CELL BIOLOGY

Protein codes promote selective subcellular compartmentalization

Henry R. Kilgore[1]†*, Itamar Chinn[2,3]†, Peter G. Mikhael[2,3]†, Ilan Mitnikov[2,3]†, Catherine Van Dongen[1], Guy Zylberberg[2,3], Lena Afeyan[1,4], Salman F. Banani[1,5], Susana Wilson-Hawken[1,6], Tong Ihn Lee[1], Regina Barzilay[2,3]*, Richard A. Young[1,4]*

Cells have evolved mechanisms to distribute ~10 billion protein molecules to subcellular compartments where diverse proteins involved in shared functions must assemble. In this study, we demonstrate that proteins with shared functions share amino acid sequence codes that guide them to compartment destinations. We developed a protein language model, ProtGPS, that predicts with high performance the compartment localization of human proteins excluded from the training set. ProtGPS successfully guided generation of novel protein sequences that selectively assemble in the nucleolus. ProtGPS identified pathological mutations that change this code and lead to altered subcellular localization of proteins. Our results indicate that protein sequences contain not only a folding code but also a previously unrecognized code governing their distribution to diverse subcellular compartments.

Groups of proteins involved in shared functions must assemble to fulfill their physiological functions (*1*). For example, the fidelity of gene transcription hinges on the assembly of more than a hundred different proteins at regulatory elements (*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 has dominated models of protein assembly, but there is now considerable evidence that large assemblies of proteins with shared functions also occur through weak multivalent noncovalent interactions (*8*–*15*). Nearly all cellular functions involve formation of such assemblies, which have been described as condensates, aggregates, puncta, hubs, and nonmembrane-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*). Integration of contributions from specific interactions (e.g., DNA-protein binding, protein-protein interactions) and nonspecific interactions (e.g., transient noncovalent interactions) is challenging to model, but protein language models provide a means to incorporate diverse contributions. If such a protein language model could be developed, it would have important implications for our understanding of cellular function and dysfunction by providing evidence of a protein code distributed throughout amino acid sequences that can guide selective distribution to subcellular compartments.

[1]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. [3]Abdul Latif Jameel Clinic for Machine Learning in Health, MIT, Cambridge, MA, USA. [4]Department of Biology, MIT, Cambridge, MA, USA. [5]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6]Computational and Systems Biology Program, MIT, Cambridge, MA, USA.
*Corresponding author. Email: hkilgore@wi.mit.edu (H.R.K.); regina@csail.mit.edu (R.B.); young@wi.mit.edu (R.A.Y.)
†These authors contributed equally to this work.

Evidence for shared protein codes in condensate compartments

To learn whether collections of proteins that assemble into specific condensate compartments have shared protein 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 protein codes embedded in the amino acid sequence of a protein. We focused our studies on a set of 5480 human protein sequences that have been annotated for 12 condensate compartments using the UniProt (Universal Protein Resource) (*23*) and CD-CODE (Crowdsourcing Condensate Database and Encyclopedia) (*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. 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 12 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 to 0.95) across the 12 different compartments (Fig. 1D). The performance of the ProtGPS model indicates that it detects patterns in the protein sequence that differentiate these condensate compartments.

We attempted to identify features that might contribute to selective compartmentalization, although extraction of the nonlinear patterns or principles learned by a machine learning classifier is a well-known challenge (*25*), owing in part to neural network architecture, to the complexity of pattern information, and to the lack of "language" to describe learned patterns outside of conventional physicochemical properties. The types of sequence features that enable transit across intracellular membranes were not immediately evident in the sets of proteins that that are found together in these compartments (fig. S1). We did observe that proteins in some compartments shared physicochemical properties such as isoelectronic point (pI) and hydrophobicity (fig. S2 and table S1). We also note that the high performance of the protein language model depended on information learned from inclusion of multiple members of protein families and that when these families were not fully represented in the training set, the performance was only somewhat better than a random forest or linear regression model (fig. S2 and table S1). This suggests to us that inclusion of multiple protein family members is informative in optimizing protein language model performance, although inclusion of this information presents some risk of overfitting. Certain amino acids were more informative to differentiate proteins found in separate compartments (fig. S3). We found little evidence to suggest that a protein distribution code can be represented with a small number of components (fig. S4). We anticipate that advances in machine learning and chemical pattern description will enable additional insights into the features learned by ProtGPS that enable its level of performance.

Guided generation of novel protein sequences for compartment selectivity

To further validate that ProtGPS has learned protein 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 (GS) algorithm (*26*) and generated eight novel proteins designed to assemble selectively into nucleoli (table S2). However, these proteins failed to assemble selectively into nucleoli (fig. S5). The failure of our initial efforts to generate proteins that selectively compartmentalize in nucleoli motivated the design of another approach that might be more successful. With GS and ProtGPS, protein sequences are generated without consideration of the chemical

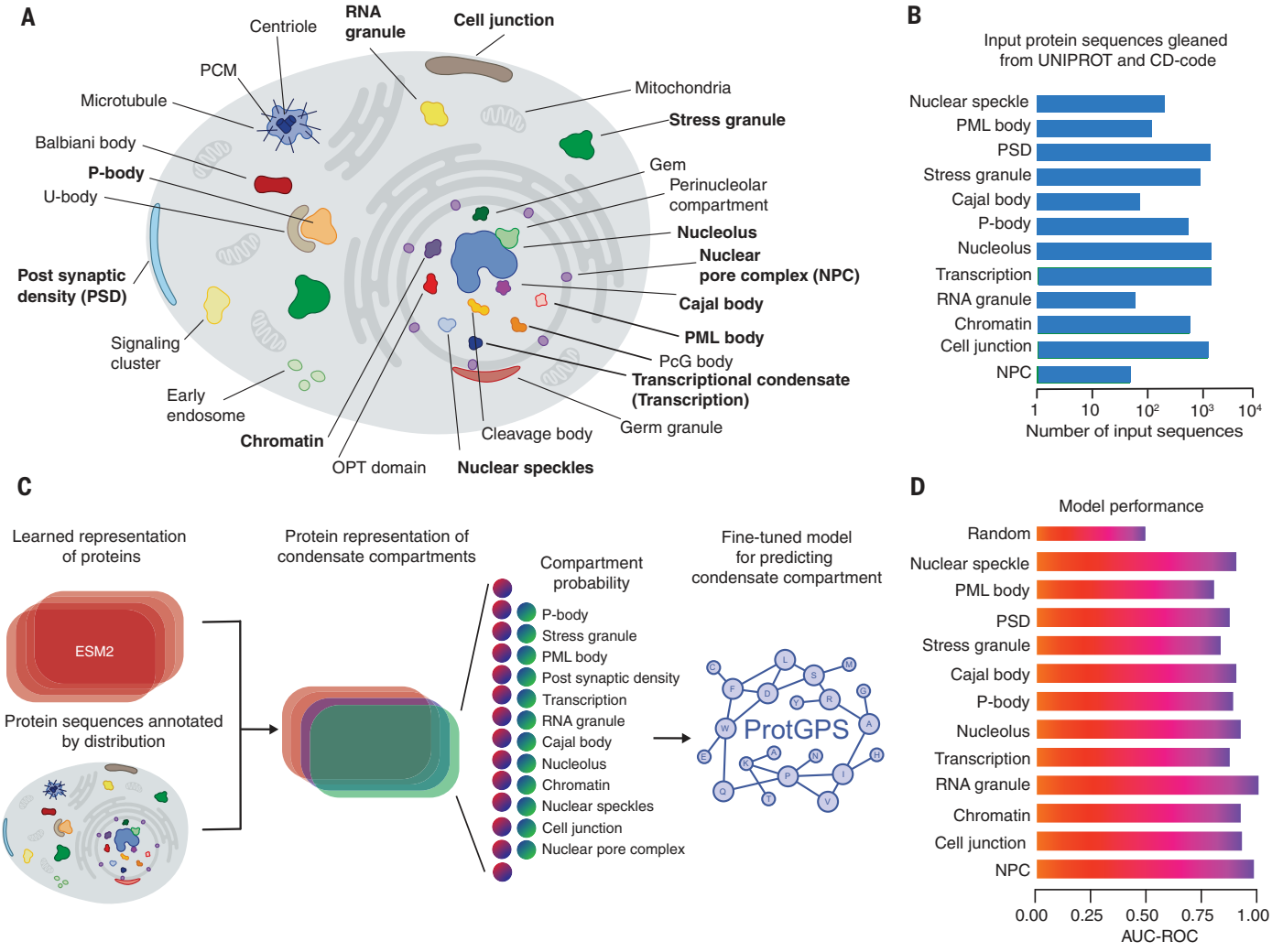

Fig. 1. ProtGPS classifies protein compartment with high performance. (**A**) Graphical depiction of some cellular compartments found in eukaryotic cells; compartments labeled in boldface were studied in this work. OPT domain, Oct1/PTF/transcription domain; P-body, processing body; PcG body, polycomb group body; PCM, pericentriolar material; PML body; promyelocytic leukemia body; U-body, uridine-rich small nuclear ribonucleoprotein–containing body. (**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 (AUC-ROC) for classification of withheld test data (15% of total) with ProtGPS.

space of proteins found in nature. We sought to create an approach that could overcome this limitation by applying a concept borrowed from medicinal chemistry, in which it is common to consider whether a molecule shares desirable physicochemical properties with others (*27*, *28*), namely, sampling from a protein chemical space with specific properties. To apply these concepts toward protein generation, we sought to constrain generation to (i) sequences in the chemical space (*29*) learned by ESM2; (ii) sequences that are intrinsically disordered (*30*), because these are less likely to introduce competing folded states and are associated with condensates (*31*, *32*); and (iii) sequences that should localize to the intended compartment. In practice, this approach integrates the starting protein sequence (mCherry) and its properties into the search for new peptide sequences that are natural, disordered, and have a compartment classification of 0.95 or greater for the target compartment. Thus, we used additional features of protein chemical space and intrinsic disorder for our Markov chain Monte Carlo (MCMC) algorithm (Fig. 2A).

We then used the MCMC algorithm to perform guided generation of proteins that would selectively assemble into a condensate compartment when appended to mCherry protein, which would allow us to follow protein distribution. The chemical properties of mCherry were therefore necessarily integrated into the resulting newly generated protein, which would then allow us to compare partitioning of the new protein with mCherry alone. We first generated proteins that were designed to selectively partition into nucleoli (*9*), which were selected because they are large, well-studied bodies with distinctive morphologies and possess unambiguous marker proteins (Fig. 2A). Ten 100 amino acid–long protein sequences targeted to nucleoli were generated (Fig. 2A, figs. S6 and S7, and tables S3 and S4). 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 was expressed in human cells together with the nucleolus marker NPM1-meGFP, and cells expressing both a test protein (mCherry) and the condensate marker (meGFP) were isolated by using flow cytometry. Imaging of cells revealed that four of 10 proteins designed to assemble into nucleoli (NUC1 to 10) showed readily visible enrichment in

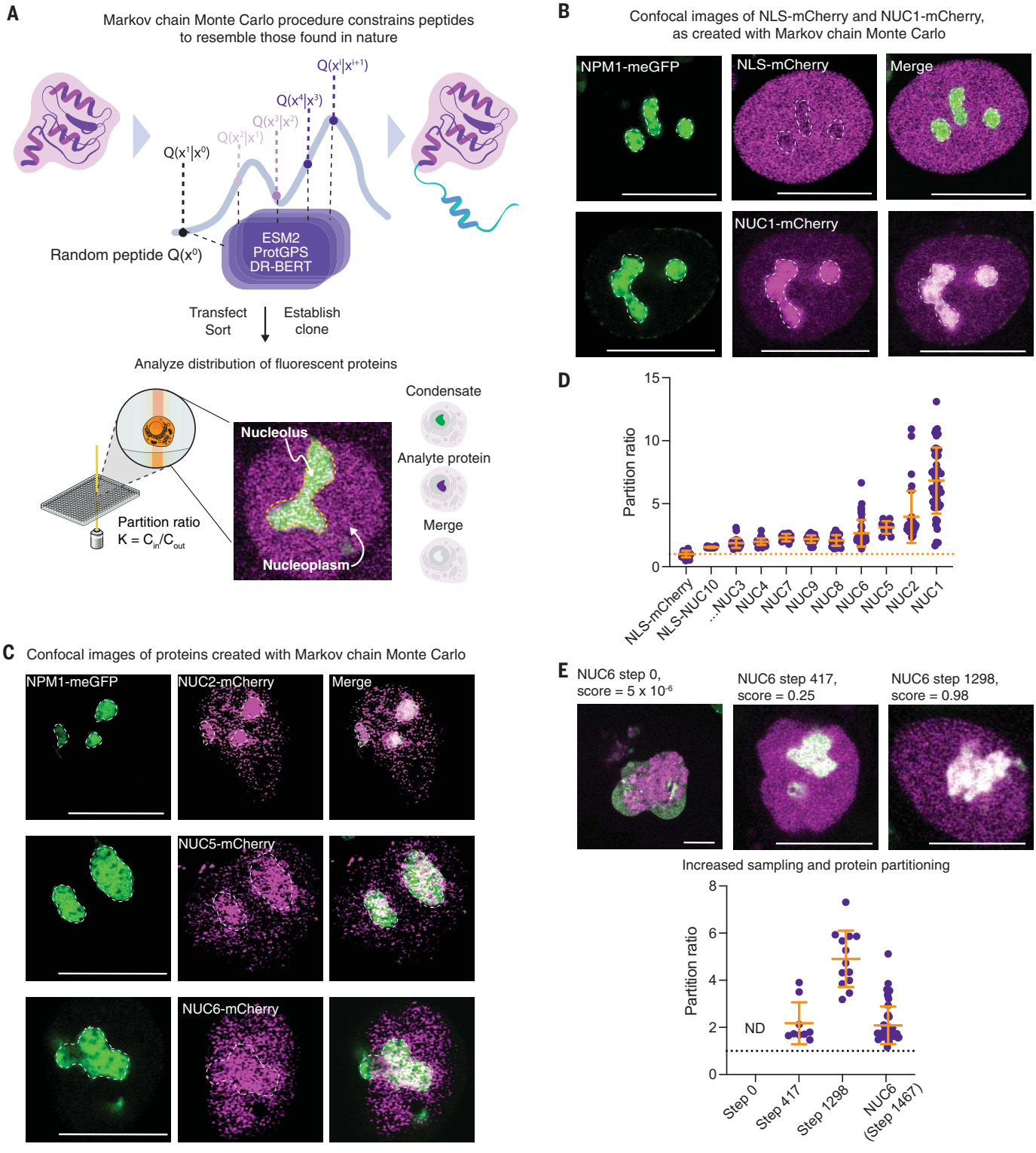

Fig. 2. Generative modeling creates novel proteins that concentrate in a desired condensate. (**A**) Schematic showing the use of Markov chain Monte Carlo to generate proteins and assay them in live cells (MCMC) (more details provided in supplementary materials). (**B**) Live-cell image of a colon cancer cell (HCT-116) tagged at the endogenous NPM1 locus with meGFP and expressing a nucleolus targeted protein NUC1-mCherry. NPM1, nucleophosmin 1; NLS, nuclear localization signal. Scale bars, 10 μm. (**C**) Live-cell confocal micrographs of NUCX-mCherry proteins in HCT-116 cells expressing NPM1-meGFP from the endogenous locus cells. Scale bars, 10 μm. (**D**) Dot plots showing the measured partition ratios of NUCX ($K_x = I_{nucleolus}/I_{nucleoplasm}$) proteins relative to the NLS-mCherry control protein; dotted line is the average value of NLS-mCherry protein (tables S5 and S6 and figs. S8 to S10 provide more information). *K*, partition ratio; $I_{nucleolus}$, intensity of light in the compartment; $I_{nucleoplasm}$, intensity of light in the nucleoplasm. (**E**) Live-cell images and quantification showing the relationship of measured partition ratios ($K_x = I_{nucleolus}/I_{nucleoplasm}$) into the nucleolus by proteins on the NUC6-mCherry trajectory to its computed probability of partitioning.

nucleolar compartments (NUC1, 2, 5, and 6) (Fig. 2, B and C, and figs. S8 to S12), and a more detailed partitioning analysis indicated that the remaining six NUC proteins exhibited more mild enrichment compared with the mCherry control [Fig. 2D; figs. S8 to S12; tables S5 and S6; and supplementary materials (SM), materials and methods].

We next tested the ability of the MCMC algorithm to guide generation of proteins that would partition into nuclear speckles, which are condensates formed by the mRNA splicing apparatus. Using the approach described for the NUC proteins, 10 proteins designed to assemble into nuclear speckles (SPL proteins) were generated and individually expressed in human cells together with SRSF2-meGFP, a marker of nuclear speckles. Imaging of cells revealed that none of the 10 sequences for SRSF2-asociated nuclear speckles became clearly concentrated in nuclear speckles, but two of the generated proteins, SPL2 and SPL3, accumulated in cytoplasmic puncta together with SRSF2-meGFP (figs. S6, S12, and S13; tables S5

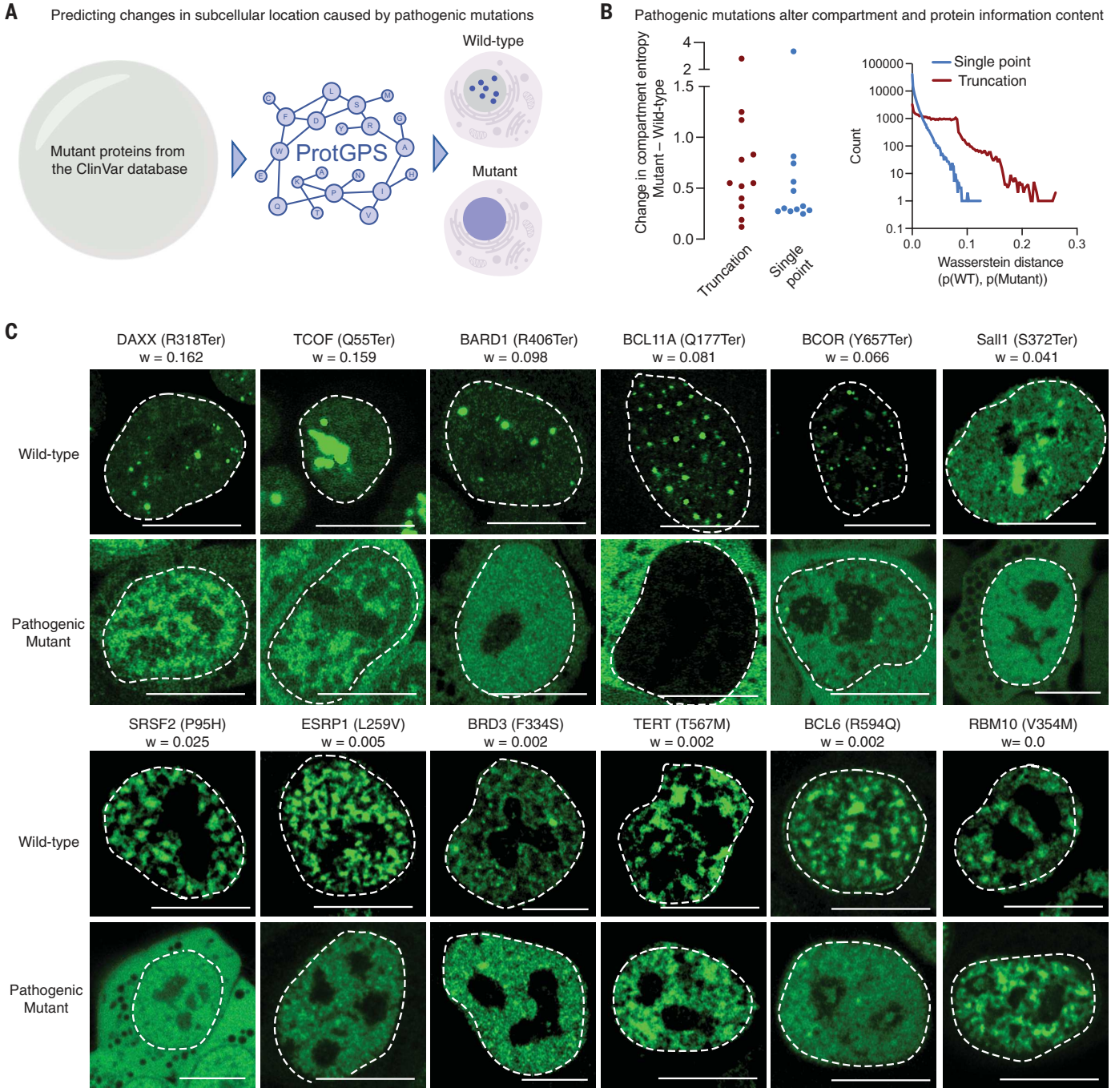

Fig. 3. Pathogenic mutations are predicted to alter protein compartmentalization. (**A**) Schematic of information flow; pathogenic ClinVar mutants caused by single-point or truncation mutations were classified with ProtGPS to determine whether the detected protein code was changed in the pathogenic 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 truncated pathogenic variants fused to meGFP; Wasserstein distance is given for each mutant as w. Scale bars, 10 μm.

and S6; and SM, materials and methods). It thus appears that SPL2 and SPL3 gained the ability to associate with the SRSF2 speckle protein in a cytoplasmic condensate but lost the ability to migrate into the nucleus, where speckles normally form. This behavior is analogous to the effect of mutations in the splicing regulator RBM20, which cause this nuclear speckle protein to accumulate in cytoplasmic puncta and concentrate other splicing proteins (*33*, *34*). These results with NUC and SPL proteins indicate that the MCMC algorithm can guide generation of proteins that selectively partition into a target compartment; but the algorithm was not fully successful in doing so, suggesting that additional training data and analytical approaches will be necessary for improved performance. Sensitivity analysis conducted on the MCMC generative process suggested that increased sampling could lead to improvements in enrichment but also found that the process was nonlinear and can lead to reduced performance, as seen for the final version selected for NUC6 (Fig. 2E and fig. S14). Generative modeling of new protein sequences is a challenging task, whose success rate can vary from <0.01 to ~70% because of the specific modeling goal, the algorithms used to generate protein sequences, and the criteria used to define success or failure (*35*–*38*).

Pathogenic mutations can alter protein codes

Mutations can create pathogenic effects by altering a protein's function or altering a protein's subcellular compartmental distribution. Because ProtGPS can accurately predict the subcellular compartmentalization 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 (*39*) database, a public archive of a vast number of human variations classified for diseases. Data were collected for 205,182 mutations, and ProtGPS was used to predict whether the changes in amino acid sequences alter the subcellular distribution of the mutant proteins (Fig. 3A). We used two approaches, first examining how changes in amino acid sequence affect ProtGPS predictions and then testing experimentally whether mutations predicted by ProtGPS to affect protein distribution can do so.

To characterize the relationship between mutations and changes in ProtGPS predictions, we used approaches applied in information theory. ProtGPS is trained on wild-type sequences and then uses learned patterns to score proteins for their likelihood of distributing to compartments. Mutations affect sequence and can be seen as a change in the information content of the sequence. Any change is thus expected to result in some change in the scoring of mutant protein compared with the wild-type. Furthermore, any changes in scoring are likely to reflect an increase in uncertainty of the prediction because mutations effectively remove information that went into the prediction for the wild-type baseline. To test this, we computed the change in Shannon entropy (*40*, *41*)—an information theory measurement of uncertainty—of the 12 condensate compartments for wild-type versus mutant proteins to ask whether mutations alter the certainty of compartment assignment for a protein (SM, materials and methods). We conducted the analysis for the truncation mutations (83,211)—which we assumed would have major effects—separately from that for the single-point mutations (121,971), which we assumed would have much smaller effects. We found that the Shannon entropy is consistently higher with mutant proteins compared with the normal proteins across all compartments, indicating that mutations are associated with decreased certainty in compartment assignment, with truncations producing larger effects than point mutations (Fig. 3B). A similar analysis was performed for individual proteins; changes in the scores between a wild-type protein and its mutant counterpart can be measured by using Wasserstein distance (*42*–*44*), a metric of dissimilarity between two probability distributions. We found that pathogenic truncation mutations, when compared with single-point mutations, tend to show larger Wasserstein distances (Fig. 3B), but both types of mutations are affecting the scores for compartmentalization. These Wasserstein distances cannot be fully explained by a model of mutations affecting well-recognized features of proteins, such as short linear motifs, residues subjected to posttranslational modifications, or buried residues that might contribute to protein stability (figs. S16 and S20, and tables S7 to S9). These measures 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 test experimentally whether pathogenic mutations predicted by ProtGPS to change protein distribution information content did so, we prepared cells ectopically expressing wild-type and pathogenic mutant proteins tagged with a fluorescent marker protein. We selected for study 20 pathogenic mutations (10 truncation and 10 single-point mutations) in proteins involved in a broad range of biological functions and diseases, whose normal cellular compartmentalization was well-known and that scored across the range of Wasserstein distances (0.162 to 0.000) (table S10). We then generated a panel of cell 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. We noted that signals for all the normal proteins occurred in the subcellular locations where they are known to reside. When comparing images of normal proteins with their mutant counterparts, we found striking differences in compartment appearance for almost all truncation mutation proteins and less-striking but clear differences in compartment appearance for point mutation proteins, except for RBM10 (V354M), which scored with a Wasserstein distance of 0 (Fig. 3C, fig. S21, and table S10). Thus, it appeared that proteins calculated to have a large Wasserstein distance tended to exhibit more dramatic changes in compartment appearance, although this relationship was imperfect (figs. S21 and S22). The effects of truncation mutations on nuclear localization sequences could not account for these results (Fig. 3C, fig. S22, and table S10). These results support the notion that ProtGPS can detect changes in protein codes resulting from 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, ESMfold, and others have learned the relationships between linear amino acid sequence and three-dimensional structure (*22*, *37*, *45*–*49*). Here, we describe ProtGPS, which can predict a protein's selective assembly into specific condensate compartments in cells. ProtGPS with the MCMC algorithm also showed reasonable success in generating novel proteins that selectively partition into the targeted condensate compartments. The complexity of the underlying physicochemical rules for both protein folding and protein localization have proven difficult to parse when 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 already described versions of amino acid codes for some compartments. Blobel and Sabatini proposed a seminal version of amino acid sequence–encoded information with their discovery of a signal peptide sequence for translocation to the endoplasmic reticulum (*50*, *51*). For the membrane-bound nucleus, there are well-known nuclear localization sequences that facilitate the transport of protein from the cytoplasm to the nucleus (*52*–*54*). 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 (*55*–*57*). 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 (*58*–*62*). Disease-related human genetic mutations have been shown to affect protein localization and provide additional experimental evidence for a protein code that contributes to compartmentalization (*62*–*64*). These observations are consistent with the concept of a protein code 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*, *61*, *65*). 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.

Patterns of amino acid sequences that occur in proteins, such as hydrophobic surface patches and blocks of charged residues or repeats, appear overall to be highly constrained in biology (*66*–*72*), 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 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 developing small-molecule therapeutics, designing proteins to preferentially distribute in biochemically relevant regions of the targeted cell population might improve upon their therapeutic properties (*16*, *65*, *73*). In addition, exploring the chemical space of proteins naturally present in specific biological compartments may provide a valuable guide to the generation of optimal chemical matter directed to target proteins in specific compartments. Indeed, there are widely used and efficacious anticancer therapeutics that concentrate in transcriptional condensates at oncogenes (*73*) owing to the chemical environment of those compartments (*16*, *65*). 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, modeling nested compartments, performing large-scale tests of generated proteins, developing robust techniques for measuring compartmentalization in vivo, deploying alternative machine learning approaches, and further exploring the effects of pathogenic mutations.

REFERENCES AND NOTES

1. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
2. S. A. Lambert *et al*., *Cell* **172**, 650–665 (2018).
3. P. Cramer, *Nature* **573**, 45–54 (2019).
4. S. Jena *et al*., *Chem. Soc. Rev.* **51**, 4261–4286 (2022).
5. E. L. Huttlin *et al*., *Nature* **545**, 505–509 (2017).
6. K. Luck *et al*., *Nature* **580**, 402–408 (2020).
7. L. J. Walport, J. K. K. Low, J. M. Matthews, J. P. Mackay, *Chem. Soc. Rev.* **50**, 12292–12307 (2021).
8. Y. Shin, C. P. Brangwynne, *Science* **357**, eaaf4382 (2017).
9. M. Feric *et al*., *Cell* **165**, 1686–1697 (2016).
10. S. Alberti, A. A. Hyman, *Nat. Rev. Mol. Cell Biol.* **22**, 196–213 (2021).
11. J.-M. Choi, A. S. Holehouse, R. V. Pappu, *Annu. Rev. Biophys.* **49**, 107–133 (2020).
12. B. Tsang, I. Pritišanac, S. W. Scherer, A. M. Moses, J. D. Forman-Kay, *Cell* **183**, 1742–1756 (2020).
13. W.-K. Cho *et al*., *Science* **361**, 412–415 (2018).
14. B. R. Sabari *et al*., *Science* **361**, eaar3958 (2018).
15. F. B. Sheinerman, R. Norel, B. Honig, *Curr. Opin. Struct. Biol.* **10**, 153–159 (2000).
16. H. R. Kilgore *et al*., *Nat. Chem. Biol.* **20**, 291–301 (2024).
17. Y. Yu, J. Wang, Q. Shao, J. Shi, W. Zhu, *Sci. Rep.* **6**, 19500 (2016).
18. A. Ben-Naim, *Biopolymers* **29**, 567–596 (1990).
19. A. M. Klibanov, *Nature* **409**, 241–246 (2001).
20. N. Prabhu, K. Sharp, *Chem. Rev.* **106**, 1616–1623 (2006).
21. A. Chandra, L. Tünnermann, T. Löfstedt, R. Gratz, *eLife* **12**, e82819 (2023).
22. Z. Lin *et al*., *Science* **379**, 1123–1130 (2023).
23. UniProt Consortium, *Nucleic Acids Res.* **49**, D480–D489 (2021).
24. N. Rostam *et al*., *Nat. Methods* **20**, 673–676 (2023).
25. S. Kruschel *et al*., Challenging the Performance-Interpretability Trade-off: An Evaluation of Interpretable Machine Learning Models. arXiv:2409.14429 [cs.LG] (2024).
26. J.-E. Shin *et al*., *Nat. Commun.* **12**, 2403 (2021).
27. C. Lipinski, A. Hopkins, *Nature* **432**, 855–861 (2004).
28. M. Beckers, N. Fechner, N. Stiefl, *J. Chem. Inf. Model.* **62**, 6002–6021 (2022).
29. P. Kirkpatrick, C. Ellis, *Nature* **432**, 823 (2004).
30. N. Ananthan, F. John Malcolm, L. Simon, M. Sergei, *Structure* **32**, 1260–1268.e3 (2023).
31. A. S. Holehouse, B. B. Kragelund, *Nat. Rev. Mol. Cell Biol.* **25**, 187–211 (2024).
32. R. van der Lee *et al*., *Chem. Rev.* **114**, 6589–6631 (2014).
33. Y. Zhang *et al*., *JCI Insight* **8**, e170001 (2023).
34. J. Kornienko *et al*., *Nat. Commun.* **14**, 4312 (2023).
35. B. L. Hie *et al*., *Nat. Biotechnol.* **42**, 275–283 (2024).
36. A. H.-W. Yeh *et al*., *Nature* **614**, 774–780 (2023).
37. J. L. Watson *et al*., *Nature* **620**, 1089–1100 (2023).
38. N. R. Bennett *et al*., *Nat. Commun.* **14**, 2625 (2023).
39. M. J. Landrum *et al*., *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
40. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379–423 (1948).
41. A. Lesne, *Math. Structures Comput. Sci.* **24**, e240311 (2014).
42. L. V. Kantorovich, *Manage. Sci.* **6**, 366–422 (1960).
43. C. Villani, in *Optimal Transport: Old and New*, vol. 338 of *Grundlehren der mathematischen Wissenschaften*, A. Chenciner, J. Coates, S.R.S. Varadhan, Eds. (Springer, 2009), pp. 93–111.
44. V. M. Panaretos, Y. Zemel, *Annu. Rev. Stat. Appl.* **6**, 405–431 (2019).
45. J. B. Ingraham *et al*., *Nature* **623**, 1070–1078 (2023).
46. S. Alamdari *et al*., Protein generation with evolutionary diffusion: sequence is all you need. bioRxiv 2023.09.11.556673 [Preprint] (2023). https://doi.org/10.1101/2023.09.11.556673.
47. S. L. *et al*., *Nat. Biotechnol.* **2024** (2024).
48. R. Krishna *et al*., *Science* **384**, eadl2528 (2024).
49. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
50. G. Blobel, D. D. Sabatini, *J. Cell Biol.* **45**, 130–145 (1970).
51. D. D. Sabatini, G. Blobel, *J. Cell Biol.* **45**, 146–157 (1970).
52. E. M. De Robertis, R. F. Longthorne, J. B. Gurdon, *Nature* **272**, 254–256 (1978).
53. C. Dingwall, S. V. Sharnick, R. A. Laskey, *Cell* **30**, 449–458 (1982).
54. J. Lu *et al*., *Cell Commun. Signal.* **19**, 60 (2021).
55. H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, L. A. Royer, *Nat. Methods* **19**, 995–1003 (2022).
56. Y. Jiang *et al*., *Comput. Struct. Biotechnol. J.* **19**, 4825–4839 (2021).
57. V. Thumuluri, J. J. Almagro Armenteros, A. R. Johansen, H. Nielsen, O. Winther, *Nucleic Acids Res.* **50**, W228–W234 (2022).
58. K. L. Saar *et al*., *Nat. Commun.* **15**, 5418 (2024).
59. A. Patil *et al*., *Cell* **186**, 4936–4955.e26 (2023).
60. H. Lyons *et al*., *Cell* **186**, 327–345.e28 (2023).
61. M. R. King *et al*., *Cell* **187**, 1889–1906.e24 (2024).
62. M. A. Mensah *et al*., *Nature* **614**, 564–571 (2023).
63. S. F. Banani *et al*., *Dev. Cell* **57**, 1776–1788.e8 (2022).
64. J. Lacoste *et al*., *Cell* **187**, 6725–6741.e13 (2024).
65. H. R. Kilgore, R. A. Young, *Nat. Chem. Biol.* **18**, 1298–1306 (2022).
66. A. I. Podgornaia, M. T. Laub, *Science* **347**, 673–677 (2015).
67. D. Repecka *et al*., *Nat. Mach. Intell.* **3**, 324–333 (2021).
68. A. J. Faure *et al*., *Nature* **634**, 995–1003 (2024).
69. J. M. Smith, *Nature* **225**, 563–564 (1970).
70. T. Hayes *et al*., *Science* **387**, eads0018 (2025).
71. S. Romero-Romero, S. Lindner, N. Ferruz, *Cold Spring Harb. Perspect. Biol.* **15**, a041471 (2023).
72. H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E. D. Levy, *Nature* **548**, 244–247 (2017).
73. I. A. Klein *et al*., *Science* **368**, 1386–1392 (2020).
74. P. G. Mikhael, H. R. Kilgore, I. Chinn, I. Mitnikov, Code and Data for "Protein Codes Promote Selective Subcellular Compartmentalization," Version v1, Zenodo (2024); https://doi.org/10.5281/zenodo.14795445.
75. H. R. Kilgore *et al*., Data for "Protein Codes Promote Selective Subcellular Compartmentalization," Figshare (2024); https://doi.org/10.6084/m9.figshare.25726581.

ACKNOWLEDGMENTS

We thank C. Lilliehook, A. Dall'Agnese, M. Gallagher, Y. Petri, J. Yang, S. Moreno, and J. Wohlwend for helpful comments, and C. Rausch and Warbler Creative for graphical artwork. **Funding:** This work was supported by NIH GM144283 (R.A.Y.), CA155258 (R.A.Y.), NSF PHY2044895 (R.A.Y.), the St. Jude Transcription Collaborative (R.A.Y), the Whitehead Innovation Initiative (H.R.K., C.V.D., T.I.L., and R.A.Y.), Damon Runyon Cancer Research Foundation Fellowship 2458-22 (H.R.K.), the DTRA Discovery of Medical Countermeasures Against New and Emerging (DOMANE) Threats program (I.C., P.G.M., I.M., and R.B.), the MIT Jameel Clinic for Machine Learning in Health (I.C., P.G.M., I.M., and R.B.), Quanta Computing (I.C., P.G.M., I.M., and R.B.), the Centurion Foundation (I.C., P.G.M., I.M., and R.B.), the Brigham and Women's Hospital Clinical Pathology Residency Program (S.F.B), and NIH National Cancer Institute (NCI) T32 CA251062-02 (S.F.B.). **Author contributions:** Conceptualization: H.R.K., R.A.Y.; Methodology: H.R.K., I.C., P.G.M., I.M.; Investigation: H.R.K., I.C., P.G.M., I.M., C.V.D., S.F.B., L.A., S.W.-H.; Visualization: H.R.K., R.A.Y.; Funding acquisition: R.B., R.A.Y.; Project administration: H.R.K., R.B., R.A.Y.; Supervision: H.R.K., R.B., R.A.Y.; Writing – original draft: H.R.K., I.C., P.G.M., I.M., R.A.Y.; Writing – review & editing: H.R.K., I.C., P.G.M., I.M., T.I.L., R.A.Y. **Competing interests:** R.A.Y. is a founder and shareholder of 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. I.C. and I.M. are founders and shareholders of Voltaris. H.R.K., R.A.Y., I.C., P.G.M., I.M., and R.B. are inventors on patent application 63/634,125 submitted by Whitehead Institute that covers protein codes involved in cellular

distribution. All other authors declare that they have no competing interests. **Data and materials availability:** Code and model weights used in this analysis are available at Zenodo (*74*) and GitHub (https://github.com/pgmikhael/protgps). Source data are available at Figshare (*75*). Reagents used are available upon reasonable request.

SUPPLEMENTARY MATERIALS

Submitted 5 May 2024; resubmitted 7 November 2024
Accepted 28 January 2025
Published online 6 February 2025
10.1126/science.adq2634

Protein codes promote selective subcellular compartmentalization

Henry R. Kilgore, Itamar Chinn, Peter G. Mikhael, Ilan Mitnikov, Catherine Van Dongen, Guy Zylberberg, Lena Afeyan, Salman F. Banani, Susana Wilson-Hawken, Tong Ihn Lee, Regina Barzilay, and Richard A. Young

Science **387** (6738), . DOI: 10.1126/science.adq2634

Editor's summary

Cells contain billions of protein molecules, the amino acid sequences of which encode their functional structures. Deep learning models can now accurately predict such structures from amino acid sequence information. Proteins with shared functions assemble into subcellular compartments to carry out their functions. Kilgore *et al*. have developed a deep learning model called ProtGPS that can predict the compartment localization of proteins based on their sequence. ProtGPS identified disease-associated mutations that change this code and lead to altered subcellular localization of proteins. These results indicate that protein sequences contain both a folding code and a code governing their distribution to subcellular compartments. —Di Jiang