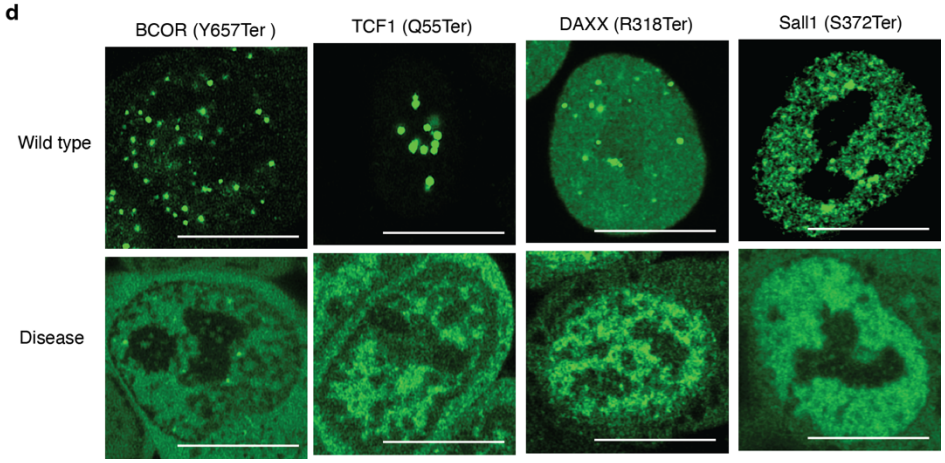

BCOR (Y657Ter)
TCF1 (Q55Ter)
DAXX (R318Ter)
Sall1 (S372Ter)
Wild type
Disease

Figure 3. Pathogenic mutations are predicted to alter protein compartmentalization. a. Schematic of information flow, pathogenic ClinVar mutants caused by single point or truncation mutations were classified with ProtGPS to determine if the detected chemical code was changed in the pathogenic variant. **b.** (*Left*) Dot plot showing the Shannon entropy change in compartment prediction due to single point or truncation mutation. (*Right*) Histogram showing the Wasserstein distance between the wild-type and mutant protein compartment probabilities. **c**. Live cell images of mESCs ectopically expressing wild type and single point mutation pathogenic variants fused to GFP as indicated, scale 10 microns. **d.** Live cell images of mESCs ectopically expressing wild type and truncated pathogenic variants fused to GFP as indicated, scale 10 microns.