MIT Open Access Articles

Protein codes promote selective subcellular compartmentalization

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,*]

[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]Abdul Latif Jameel Clinic for Machine Learning in Health, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

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

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

[6]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).

Abstract

Cells have evolved mechanisms to distribute ~10 billion protein molecules to subcellular compartments where diverse proteins involved in shared functions must assemble. Here, we demonstrate that proteins with shared functions share amino acid sequence codes that guide them to compartment destinations. A protein language model, ProtGPS, was developed 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 over 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 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 noncovalent 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*). 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.

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 5,480 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. 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 sequence that differentiates these condensate compartments.

    We attempted to identify features that might contribute to selective compartmentalization, although extraction of the non-linear patterns or principles learned by a machine learning classifier is a well-known challenge (*25*), due 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 pI and hydrophobicity (Fig. S2, 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 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, 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 in chemical pattern description will enable additional insights into the features that have been 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 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, where 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 (1) sequences in the chemical space (*29*) learned by ESM2, (2) sequences that are intrinsically disordered (*30*) because these are less likely to introduce competing folded states and are associated with condensates (*31, 32*), and (3) 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 (Table S3, Fig. 2A, Fig S6-7, Table 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 using flow cytometry. Imaging of cells revealed that four of ten proteins designed to assemble into nucleoli (NUC1-10) showed readily visible enrichment in nucleolar compartments (NUC1, 2, 5, 6) (Fig. 2B-C, S8-12), and a more detailed partitioning analysis indicated that the remaining six NUC proteins exhibit more mild enrichment compared to the mCherry control (Fig. 2D, Fig. S8-12, Table S5-6, 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 mRNA splicing apparatus. Using the approach described for the NUC proteins, ten 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 ten 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 (Figure S6, S12-13, Table S5-6, 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 it 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 the process was non-linear and can lead to reduced performance, as seen for the final version selected for NUC6 (Fig. 2E, S14). Generative modeling of new protein sequences is a challenging task whose success rate can vary from less than 0.01% to approximately 70% due to 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 compartment 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 if the changes in amino acid sequences alter the subcellular distribution of the mutant proteins (Fig. 3A). We employed 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 to the wild-type. Furthermore, any changes in scoring are likely to reflect an increase in uncertainty of the prediction, as 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 twelve condensate compartments for wild-type versus mutant proteins to ask if mutations alter the certainty of compartment assignment for a protein (Methods). We conducted this analysis separately for the truncation mutations (83,211), which we assumed would have major effects, from the single point mutations (121,971), which we assumed would have much smaller effects. We find that the Shannon entropy is consistently higher with mutant proteins compared to the normal proteins across all compartments, indicating 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 using Wasserstein distance (*42-44*), a metric of dissimilarity between two probability distributions. We find that pathogenic truncation mutations, when compared to 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 post-translational modifications or buried residues that might contribute to protein stability (Fig. S16-20, Table S7-9). 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 if pathogenic mutations predicted by ProtGPS to change protein distribution information content did so, we prepared cells ectopically expressing wild-type and pathogenic mutant proteins from 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 (0.162-0.000) (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 zero (Fig. 3C, Fig. S21, 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 (Fig. S21-22). The effects of truncation mutations on nuclear localization sequences could not account for these results (Fig. 3C, Figure S22, Table S10). These results support the notion that ProtGPS can detect changes in protein 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, ESMfold, and others have learned the relationships between linear amino acid sequence and 3D structure (*22, 37, 45-49*). We here 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 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, 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 anti-cancer therapeutics that

concentrate in transcriptional condensates at oncogenes (*73*) due 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, Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular and Cell Biology* **18**, 285-285 (2017).
2. S. A. Lambert *et al.*, The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
3. P. Cramer, Organization and regulation of gene transcription. *Nature* **573**, 45-54 (2019).
4. S. Jena *et al.*, Noncovalent interactions in proteins and nucleic acids: beyond hydrogen bonding and π-stacking. *Chemical Society Reviews* **51**, 4261-4286 (2022).
5. E. L. Huttlin *et al.*, Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-509 (2017).
6. K. Luck *et al.*, A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020).
7. L. J. Walport, J. K. K. Low, J. M. Matthews, J. P. Mackay, The characterization of protein interactions – what, how and how much? *Chemical Society Reviews* **50**, 12292-12307 (2021).
8. Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
9. M. Feric *et al.*, Coexisting liquid phases underlie nucleolar subcompartments. *Cell* **165**, 1686-1697 (2016).

10. S. Alberti, A. A. Hyman, Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nature Reviews Molecular Cell Biology* **22**, 196-213 (2021).

11. J.-M. Choi, A. S. Holehouse, R. V. Pappu, Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annual Review of Biophysics* **49**, 107-133 (2020).

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

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

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

15. F. B. Sheinerman, R. Norel, B. Honig, Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology* **10**, 153-159 (2000).

16. H. R. Kilgore *et al.*, Distinct chemical environments in biomolecular condensates. *Nature Chemical Biology* **20**, 291-301 (2023).

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

18. A. Ben-Naim, Solvent effects on protein association and protein folding. *Biopolymers* **29**, 567-596 (1990).

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

20. N. Prabhu, K. Sharp, Protein−Solvent Interactions. *Chemical Reviews* **106**, 1616-1623 (2006).

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

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

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

24. N. Rostam *et al.*, CD-CODE: crowdsourcing condensate database and encyclopedia. *Nature Methods* **20**, 673-676 (2023).

25. S. Kruschel *et al.*, Challenging the Performance-Interpretability Trade-off: An Evaluation of Interpretable Machine Learning Models. *arXiv preprint arXiv:2409.14429*, (2024).

26. J.-E. Shin *et al.*, Protein design and variant prediction using autoregressive generative models. *Nature Communications* **12**, 2403 (2021).

27. C. Lipinski, A. Hopkins, Navigating chemical space for biology and medicine. *Nature* **432**, 855-861 (2004).

28. M. Beckers, N. Fechner, N. Stiefl, 25 Years of Small-Molecule Optimization at Novartis: A Retrospective Analysis of Chemical Series Evolution. *Journal of Chemical Information and Modeling* **62**, 6002-6021 (2022).

29. P. Kirkpatrick, C. Ellis, Chemical space. *Nature* **432**, 823-823 (2004).

30. N. Ananthan, F. John Malcolm, L. Simon, M. Sergei, DR-BERT: A Protein Language Model to Annotate Disordered Regions. *bioRxiv*, 2023.2002.2022.529574 (2023).

31. A. S. Holehouse, B. B. Kragelund, The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* **25**, 187-211 (2024).

32. R. van der Lee *et al.*, Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **114**, 6589-6631 (2014).

33. Y. Zhang *et al.*, Disruption of the nuclear localization signal in RBM20 is causative in dilated cardiomyopathy. *JCI Insight* **8**, e170001 (2023).

34. J. Kornienko *et al.*, Mislocalization of pathogenic RBM20 variants in dilated cardiomyopathy is caused by loss-of-interaction with Transportin-3. *Nature Communications* **14**, 4312 (2023).

35. B. L. Hie *et al.*, Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology* **42**, 275-283 (2024).

36. A. H.-W. Yeh *et al.*, De novo design of luciferases using deep learning. *Nature* **614**, 774-780 (2023).

37. J. L. Watson *et al.*, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023).

38. N. R. Bennett *et al.*, Improving de novo protein binder design with deep learning. *Nature Communications* **14**, 2625 (2023).

39. M. J. Landrum *et al.*, ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, D1062-D1067 (2018).

40. C. E. Shannon, A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379-423 (1948).

41. A. Lesne, Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Mathematical Structures in Computer Science* **24**, e240311 (2014).

42. L. V. Kantorovich, Mathematical Methods of Organizing and Planning Production. *Manage. Sci.* **6**, 366–422 (1960).

43. C. Villani, in *Optimal Transport: Old and New,* C. Villani, Ed. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009), pp. 93-111.

44. V. M. Panaretos, Y. Zemel, Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application* **6**, 405-431 (2019).

45. J. B. Ingraham *et al.*, Illuminating protein space with a programmable generative model. *Nature* **623**, 1070-1078 (2023).

46. S. Alamdari *et al.*, Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.2009.2011.556673 (2023).

47. L. Sidney Lyayuga *et al.*, Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion. *Nature Biotechnology* (2023).

48. R. Krishna *et al.*, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).

49. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

50. G. Blobel , D. D. Sabatini CONTROLLED PROTEOLYSIS OF NASCENT POLYPEPTIDES IN RAT LIVER CELL FRACTIONS : I. Location of the Polypeptides within Ribosomes. *Journal of Cell Biology* **45**, 130-145 (1970).

51. D. D. Sabatini , G. Blobel CONTROLLED PROTEOLYSIS OF NASCENT POLYPEPTIDES IN RAT LIVER CELL FRACTIONS : II. Location of the Polypeptides in Rough Microsomes. *Journal of Cell Biology* **45**, 146-157 (1970).

52. E. M. De Robertis, R. F. Longthorne, J. B. Gurdon, Intracellular migration of nuclear proteins in Xenopus oocytes. *Nature* **272**, 254-256 (1978).

53. C. Dingwall, S. V. Sharnick, R. A. Laskey, A polypeptide domain that specifies migration of nucleoplasmin into the nucleus. *Cell* **30**, 449-458 (1982).

54. J. Lu *et al.*, Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell Communication and Signaling* **19**, 60 (2021).

55. H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, L. A. Royer, Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature Methods* **19**, 995-1003 (2022).

56. Y. Jiang *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 (2021).

57. V. Thumuluri, J. J. Almagro Armenteros, Alexander R. Johansen, H. Nielsen, O. Winther, DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research* **50**, W228-W234 (2022).

58. K. L. Saar *et al.*, Protein Condensate Atlas from predictive models of heteromolecular condensate composition. *Nature Communications* **15**, 5418 (2024).

59. A. Patil *et al.*, A disordered region controls cBAF activity via condensation and partner recruitment. *Cell* **186**, 4936-4955.e4926 (2023).

60. H. Lyons *et al.*, Functional partitioning of transcriptional regulators by patterned charge blocks. *Cell* **186**, 327-345.e328 (2023).

61. M. R. King *et al.*, Macromolecular condensation organizes nucleolar sub-phases to set up a pH gradient. *Cell* **187**, 1889-1906.e24 (2024).

62. M. A. Mensah *et al.*, Aberrant phase separation and nucleolar dysfunction in rare genetic diseases. *Nature* **614**, 564-571 (2023).

63. S. F. Banani *et al.*, Genetic variation associated with condensate dysregulation in disease. *Developmental Cell* **57**, 1776-1788.e1778 (2022).

64. J. Lacoste *et al.*, Pervasive mislocalization of pathogenic coding variants underlying human disorders. *Cell* **187**, 6725-6741.e6713 (2024).

65. H. R. Kilgore, R. A. Young, Learning the chemical grammar of biomolecular condensates. *Nature Chemical Biology*, 1298-1306 (2022).

66. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673-677 (2015).

67. D. Repecka *et al.*, Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence* **3**, 324-333 (2021).

68. J. F. Andre, M.-A. Aina, H.-C. Cristina, M. S. Jörn, L. Ben, The genetic architecture of protein stability. *bioRxiv*, 2023.2010.2027.564339 (2023).

69. J. Maynard Smith, Natural Selection and the Concept of a Protein Space. *Nature* **225**, 563-564 (1970).

70. T. Hayes *et al.*, Simulating 500 million years of evolution with a language model. *Science*, eads0018 (2025).

71. S. Romero-Romero, S. Lindner, N. Ferruz, Exploring the Protein Sequence Space with Global Generative Models. *Cold Spring Harbor Perspectives in Biology* **15**, (2023).

72. H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E. D. Levy, Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244-247 (2017).

73. I. A. Klein *et al.*, Partitioning of cancer therapeutics in nuclear condensates. *Science* **368**, 1386 (2020).

74. H. R. Kilgore *et al.*, Protein codes promote selective subcellular compartmentalization. *Zenodo*, (2025).

75.   L. N. Randolph, X. Bao, C. Zhou, X. Lian, An all-in-one, Tet-On 3G inducible PiggyBac system for human pluripotent stem cells and derivatives. *Scientific Reports* **7**, 1549 (2017).

76.   D. R. Stirling *et al.*, CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 433 (2021).

77.   R. M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**, 610-621 (1973).

78.   E. Cerami *et al.*, The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2**, 401-404 (2012).

79.   A. P. G. C. The *et al.*, AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discovery* **7**, 818-831 (2017).

80.   P. D. Stenson *et al.*, The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Human Genetics* **139**, 1197-1207 (2020).

81.   R. Kundra *et al.*, OncoTree: A Cancer Classification System for Precision Oncology. *JCO Clinical Cancer Informatics*, 221-230 (2021).

82.   S. Köhler *et al.*, Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research* **47**, D1018-D1027 (2019).

83.   J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, A. Hamosh, OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**, D789-D798 (2015).

84.   K. A. Hoadley *et al.*, Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e296 (2018).

85.   W. J. Kent *et al.*, The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006 (2002).

86.   A. D. Yates *et al.*, Ensembl 2020. *Nucleic Acids Research* **48**, D682-D688 (2020).

87.   M. Griffith *et al.*, CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics* **49**, 170-174 (2017).

88.   D. Chakravarty *et al.*, OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, 1-16 (2017).

89.   M. M. Li *et al.*, Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American

Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics* **19**, 4-23 (2017).

90. S. Richards *et al.*, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**, 405-424 (2015).

91. R. G. H. Lindeboom, F. Supek, B. Lehner, The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature Genetics* **48**, 1112-1118 (2016).

92. A. Henrie *et al.*, ClinVar Miner: Demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. *Human Mutation* **39**, 1051-1060 (2018).

93. M. Bernhofer *et al.*, NLSdb—major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Research* **46**, D503-D508 (2018).

94. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, (2014).

95. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028 (2017).

96. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

97. M. A.J. (GitHub, 2023), vol. 1.2.1.

98. S. A. K. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, Z. Cao, Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics* **8**, 300 (2007).

99. A. McKenna, S. Dubey, Machine learning based predictive model for the analysis of sequence activity relationships using protein spectra and protein descriptors. *Journal of Biomedical Informatics* **128**, 104016 (2022).

100. M. Sundararajan, A. Taly, Q. Yan, in *International conference on machine learning*. (PMLR, 2017), pp. 3319-3328.

101. N. Kokhlikyan *et al.*, Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, (2020).

102. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261-272 (2020).

103. L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, (2018).

104. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

105. N. Halko, P.-G. Martinsson, J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* **53**, 217-288 (2011).

106. V. Robert *et al.*, Language models generalize beyond natural proteins. *bioRxiv*, 2022.2012.2021.521521 (2022).

107. G. Peyré, M. Cuturi, Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning* **11**, 355-607 (2019).

108. M. Kumar *et al.*, ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Research* **52**, D442-D455 (2024).

109. R. Wang, M. G. Brattain, The maximal size of protein to diffuse through the nuclear pore is larger than 60 kDa. *FEBS Letters* **581**, 3164-3170 (2007).

110. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).

111. M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, C. O. Wilke, Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLOS ONE* **8**, e80635 (2013).

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.

Supplementary materials:
Materials and Methods
Supplementary Text
Fig. S1 to S22
Tables S1 to S10
References (75-111)
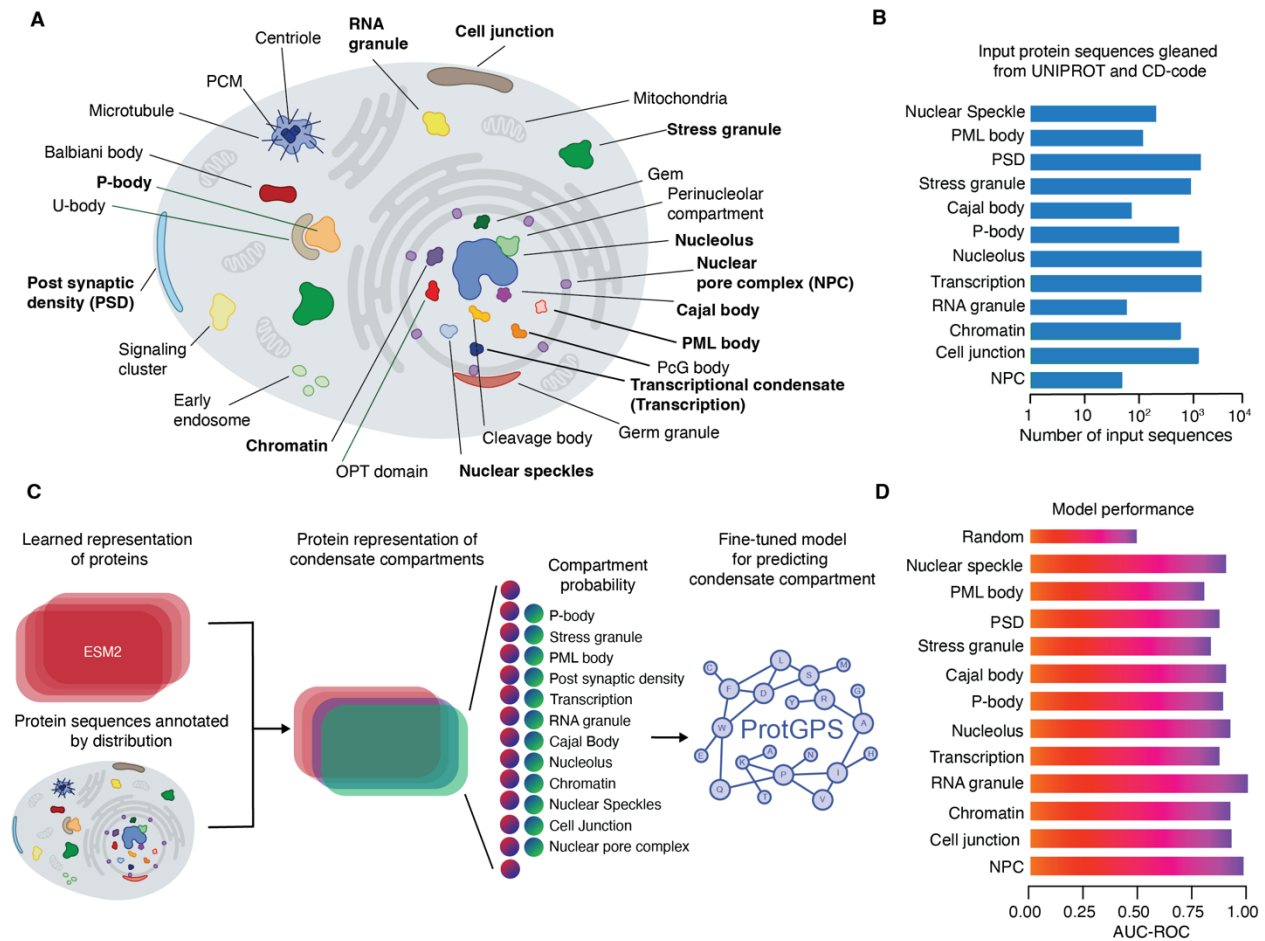
Fig 1. ProtGPS classifies protein compartment with high performance. A.
Graphical depiction of some cellular compartments found in eukaryotic cells,
compartments in bold were studied in this work. **B**. Bar graph showing the number
of protein sequences gathered from UniProt and the CD-code database 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.

A Markov chain Monte Carlo procedure constrains peptides to resemble those found in nature

$Q(x^1|x^0)$

$Q(x^1|x^0)$ Random peptide $Q(x^0)$

ESM2 ProtGPS DR-BERT

$Q(x^2|x^1)$ $Q(x^3|x^2)$ $Q(x^4|x^3)$ $Q(x^i|x^{i+1})$

Transfect Sort Establish clone

Analyze distribution of fluorescent proteins

Condensate

Analyte protein

Merge

Nucleolus
Nucleoplasm

Partition ratio $K = C_{in}/C_{out}$

B Confocal images of NLS-mCherry and NUC1-mCherry, as created with Markov chain Monte Carlo

NPM1-meGFP | NLS-mCherry | Merge
NUC1-mCherry

C Confocal images of proteins created with Markov chain Monte Carlo

NPM1-meGFP | NUC2-mCherry | Merge
NUC5-mCherry
NUC6-mCherry

D Partition ratio

NLS-mCherry, NLS-NUC10, ...NUC3, NUC4, NUC7, NUC9, NUC8, NUC6, NUC5, NUC2, NUC1

E

NUC6 step 0, score = 5 x 10^{-6}   NUC6 step 417, score = 0.25   NUC6 step 1298, score = 0.98

Increased sampling and protein partitioning

Partition ratio

ND

Step 0, Step 417, Step 1298, NUC6 (Step 1467)

19

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) (*see supporting information for more details*). 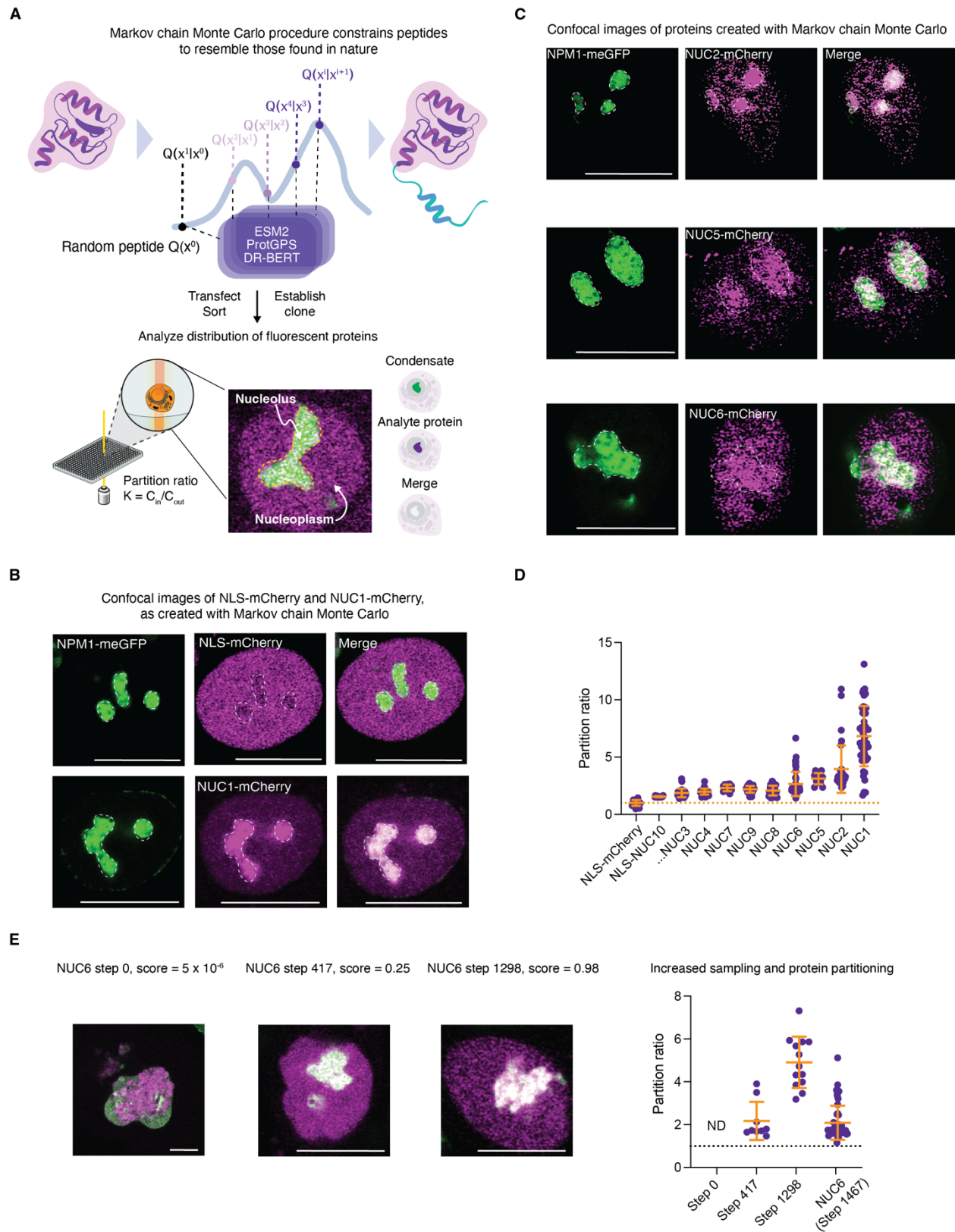**B.** Live cell image of a colon cancer cell (HCT-116) tagged at the endogenous NPM1 locus with meGFP and expressing a nucleolus targeted protein NUC1-mCherry, scale: 10 microns**. C.** Live cell confocal micrographs of NUCX-mCherry proteins in HCT-116 cells expressing NPM1-meGFP from the endogenous locus cells, scale: 10 microns. **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. See Tables 5-6 and Fig. S8-10 for more information. **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.
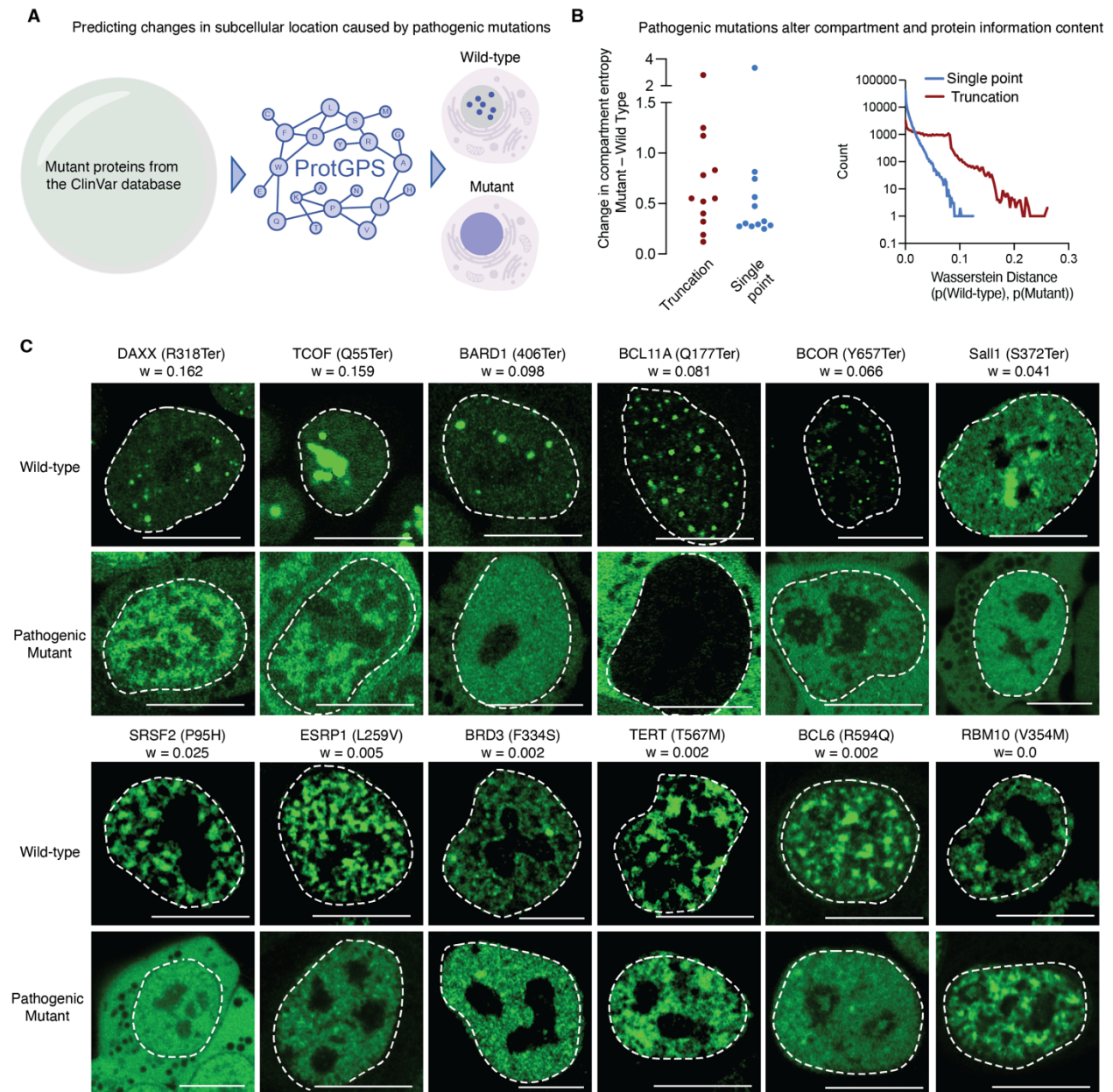
Fig 3. Pathogenic mutations are predicted to alter protein compartmentalization. A. Schematic of information flow, pathogenic ClinVar mutants caused by single point or truncation mutations were classified with ProtGPS to determine if the detected protein code was changed in the pathogenic variant. **B.** (*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 10 microns.

Science

AAAS

Supplementary Materials for

Protein codes promote selective subcellular compartmentalization

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 F. 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]Abdul Latif Jameel Clinic for Machine Learning in Health, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

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

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

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

Corresponding author. Email: hkilgore@wi.mit.edu (H.R.K), regina@csail.mit.edu (R.B.), young@wi.mit.edu (R.A.Y).

The PDF file includes:
Supplementary materials:
Materials and Methods
Text
Figs. S1 to S22
Tables S1 to S10
Supplemental references 75-111

Materials and Methods

Cloning

Gene fragments were codon optimized for humans and purchased from Integrated DNA Technologies (IDT). Gene fragments were assembled into destination plasmids using the NEBuilder® HiFi DNA assembly kit with a molar ratio of 3:1 insert:back bone DNA (New England Biolabs, E5520S). Double stranded DNA products for CRISPR/Cas9 guide RNAs were purchased as monomer oligos from IDT, annealed in 10 mM Tris, 1 mM EDTA, 50 mM salt, heated to 95°C, and allowed to cool until reaching 25°C, over 25 minutes. Assembled duplexes were then ligated using the Quick Ligation™ Kit (New England Biolabs, M2200S) following the standard protocol provided with the product. Chemically competent *E. coli* cells were allowed to incubate on ice for 20 minutes with plasmids, heat shocked at 45°C for 30 seconds, before recovering on ice for 5 minutes. The transformed bacteria were then allowed to recover or SOC outgrowth medium (New England Biolabs, B9020S) for 1-2 hours, diluted 1:10, and 50 µL was spread over a 2% agar plate containing the appropriate antibiotic selection marker (Ampicillin, 100 µg/mL).

Transformed bacteria on antibiotic selection plates were then allowed to incubate overnight at 37°C. Single colonies of bacteria appearing on antibiotic selection plates were used to inoculate 5 mL of LB media with ampicillin (100 µg/mL) to create an overnight culture. Overnight cultures were allowed to incubate at 37°C for 16 hours, cultures were pelleted immediately, and plasmid DNA was purified using a PureLink™ MiniPrep Kit (Invitrogen, K210011). Isolated plasmids were sequenced using whole plasmid sequencing. Restriction digests were performed to generate DNA backbones appropriate for ligation chemistry or Gibson assembly reactions. Backbone plasmids were digested using restriction enzymes AfeI, BsrGI, SpeI (New England Biolabs, R0652L, R3575L, R3133L). Digest products were isolated using DNA gel electrophoresis, using a 120 V potential over 60 minutes as supplied by a Thermo scientific EC300 XL. Agarose gels were created with a 1% solution of SeaKem LEAgarose (Lonza 50004) in Tris-Acetate-EDTA buffer (Millipore-Sigma, T9650) with the addition of ethidium bromide solution 10 mg/mL, to 1 part per 20,000 (Millipore-Sigma, E1510). Gels were imaged using a Biorad-Chemidoc XRST, and bands were excised while wearing UV ray eye protection. Relevant bands were isolated from gels after each run and then extracted with a razor blade from the larger gel. Gel chunks containing desired DNA bands were carefully weighed and extracted using a Monarch DNA gel extraction kit according to the manufacturer's specifications (New England Biolabs, T1020L). An estimate of DNA concentration was collected from an absorbance reading for double stranded DNA using a Nanodrop one^C (Thermo scientific, ND-ONEC-W) and product was stored at -20°C.

Protein design and testing

Proteins designed in our assays consisted of 3 (NLS-mCherry) or 4 components (all other protein sequences). These components were arranged in the following order: SV-40 NLS signal ('PKKKRKV'), an aqueously soluble and flexible linker ('SGSGSG'), a generated protein fragment, and an mCherry protein (see Table S2 and S3 for corresponding sequences). NLS-sequences were attached to each protein fragment in improve the tendency for a protein to accumulate in the target compartment. Every reference to a SPLX or NUCX protein has this specific design construction.

The NLS signal is added to ensure delivery to the nucleus, helping to subsequently assay the ability of the generated sequence to assemble in the target compartment. We note that ProtGPS considers each compartment an independent entity. Characteristics that maximize the probability for one compartment are not expected to have relevance for a different compartment. We would expect that the model would have learned to target sequences to the nucleus and then to a subcompartment only if the characteristics that dictated association with subcompartment were *also* those that dictated association with nucleus.

Democratization of generative modeling strategies and their experimental validation would be enabled by a cost reduction in the infrastructure and reagents required for creating and testing generated information.

Selection of compartments to test success of protein design

Nucleoli and nuclear speckles were chosen for study because these compartments are relatively large, stable and possess readily discernible boundaries. These characteristics make it possible to identify a discrete compartment with confidence using a meGFP-tagged marker protein, and then to obtain robust measurements of mCherry signal inside and outside as a means to quantify enrichment of the protein. Other condensates are much smaller and more dynamic, making it much more challenging to obtain robust measurements of signal inside and outside. Nonetheless, we did attempt to generate de novo sequences for smaller condensates, such as transcriptional condensates marked by MED1 protein and chromatin compartments marked by the HP1α protein. We found that we could not discern with confidence whether or not there is enrichment of mCherry signal in these small puncta, using Zeiss LSM980 with Airyscan microscopy.

Tumor cell tissue culture

Human colorectal cancer cells (HCT-116 American Tissue Culture Catalog CCl-247TM) and human breast cancer cells (MCF7, American Tissue Culture Catalog HTB-22) were cultured in sterile 10 or 15 cm plates with 15 or 35 mL of DMEM (Gibco, 11965084) media supplemented with 10 % Fetal bovine serum (FBS) (Sigma F2442) and 100 units/mL penicillin (Life Technologies, 15140122), and 100 μg/mL streptomycin (Life Technologies, 15140122). Cells were cultured at 37 °C and 5 % v/v CO_2 in a humidified cell culture incubator and passaged at 75 % confluency. Cells were counted to determine seeding density using a CountessTM II automated cell counter, employing trypan blue and disposable countess chamber slides according to manufacturer recommendations. Cells were tested regularly for mycoplasma using the MycoAlert Mycoplasma Detection Kit (Lonza LT07-218) and found to yield negative results. HCT-116 cells expressing NPM1-, and SRSF2-meGFP from the endogenous gene locus were previously reported.

Stem cell tissue culture

In these studies, we employed V6.5 mouse embryonic stem cells, a kind gift from R. Jaenisch. These cells were authenticated by short tandem repeat (STR) analysis compared to commercially acquired cells with the same name. Cells were passaged every 1–2 days by dissociation using TrypLE Express (Gibco, catalog no. 12604), the dissociation reaction was quenched using serum/LIF medium. Stem cells were cultured in 2i/leukemia inhibitor factor (LIF) medium on tissue culture-treated plates coated with 0.2% gelatin (Sigma, catalog no. G1890) in a humidified incubator at 37 °C and 5% v/v CO_2. Cultured cell lines were tested for mycoplasma regularly using the MycoAlert Mycoplasma Detection Kit (Lonza, catalog no. LT07-218) and found to yield negative results.

The composition of 2i/LIF medium is defined as 3 μM CHIR99021 (Stemgent, catalog no. 04-0004), 1 μM PD0325901 (Stemgent, catalog no. 04-0006) and 1,000 U ml−1 LIF (ESGRO, catalog no. ESG1107) in N2B27 medium.

In these experiments N2B27 medium was defined as follows: DMEM/F12 (Gibco, catalog no. 11320) supplemented with 0.5-fold N2 supplement (Gibco, catalog no. 17502), 0.5-fold B27 supplement (Gibco, catalog no. 17504), 2 mM L-glutamine (Gibco, catalog no. 25030), onefold MEM nonessential amino acids (Gibco, catalog no. 11140), 100 U ml−1 penicillin-streptomycin (Gibco, catalog no. 15140) and 0.1 mM 2-mercaptoethanol (Sigma, catalog no. m7522).

Preparation of Serum/LIF medium used KnockOut DMEM (Gibco, catalog no. 10829) supplemented with 15% FBS (Sigma, catalog no. F4135), 2 mM L-glutamine (Gibco, catalog no. 25030), onefold MEM nonessential amino acids, 100 U ml−1 penicillin-streptomycin, 100 μM 2-mercaptoethanol (Sigma, catalog no. M7522) and 1,000 U ml−1 LIF (ESGRO, catalog no. ESG1107).

Cell line generation

Dox inducible cell lines were generated using the Super Piggybac Transposase Expression Vector (Systems Bioscience, PB210PA-1), in conjunction with protein x-lone (*75*) expression cassettes. These reagents transposed proteins engineered in this study under a TR3GS doxycycline inducible promoter system. Plasmids were combined with Lipofectamine3000 reagent and Optimem media (Invitrogen, L3000015) and added to cells plated the day before at 50,000 cells/mL in DMEM (Gibco, 11965084) supplemented with 10% FBS in either 6-well or 10cm

plates in accordance with manufacturer specifications. After 24 hours, their media was changed to DMEM supplemented with 10% FBS, 100 units/mL of penicillin, 100 units/mL streptomycin, and 1000-2000 ng/μL doxycycline hyclate (Millipore-Sigma, D9891) in water.

Twenty-four hours after induction with doxycycline, cells were prepared for sorting by washing cells twice with 10 mL of phosphate buffered saline prior to the addition of 1.5-3 mL of TrypLE to trypsinize the adherent cells for 5-10 minutes at 37 °C. The trypsin reaction was then quenched by the addition of 5 mL of DMEM (Gibco, 11965084) containing 10% FBS, 100 μg/mL of penicillin and streptomycin (Life Technologies, 15140122). Cells were pelleted in 15 mL conical vials at 500 RPM using a table top centrifuge, and resuspended in Dulbecco's phosphate buffered saline containing magnesium and calcium (GibCo 14040117), and filtered into a 5 mL polystyrene round-bottom tube outfit with a cell straining cap (Corning, 352235).

Cells were sorted by flow cytometry as described below for double positives colon cancer (HCT-116) cells expressing green fluorescent protein tagged SRSF2 or NPM1 from the endogenous locus and the target mCherry protein (see Table S1 and S2 for sequences). A homogenous population of cells expressing only SRSF2-meGFP or NPM1-meGFP from the endogenous locus was used as a positive control for meGFP expression and a negative control for mCherry expression. Double positives were collected into 1.5 mL Eppendorf tubes containing 500 μL DMEM containing 10% FBS, 100 μg/mL of penicillin and streptomycin and stored on ice until they could be transferred into 12-well dishes. Sorted cells were cultured for 7 days or until approaching confluency in 12-well dishes.

At approximately 75 % confluency, cells were taken up into solution following the protocol for TrypLE and washes given above, the concentration of cells was established using a CountessTM II automated cell counter, employing trypan blue and disposable countess chamber slides according to manufacturer recommendations. Each population of population of double positive cells was then diluted to 0.85 cells / 100 μL in DMEM containing 10% FBS, 100 μg/mL of penicillin and streptomycin. A multichannel pipette was then used to transfer 100 μL of the diluted cell solution into four 96-well plates and cells were allowed to grow for 7-14 days until single colonies were identified, with media changes occurring on days, 4, 8, and 11. Clonal cell populations were replated in 96-well imaging plates and imaged with confocal microscopy to identify those clonal populations possessing the desired double positive phenotype. Chosen clonal populations were then transferred into 12 wells and allowed to grow to confluency before replating in 10 cm dishes for analysis.

Production of stable mouse embryonic stem cell lines was performed by cloning WT and mutant gene sequences using NEBuilder HiFi DNA Assembly (NEB) into a doxycycline-inducible, N-terminal mEGFP-tagged expression construct with a hygromycin-resistance gene (pbfh-GFP), which was integrated into mESCs using the PiggyBac transposon system (Systems Biosciences). To perform a routine transfection, 0.5×10^6 wildtype mESCs were plated in 6-well format and simultaneously transfected with 1 μg of the expression vector and 1 μg of the PiggyBac transposase using Lipofectamine 3000 (ThermoFisher, L3000001), according to manufacturer instructions in serum/LIF media. The next day, media was changed to 2i, and cells were split into 100 mm gelatin-coated plates with 2i-media supplemented with 500 μg/mL hygromycin (ThermoFisher, 10687010) for selection. Selection media was exchanged every day and un-transfected control cells were monitored to assess selection.

Flow cytometry

Samples were sorted using a BD FACS Aria. Green fluorescent protein and mCherry signal was used to identify colon cancer (HCT-116) that were expressing NPM1 and SRSF2-meGFP fusion proteins in addition to the mCherry proteins incorporated as described in the section "Doxycycline inducible line generation." Double positive cells were collected when both channels had relative signal 10-fold above the background signal produced in the absence of mCherry and within the region defined by the signal found in the meGFP-SRSF2 and meGFP-NPM1 control cell lines. Double positive lines were sorted into 1.5 mL Eppendorf tubes containing in 1 mL of media and stored on ice until plating.

Live cell imaging

Endogenously tagged HCT-116 cells expressing NPM1-meGFP or SRF2-meGFP were seeded at 50,000 cells/mL on an imaging plate to create 3 technical replicates. Imaging plates used were sterile Cellvis 96-well glass (Cellvis,

P96-1.5H-N) bottom plates with #1.5 high performance cover glass (0.17 ± 0.005 mm), or sterile Cellvis 384-well (Cellvis, P384-1.5H-N) glass bottom plates with #1.5 high performance cover glass (0.17 ± 0.005 mm). Cells were plated 48 hours prior to the experiment in DMEM containing 10% FBS, 100 µg/mL of penicillin and streptomycin. Cell lines were induced to express generated protein sequences or NUC1-meGFP charge variants 24 hours before imaging by changing cell media to DMEM containing 2000 ng/µL doxycycline halcylate in water (Millipore-Sigma, D9891) 10% FBS, 100 µg/mL of penicillin and streptomycin. Cells were maintained at 37 °C with 5 % v/v CO_2 in a humidified chamber over the course of the imaging experiment. Experiments were performed at least 3 times on different dates.

Imaging instrumentation

Live cell confocal micrographs were recorded with a Zeiss LSM 980 Airyscan 2 Laser Scanning confocal with a 1.4 NA ×63 Plan Apo objective and running Zeiss Zen Blue v.3.5. Cells were maintained at 37 °C and 5% v/v CO_2 in a humidified chamber throughout the experiment. Images were recorded using 405 nm at 25 mW, 488 nm at 25 mW, 561 nm at 25 mW or 639 nm at 25 mW diode lasers as required.

Imaging data analysis

Our image analysis approach was designed to compute the partition ratio of nucleolus and splicing speckle targeted proteins as compared to the nucleoplasm in each cell. Regions were defined using Zeiss Zen Blue image analysis software. Nucleophosmin (NPM1) is a scaffold and marker of the granular cluster of the nucleolus and serine arginine-rich splicing factor 2 (SRSF2) is a marker for nuclear speckles. Nucleoli and nuclear speckles were identified using the 488 nm excitation band, which could indicate the distribution of nucleophosmin (NPM1)-meGFP and SRSF2-meGFP fusion proteins in the cell. Global threshold-based detection using the following options enabled identification of the nucleolus: a three-sigma threshold approach, a minimum object area of 10 pixels^2 (a size constraint of 50 pixels^2 was used to cull erroneous calls of nuclear speckles), objects were expanded by employing a closed binary criterion, gaussian smoothing, and signal segmentation was performed using watersheds. Regions found outside of the target condensates were identified using Otsu thresholding (light-regions), without gaussian smoothing, object expansion was set to none, and watersheds.

Average signal was computed for inside of the nucleolus ($I_{nucleolus}$) and in the nucleoplasm ($I_{nucleoplasm}$) to compute a partition ratio $K_{nucleolus} = I_{nucleolus} / I_{nucleoplasm}$. Images collected of mCherry signal using the 561 nm excitation laser were analyzed used to calculate, $K_{nucleolus}$, providing the partition ratio of mCherry proteins in regions defined above.

Average signal within splicing speckles and cytoplasmic regions defined by the accumulation of SRSF2-meGFP was computed using I_{SRSF2}. Reference regions, such as the nucleoplasm were using $I_{nucleoplasm}$ or $I_{cytoplasm}$, which was manually defined in Zen blue. Images collected of mCherry signal using the 561 nm excitation laser were analyzed used to calculate, K_{SRSF2}, providing the partition ratio of mCherry proteins in regions defined above.

Images were analyzed to evaluate the correlation of condensate marker protein signal and signal generated from de novo generated protein sequences across the nucleus using Cell Profiler(76) (v.4.2.8). Image textural features(77) used to analyze pathogenic mutant cells (signal homogeneity and entropy) were computed from the gray level co-occurrence matrix as implemented in Cell Profiler (v.4.2.8). Co-occurrence matrices embed information about the signal intensity of pixels relative to each other. Signal homogeneity is measurement of how homogenous a signal is within a defined region of an image; uniform signal has a homogeneity equal to 1. Entropy measurements compute the degree of randomness or order within a defined region of an image; higher values indicate more random signals. Spearman r-correlations and line plots from imaging were computed and displayed with GraphPad Prism (V.10.2.3) from the signal generated from line-plots using Fiji image analysis software.

Statistical analysis of imaging data

Statistical testing of data was performed using unpaired non-parametric t-tests (Kolmogorov-Smirnov test), which were performed using GraphPad Prism (V.10.2.3), as indicated. Kolmogorov-Smirnov tests (KS-test) ask how similar are two different cumulative distributions. In the context of this work, p-values computed with a KS-test ask how significant are the distribution of measurements for each protein's enrichment in a target compartment

compared to a control NLS-mCherry protein's enrichment in a target compartment. P-values, statistical tests, and correlation measurements are reported for select data in Tables S5.

Spearman r correlations were computed along lines intersecting different condensate compartments. Spearman correlations reflect a monotonic relationship between two variables even if that relationship is not linear in nature. Spearman coefficients generalize to non-linear correlations by computing correlations from the rank values of two variables. The reported spearman correlation reflects a typical value for a compartment and is specific to the example provided. For Table S6, single cross-sections of 4 cells were examined for each protein. The average Spearman correlation and its standard deviation are shown in the final column.

Identification of benign, uncertain, and pathogenic variants

Pathogenic mutations were collected from ClinVar database and annotated following the approach of Banani et al. 2022(*39, 78-83*). Variants associated with Mendelian diseases were obtained from HGMD v2020.4(*80*), ClinVar(*39*) and in hg38. AACR Project GENIE v8.1(*79*) and various TCGA(*78*)·(*84*) and TARGET studies via cBioPortal were used to collect cancer variants. (cBioPortal study identifiers:
ucec_tcga_pan_can_atlas_2018, skcm_tcga_pan_can_atlas_2018, coadread_tcga_pan_can_atlas_2018, luad_tcga_pan_can_atlas_2018, stad_tcga_pan_can_atlas_2018, lusc_tcga_pan_can_atlas_2018, blca_tcga_pan_can_atlas_2018, brca_tcga_pan_can_atlas_2018, hnsc_tcga_pan_can_atlas_2018, cesc_tcga_pan_can_atlas_2018, gbm_tcga_pan_can_atlas_2018, lihc_tcga_pan_can_atlas_2018, ov_tcga_pan_can_atlas_2018, lgg_tcga_pan_can_atlas_2018, esca_tcga_pan_can_atlas_2018, prad_tcga_pan_can_atlas_2018, paad_tcga_pan_can_atlas_2018, kirp_tcga_pan_can_atlas_2018, kirc_tcga_pan_can_atlas_2018, sarc_tcga_pan_can_atlas_2018, thca_tcga_pan_can_atlas_2018, acc_tcga_pan_can_atlas_2018, ucs_tcga_pan_can_atlas_2018, laml_tcga_pan_can_atlas_2018, dlbc_tcga_pan_can_atlas_2018, thym_tcga_pan_can_atlas_2018, meso_tcga_pan_can_atlas_2018, kich_tcga_pan_can_atlas_2018, tgct_tcga_pan_can_atlas_2018, chol_tcga_pan_can_atlas_2018, pcpg_tcga_pan_can_atlas_2018, uvm_tcga_pan_can_atlas_2018, wt_target_2018_pub, all_phase2_target_2018_pub, aml_target_2018_pub, nbl_target_2018_pub, and rt_target_2018_pub).

Liftover (*85*) was used to convert the genomic coordinates for different cancer variants from hg19 to hg38. We did not consider deletions larger than 100kb in this analysis. Protein coding sequences changes associated with variants in our study were mapped to the set of 20,394 human proteins using Ensemble VEP v102 and ID mappings between Ensemble and UniProt (*86*). We considered the pathogenic mutations in the context of the canonical isoforms in this study, which represent the best characterized set of isoforms. Isoforms are selected from criteria such as prevalence, similarity to other homologs and without consideration of other information (e.g., sequence length) (*23*). A collection of n= 2,644,688 DNA variants (62% of all variants located within source data sets) were mapped onto the 20,394 canonical protein isoforms found within UniProt. All variant were counted as protein variants—i.e., DNA variants resulting in the same protein-coding alteration on different DNA sequences, were counted as the same. Synonymous variants were excluded from our analysis. For non-synonymous variants, only the primary and most severe protein-coding change associated with a variant was considered based on the established hierarchy of mutation effect severity conveyed by variant annotations in Ensemble.

Mendelian variant pathogenicity was classified from the designations of their clinical significance for ClinVar variants (pathogenic or likely pathogenic) or of variant class for HGMD variants (DM or DM?). Cancer variant pathogenicity was determined by assessment of variants for their inclusion in CIViC (*87*), their inclusion in the list of CGI's validated oncogenic mutations or oncogenicity designation in OncoKB v2.10 (predicted oncogenic, likely oncogenic, or oncogenic) (*88*). Definitions of pathogenicity rely on computation prediction of pathogenicity, but are less dependent upon computation prediction than clinical biological/functional or evolutionary evidence of pathogenicity (*89, 90*).

Among pathogenic mutations, we chose to investigate those that might be readily discernable as influencing structure and assembly. Nonsense and frameshift variants were considered together to be truncating variants and assessed for their predicted propensity to elicit NMD. Predictive rules for NMD were obtained from prior work (*91*). A truncating variant was considered to elicit NMD if the corresponding premature stop codon it introduced occurred (i) >200 residues C-terminal to the start codon; (ii) >50 residues N-terminal to the final exon-exon junction; and (iii) in an exon ≤400 base pairs in length. Mutations were then be subsampled to include only those identified as a single

point or truncation mutations, leading to 205,182 protein sequences. The resulting mutant protein sequences were then classified using ProtGPS.

Benign and uncertain significance mutations were identified using ClinVar miner (*92*), unique variants were filtered by significance labeled as "benign" or "uncertain significance." Only missense mutations causing a single point mutation in the coding region of a protein were included in the set of 26,848 benign and 23,538 uncertain mutations analyzed in this work.

Bioinformatics analysis of nuclear signal peptides

Nuclear export and localization signals identified from UniProt *motifs* possessing a description "Nuclear localization signal." Nuclear localization and nuclear export signals within NLSdb (*93*) were filtered to be comprised of subsequences annotated as "Expert verified", "experimental", or "potential".

Compartment Classification

To train ProtGPS's compartment classifier module, we collected a dataset of 5,480 proteins from UNIPROT and CD-CODE (*23, 24*) covering 12 condensates, consisting of nuclear speckles, p-bodies, PML-bodies, post synaptic densities, stress granules, chromatin, nucleoli, nuclear pore complexes, Cajal bodies, RNA granules, cell junctions, and transcriptional condensates. Explicit incorporation of a nested or hierarchical cellular structure was avoided, as our goal was to learn the discriminating characteristics of condensate compartments. Signal sequences could be included downstream to enable enrichment into a membrane bound compartment. We randomly assigned 70% of protein sequences to training, 15% to development and 15% to test, yielding 3,834, 823 and 823 sequences in each split, respectively. Random assignment of sequences to training, development and test sets is assumed to control for potential biases such as length distribution. Furthermore, the use of a development set helps address concerns of potential overfitting of ProtGPS on patterns found in the protein sequences in the training set. We note that proteins with similar functions (as implied by their common presence in a compartment) may also share sequence homology. This homology, while likely reflective of the underlying biology, may produce a degree of bias in classification performance when examining other related proteins in the test set.

Our model utilizes only the protein sequence to obtain a binary prediction for each of the 12 condensates. Specifically, we initialize a sequence encoder using the protein language model ESM-2 with 8 million parameters (esm2_t6_8M_UR50D) (*22*), and utilize a 2-layer feed-forward neural network with a hidden dimension of 512 as the classifier head. The classifier is implemented with batch normalization. From the ESM-2 model, we obtain a 320-length feature vector per residue. We take the mean embedding across residues to obtain a 320 embedding of the protein sequence, and pass it to the MLP. We train the model end-to-end for 90 epochs in half precision. We use a batch size of 10, an initial learning rate of 0.001, an exponentially decaying learning rate schedule with a decay rate of 0.91, and a dropout rate of 0.1. The model is optimized with the Adam algorithm (*94*) with default parameters. All models are implemented in PyTorch (v2.0.0+cu117) and PyTorch Lightning (v1.6.4).

Compartment classification with clustered train-test split

We consider a train-test split according to sequence similarity and report the performance of a model trained on this data split. We cluster all sequences using MMSeqs2 (*95*) with a 30% sequence identity and 80% coverage, yielding 3,166 clusters. Then, sequences belonging to the same cluster were assigned the same data split yielding 3,834 sequences (2,136 clusters) in the training set, 823 sequences (504 clusters) in the development set, and 823 sequences (526 clusters) in the test set. We train a new model with the same architecture as ProtGPS and hyperparameters on this new dataset.

Compartment classification with physicochemical properties

We evaluate the performance of non-deep learning models on predicting condensate localization, using the same training, development and test sets as those used for ProtGPS (random split of proteins among sets) and for those clustered as described in "*Compartment classification with clustered train-test split*". We train a random forest and a logistic regression model in SciKit-Learn (*96*) (v1.5.0) that receive as input a set of physicochemical features associated with each protein that were calculated with the ProtPy package (*97*) (v1.2.1). Specifically, for each

sequence, we calculate the amino acid composition, and the composition, transition, and distribution (CTD) of the sequence's hydrophobicity, polarity, charge, solvent accessibility, and polarizability(*98, 99*). A description of how ProtPy features were used follows below. We optimized the models on multi-label classification and compute the AUC-ROC for each condensate separately.

The protPy features used are:
1. Amino acid composition: how often each amino acid type appears within the protein sequence

For each of "hydrophobicity", "polarity", "charge", "solvent accessibility", "polarizability", we also calculate the following descriptors:
2. Composition: proportion of the sequence with a particular property. This consists of 3 total features (e.g., for hydrophobicity, this is the fractions of residues that are hydrophobic, neutral or polar).
3. Transition: how often there is a change in particular property along the sequence. This consists of 3 total features (e.g., transition from neutral to polar).
4. Distribution: the percent of the sequence length which contains the first 1%, 25%, 50%, 75%, and 100% of amino acids with a specific property. This consists of 15 (3x5) total features (e.g., the length of the chain needed to capture 75% of hydrophobic residues).

Calculation of attribution scores

We used the trained ProtGPS model to generate attribution scores for amino acids within protein sequences across different compartments. Attribution scores were computed using the Integrated Gradients method (*100*), implemented with the `captum` library version 0.7.0 (*101*). The baseline sequence for generating attributions consisted of mask tokens, ensuring a neutral starting point for comparison. Integrated Gradients interpolated between the masked sequence and the actual sequence, accumulating gradients to highlight the contributions of individual amino acids to model predictions. For each sequence, residue-level attributions were aggregated by compartment for interpretation. To assess the significance of the attribution scores, we calculated p-values using two-sample t-tests. For each amino acid in a specific compartment, its attribution scores were compared to those of the same amino acid across other compartments. The t-statistic and p-values were computed with the `ttest_ind` function from `scipy.stats` version 1.11.2 (*102*).

Latent space analysis of ProtGPS

We used the trained ProtGPS model to analyze the latent space representations of protein sequences by compartment. To visualize the high-dimensional embeddings, we applied the Uniform Manifold Approximation and Projection (UMAP) method (*103*), which provides a two-dimensional projection of the latent space while preserving as much of the original structure as possible. UMAP was implemented using the umap-learn library, version 0.5.7, with default hyperparameters (*104*).

To further investigate how compartmental labels relate to the model's latent structure, we computed the mutual information between these labels and components extracted via singular value decomposition (SVD) (*105*). SVD was applied to decompose the latent space, yielding 320 eigenvectors that capture the modes of variation within the embeddings. We then computed the mutual information between each eigenvector and the compartment labels to assess whether specific components held significant compartment-related information. Mutual information scores were calculated with the mutual_info_score function from the scikit-learn module, version 1.5.2 (*96*), providing a measure of association between the SVD components and compartment labels.

Protein generation with ProtGPS: Autoregressive Greedy Search Generation

Our first attempt at generating proteins possessing chemical codes for different compartments utilized a greedy search algorithm. Given the mCherry sequence, we add to the N-terminus a random subsequence of length $\ell = 150$. This sequence is then iteratively mutated at each position of the subsequence. At each step, we predict the localization of the protein when mutating the current position to all 20 possible amino acids. We keep the top 3 sequences predicted to localize to the desired compartment. For each of those 3 sequences, we repeat the process of mutating the next position (obtaining 3 x 20 sequences) and keeping only the top 3 scoring proteins. Once all ℓ

positions are explored, we choose the single protein most likely to localize to the target compartment among all those generated.

Protein generation with ProtGPS: Markov chain Monte Carlo Generation

We adapt the framework presented in Verkuil et al. (*106*) to generate novel sequences that lead to the localization of mCherry to specific condensates. In particular, we aim to sample sequences x, where the first ℓ amino acids are designed computationally and the rest of the protein corresponds to the mCherry sequence. We guide the generation such that (1) the newly generated subsequence follows the natural distribution of protein sequences, (2) the subsequence is predicted to be disordered and (3) the full protein is predicted to have the desired localization phenotype (e.g., localizing to the nucleolus). We use blocked Gibbs sampling with MCMC (*106*) where we start from a random subsequence, sample a backbone structure y, then update the sequence given the current backbone. This process generates sequences according to the data distribution defined by the proteins that ESM-2 was trained on. Doing so results in sequences that are expected to follow the distribution found in the natural world.

However, our aim is to specifically generate IDRs that are consistent with the chemical space of the condensate we are targeting. To do so, we use ProtGPS to condition the generation process on the likelihood that the full sequence (with mCherry) localizes to the desired condensate. In other words, we sample subsequences from the space of proteins that ProtGPS predicts to localize in our target condensate. To ensure the novel subsequence does not have a definite 3D fold, we use a predictor of protein disorder as a further constraint. Formally, we sample from the joint distribution

$$x, y \sim p(x, y | c = C, d = 1)$$

where c is the condensate compartment we are targeting, and d indicates whether the generated subsequence is disordered. Since the sequence can fully determine structure, the backbone structure $y_{sampled}$ is obtained as in(*106*):

$$y_{sampled} \sim p(y|x)$$

However, we sample a new amino acid sequence (keeping the mCherry sequence fixed) as:

$$x' \sim p\left(x | y = y_{sampled}, c = C, d = 1\right)$$

We consider the likelihoods that a sequence localizes to a specific condensate and that it contains an IDR to be conditionally independent. So, we obtain:

$$p\left(x | y_{sampled}, \ c = C, \ d = 1\right) \propto p\left(c = C | x, y_{sampled}\right) p\left(d = 1 | x, y_{sampled}\right) p\left(x | y_{sampled}\right)$$

Therefore, we add two terms, $E_{condensate}$ and E_{IDR}, to the original energy-based MCMC sampling(*106*):

$$E(x) \ = \ \lambda_p \, E_{projection}(y = Y|x) \ + \ \lambda_{LM} \, E_{LM}(x) \ + \ \lambda_n \, E_{ngram}(x)$$

$$+ \ \lambda_c E_{condensate}(x) + \lambda_{IDR} E_{IDR}(x)$$

where

$$E_{condensate}(x) = -log \ p(c = k \ | \ x)$$

$$E_{IDR}(x) = -\sum_{i=1}^{\ell} \ log \ p(x\_i \ \in IDR|x)$$

Note that we use the full sequence to predict localization, but we calculate disorder only for the first ℓ residues, where ℓ is the length of the IDR we seek to generate.

As in (*106*), we use ESM-2 (esm2_t33_650M_UR50D) for the language model and protein structure samplers. We use ProtGPS to estimate the likelihood the generated sequence localizes correctly ($p(c = k \mid x)$), and the DR-BERT model (*30*) to predict the disorder of each residue ($p(x\_i \in IDR|x)$). We generate sequences of length 100 at the N-terminus of the mCherry protein sequence, keeping the mCherry sequence fixed throughout the process. We set the weights for each energy term as $\lambda_p = 3$, $\lambda_{LM} = 2$, $\lambda_n = 1, \lambda_c = 1, \lambda_{IDR} = 1$. Since we do not intend for the sequence that we generate to have a highly ordered structure, we stop the generation process when the sequence has a likelihood of $p(c = k|x) > 0.85$ for 10 consecutive steps (instead of performing 170,000 MCMC steps). We use a warm-up of 1000 steps. For all other parameters, we use the default values. We use a different seed to initialize each process.

Computation of Wasserstein distance and compartment entropy

To provide a metric for the potential change in compartmentalization that could be attributed to a mutation in a protein, we calculate the Wasserstein distance (*44*) between the predicted scores of the two sequences. For each wild-type protein, ProtGPS produces a set of probabilities for the assignment of a protein to each of the compartments studied here. This set of probabilities is the predicted localization for a sequence by ProtGPS. The process is performed for wild-type and mutant protein sequences providing two sets of probabilities. The distance between the predicted localization of wild-type and mutant protein sequences made by ProtGPS can be computed using the Wasserstein distance (*43, 44*), a distance function from optimal transport (*107*) that computes a distance between the two sets of probabilities. In this application, we note it can be intuitively thought of as the change caused in the protein distribution code due to a specific pathological mutation.

To compute Shannon entropy (*40, 41*) changes for each compartment between wild-type and mutant proteins, we make predictions with ProtGPS that gives a separate set of probabilities for all wild-type and mutant proteins that describes if they would be anticipated to be found in each compartment. For each compartment, the probabilities from every wild-type and mutant protein are then binned to create histograms for wild-type and mutant proteins for each compartment. Those histograms were then used to compute a Shannon entropy for each compartment from the wild-type and mutant protein compartment histograms. Shannon entropy describes the information required (in binary, this is bits of 0 or 1) to represent a "source", here, a source is defined as the histograms constructed from wild-type and mutant protein predictions for each compartment. When two Shannon entropies are compared describing separate states, here, wild-type and mutant proteins, a positive increase in Shannon entropy conveys an increase in the uncertainty where a negative change indicates a decrease in uncertainty.

Frequency of recognized protein features in pathogenic mutations

For frequency of mutations affecting SLiMs, we used a set of over 350 SLiMs annotated in the Eukaryotic linear motif database (*108*) (accession date, July 14, 2024) and mapped their locations on proteins (Supplementary Tables S7,8; Supplementary Figure S17). We then checked the 2,057 pathogenic mutations found in the test set of proteins to see how many would potentially affect one or more SLiMs. We found 15% of pathogenic single point mutations overlap with one or more SLiMs, suggesting that aberrant SLiM-mediated function may explain a small fraction of pathogenic mutations.

For frequency of mutations affecting PTM's, we used a set of 1,127 potential PTM sites identified in the UNIPROT database (*23*) and mapped their locations on the test set proteins (Supplementary Table S9; Supplementary Figure S18). We then checked the 2,057 pathogenic single point mutations found in the test set of proteins to see how many would potentially affect one or more PTM sites. We found 0.97% of pathogenic mutations overlap with a PTM site, suggesting that aberrant PTM-mediated function might explain only a very small fraction of pathogenic mutations in these data.

We reasoned that mutations in buried regions are more likely to contribute to stability defects than mutations in solvent exposed regions due to their increased potential to disrupt the protein's hydrophobic core. For frequency of mutations that occur within buried regions, we asked what percentage of the 2,057 pathogenic mutations have no predicted solvent exposure. Approximately 35% of pathogenic mutations qualify, consistent with the average fraction of amino acids expected to be found in a protein's hydrophobic core (40-50%).

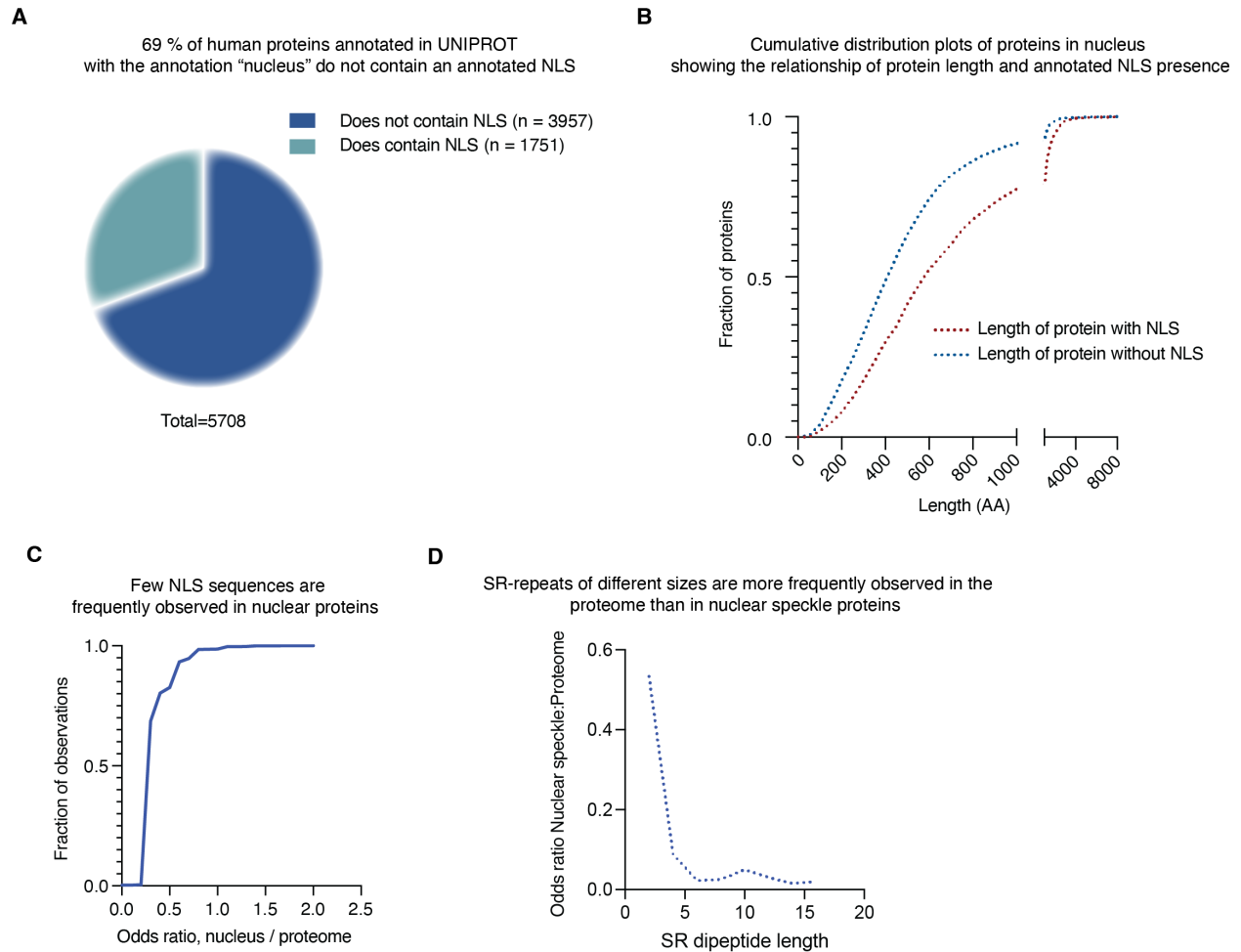Analysis of protein sequences and composition

To determine the tendency for a mutation to occur at disordered or folded domain, we computed a disorder score across every protein in the human proteome using the disorder prediction tool, DR-BERT (*30*). It was then possible to ask if a mutation occurred within an ordered region (DR-BERT score < 0.5) or a disordered region (DR-BERT score > 0.50).

Supplementary Text

<u>Sensitivity analysis of NUC protein sequences</u>

We conducted a sensitivity analysis for the MCMC generative process. In the multistep optimization process for each generated protein, we might expect that continuous improvement in the score computed during the optimization process should reflect the ability to generate proteins with improved compartmentalization phenotypes. As a test of this prediction, we investigated nucleolar partitioning of proteins generated at different steps during the optimization trajectory for NUC1 and NUC6 (Fig. 2E, Fig. S14). Random sequences appended to mCherry, those at step 0, did not show nucleolar compartmentalization. Greater scores produced precursors to NUC1 and NUC6 proteins that tended to show improved nucleolar compartmentalization, although improvement was not continuous (Fig. 2E, Fig. S14). These results suggest sampling for greater periods of time will tend to increase the likelihood of generating protein sequences with desired properties, although this is nonlinear and can lead to reduced performance, as seen for the final version selected for NUC6 (Fig. 2E).

A

69 % of human proteins annotated in UNIPROT
with the annotation "nucleus" do not contain an annotated NLS

Does not contain NLS (n = 3957)
Does contain NLS (n = 1751)

Total=5708

B

Cumulative distribution plots of proteins in nucleus
showing the relationship of protein length and annotated NLS presence

Fraction of proteins

Length of protein with NLS
Length of protein without NLS

Length (AA)

C

Few NLS sequences are
frequently observed in nuclear proteins

Fraction of observations

Odds ratio, nucleus / proteome

D

SR-repeats of different sizes are more frequently observed in the
proteome than in nuclear speckle proteins

Odds ratio Nuclear speckle:Proteome

SR dipeptide length

Supplementary Figure 1. Prevalence of common motifs associated with protein compartmentalization.
Protein sequence features, such as localization signals and SR-dipeptide repeats are typically with the nucleus and nuclear speckles, however, known examples tend to be poorly predictive of those subcellular compartments. **A.** Pie-chart showing the fraction of proteins annotated to reside in the nucleus within the UniProt database that possess an "expert verified", "experimental", or "potential" NLS motif present in the NLS-DB (accession date Feb. 2024). Protein length has been associated with different mechanisms for nuclear import and export (*109*).**B.** Cumulative distribution plot showing the length of proteins annotated to reside in the nucleus within the UniProt database and possess one or more NLS motifs (···), or without an NLS motif (···) identified in the NLS-DB. Statistical testing between cumulative distributions "Length with NLS" and "Length without NLS" was performed with a Kolmogorov-Smirnov test, p-value < 0.0001, KS-statistic = 0.33. **C.** Cumulative distribution plot of the odds ratios for the presence of any NLS-motif studied in Figure S1A in nuclear proteins and the human proteome. SR-dipeptide repeats are associated with a subset of proteins identified in nuclear speckles are more likely to be observed in the proteome than in nuclear speckle proteins. **D.** Plot showing the odds ratios for the presence of SR-dipeptide repeats in nuclear speckles and the human proteome.
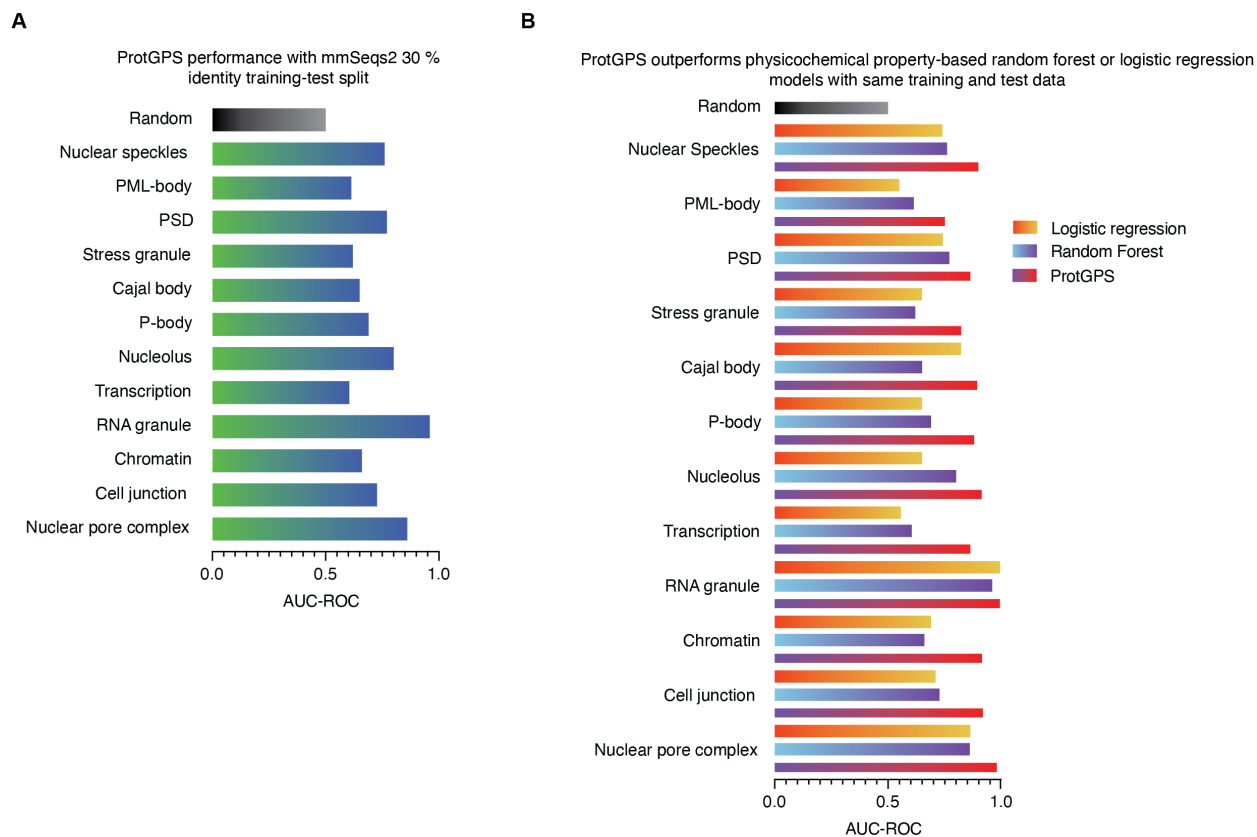
A

ProtGPS performance with mmSeqs2 30 % identity training-test split

B

ProtGPS outperforms physicochemical property-based random forest or logistic regression models with same training and test data

Legend (B):
- Logistic regression
- Random Forest
- ProtGPS

Figure S2. Performance of training data in different strategies. Shared sequence identity and protein physicocemical properties might be expected to be sufficient to predict compartmentalization. For more information on sequence similarity split compartment classification (see *Compartment classification with train-test split*). **A.** Bar-graph showing ProtGPS architecture performance (area under the receiver operator curve) with 30 % sequence identity cluster split, random = 0.5 defines theoretical minimum performance. A random forest and logistic regression model were trained on a physiochemical property-based representation of the same protein sequences used to train ProtGPS to test if this information was sufficient to achieve performance similar to ProtGPS (see *Compartment classification with physicochemical properties* for more information). **B.** Bar graph showing performance (area under the receiver operator curve) of physicochemical property-based model performance with a random forest or logistic regression model (blue gradient, random forest. Sunburst gradient, logistic regression), random = 0.5 defines theoretical minimum performance.
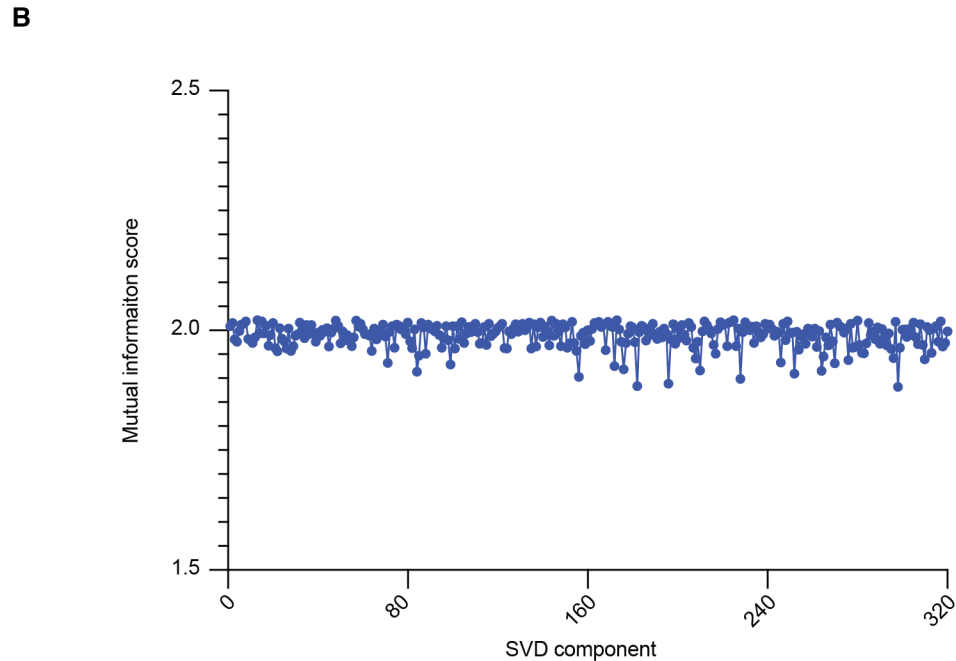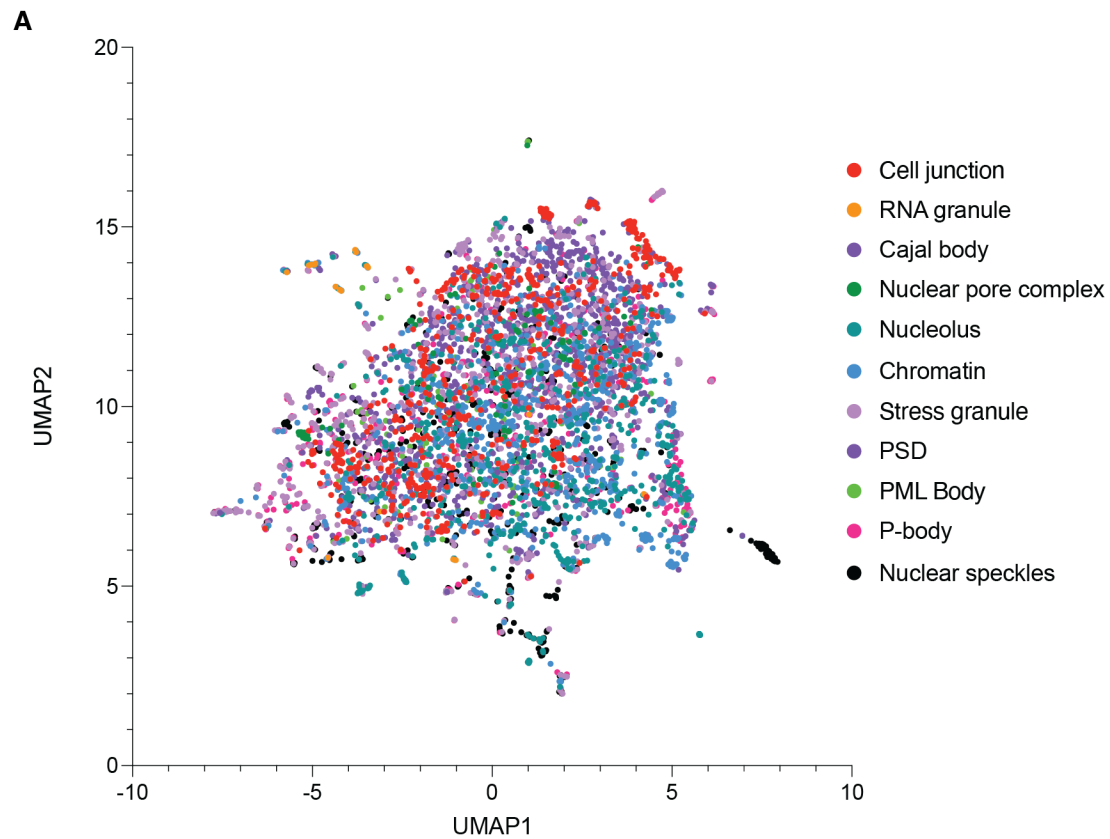
A

Average attribution score of amino acids to compartment classification
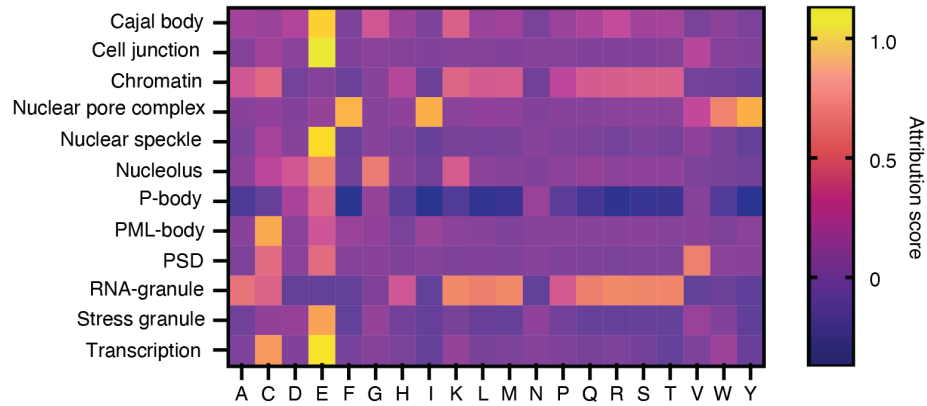
B

Compartments can be clustered and delineated by attribution scores
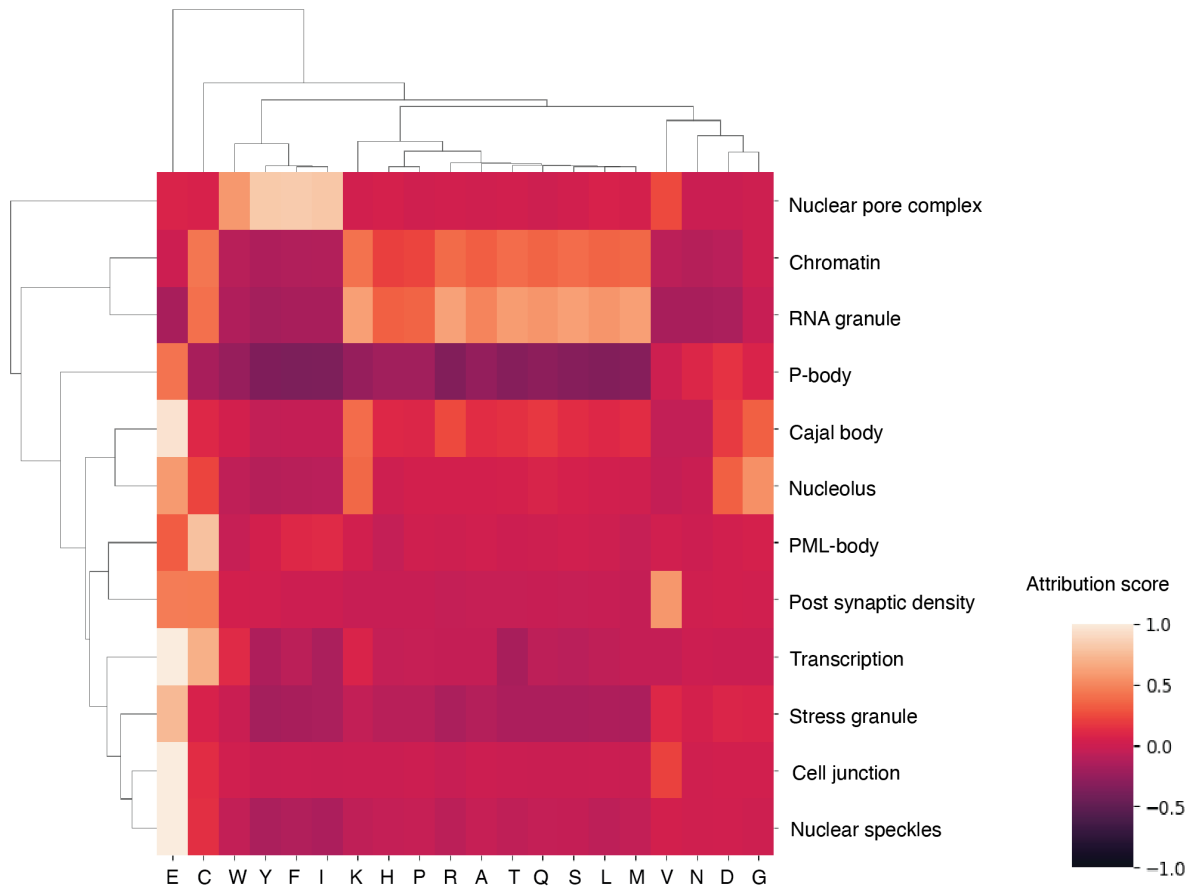
Supplementary Figure 3. Average attribution scores of amino acids for condensate compartment predictions.
A. Heat map showing the attribution of different amino acids to the ProtGPS score, mean attribution score for amino acids normalized by protein sequence length. **B.** Clustering of condensate compartments modeled with ProtGPS by attribution scores of amino acids.

A

Legend:
- Cell junction
- RNA granule
- Cajal body
- Nuclear pore complex
- Nucleolus
- Chromatin
- Stress granule
- PSD
- PML Body
- P-body
- Nuclear speckles

B

Supplementary Figure 4. Investigation of ProtGPS' hidden layers. In principle, investigating the latent space of a neural network can help to reveal relationships between data and provide insight into the performance and behavior of a classifier model. Embeddings for individual proteins from the hidden layers of ProtGPS' neural network architecture were projected onto a 2-dimensional surface using universal manifold projection components (UMAP). However, as multiple layers of a neural network architecture are used to make predictions by ProtGPS, we

might expect the poor separation of proteins. **A**. UMAP projection of the embeddings of proteins used in the training, test, and development of ProtGPS. Mutual information scores can help deduce how, or if, features are related. A singular value decomposition was performed on the latent space of ProtGPS' neural network and SVD components were extracted. A mutual information score was then computed to see if there might exist key eigenvectors contributing to the protein distribution code. **B**. Plot of mutual information score (y-axis) against SVD component, which shows that there are only small differences in the mutual information content of any one component.
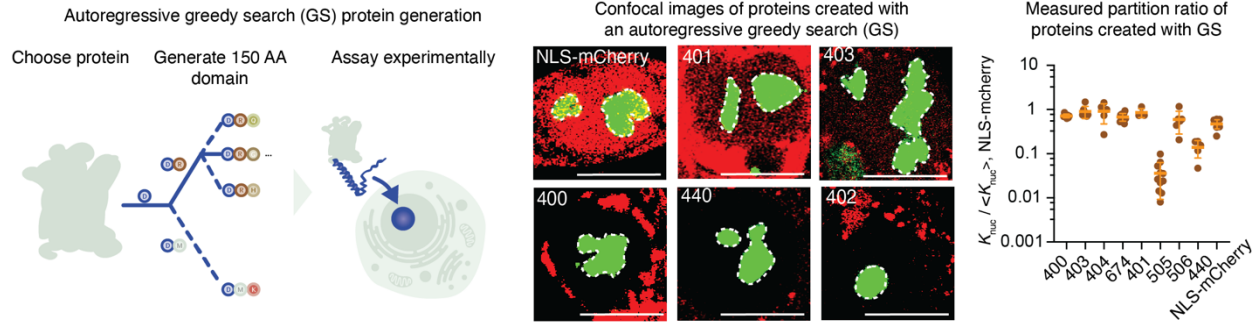
Supplementary Figure 5. Protein generation using an autoregressive greedy search algorithm. (*Left*) Schematic showing the approach to generating proteins using an autoregressive greedy search algorithm guided by ProtGPS. (*middle*) The nucleolus is shown in green (indicated by NPM1-GFP) proteins were generated to target the nucleolus. Confocal micrographs of GS proteins targeted to the nucleolus expressed in colon cancer (HCT-116) cells tagged at the endogenous locus of nucleophosmin (NPM1) with green fluorescent protein (GFP) to indicate the nucleolus (488 nm excitation, green, 561 nm excitation red, overlap, yellow). Dashed lines indicate the perimeter of the nucleolus. scale: 10 microns. (*Right*) Dot plot on a log scale showing the partition ratios of GS proteins in the nucleolus relative to the nucleoplasm ($K = I_{nucleolus} / I_{nucleoplasm}$).

A

B
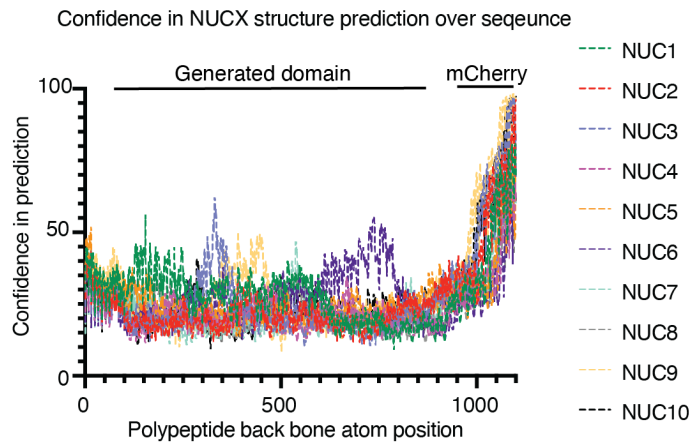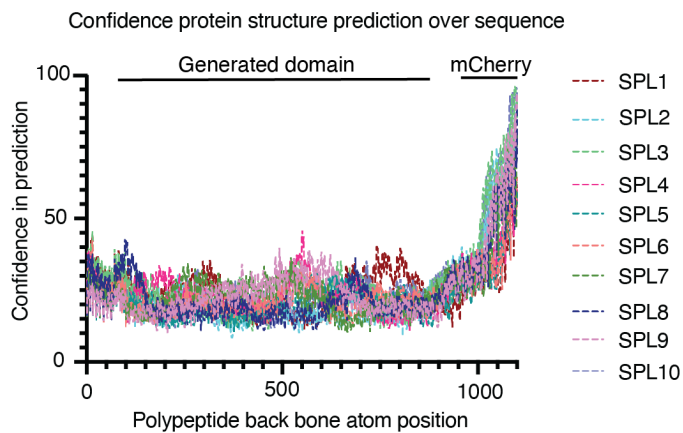
Supplementary Figure 6. Alphafold3 prediction and confidence over sequence appended to the N-terminus of mCherry. Per residue local confidence in prediction (pLDDT) is plotted for newly generated protein sequences over the first 1,100 atoms in the polypeptide backbone of the **A.** NUCX, **B.** SPLX proteins. Confidence is correlated with a tendency to be disordered disordered (*110*). Unique nucleolus or nuclear speckle targeting sequence begins at position 70, ends at position 920).

A

Frequency of SLiMs across all NUCX proteins defined in ELM database

B

NUCX proteins clustered by the presence and absence of different SLiM motifs

Eukaryotic short linear motifs
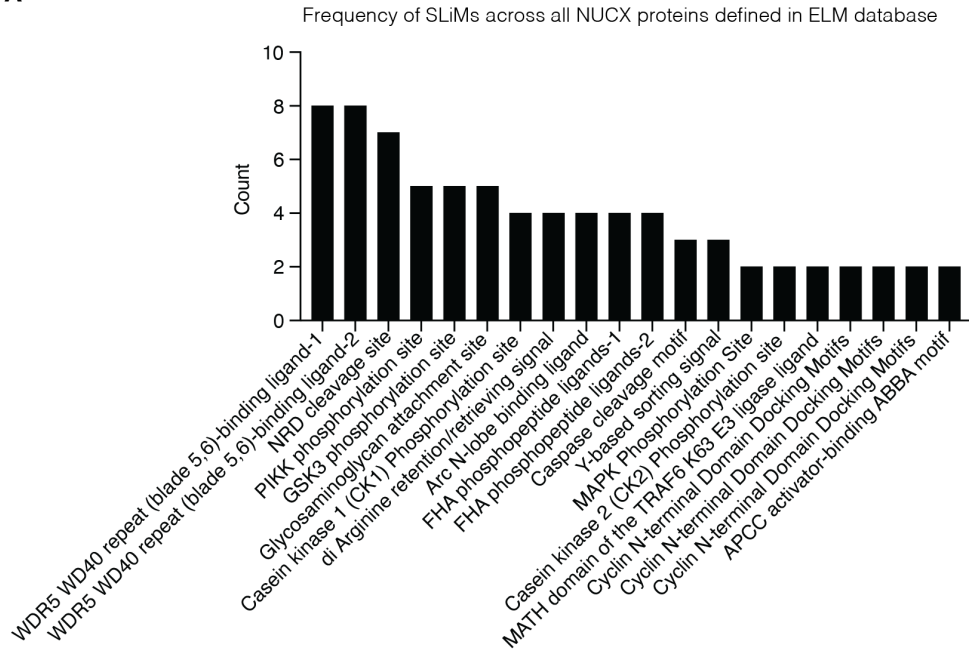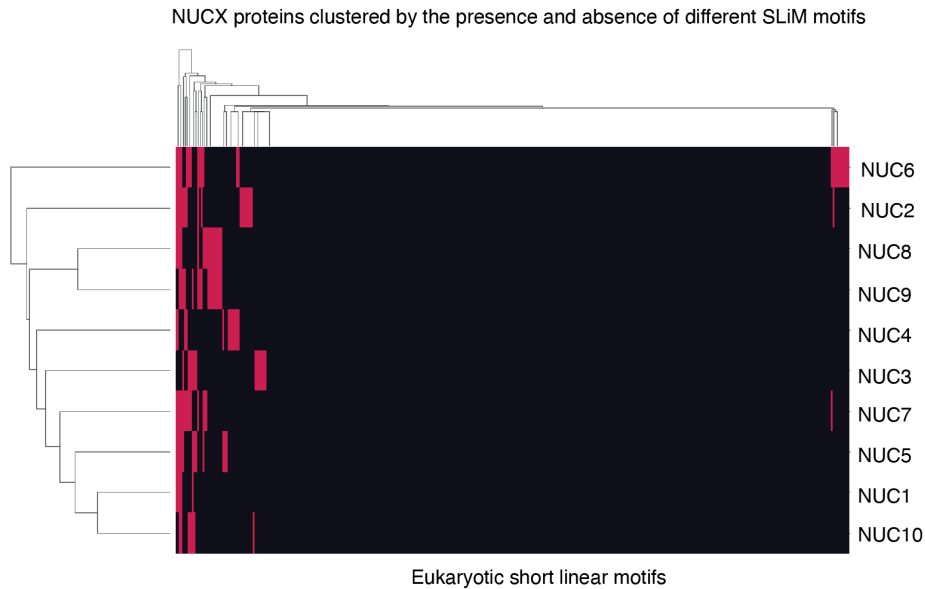
Supplementary Figure 7. SLiM motifs were identified in different NUCX proteins. Short linear motifs (SLiMs) might constitute a subset of the protein distribution code, we looked for the presence of 352 different short SLiMs in the NUCX proteins to look for evidence they're associated with distribution. **A**. Bar-graph showing the count of the top 20 most frequently identified SLiMs in NUCX proteins. **B**. Plot showing the outcome of hierarchical clustering of NUCX proteins by different SLiMs found in their sequences, at most one SLiM of each type was found in each sequence. Red bars, indicate a SLiM is found, black bars indicate the corresponding SLiM is absent. See Table S4 for the list of SLiMs found in NUCX proteins.
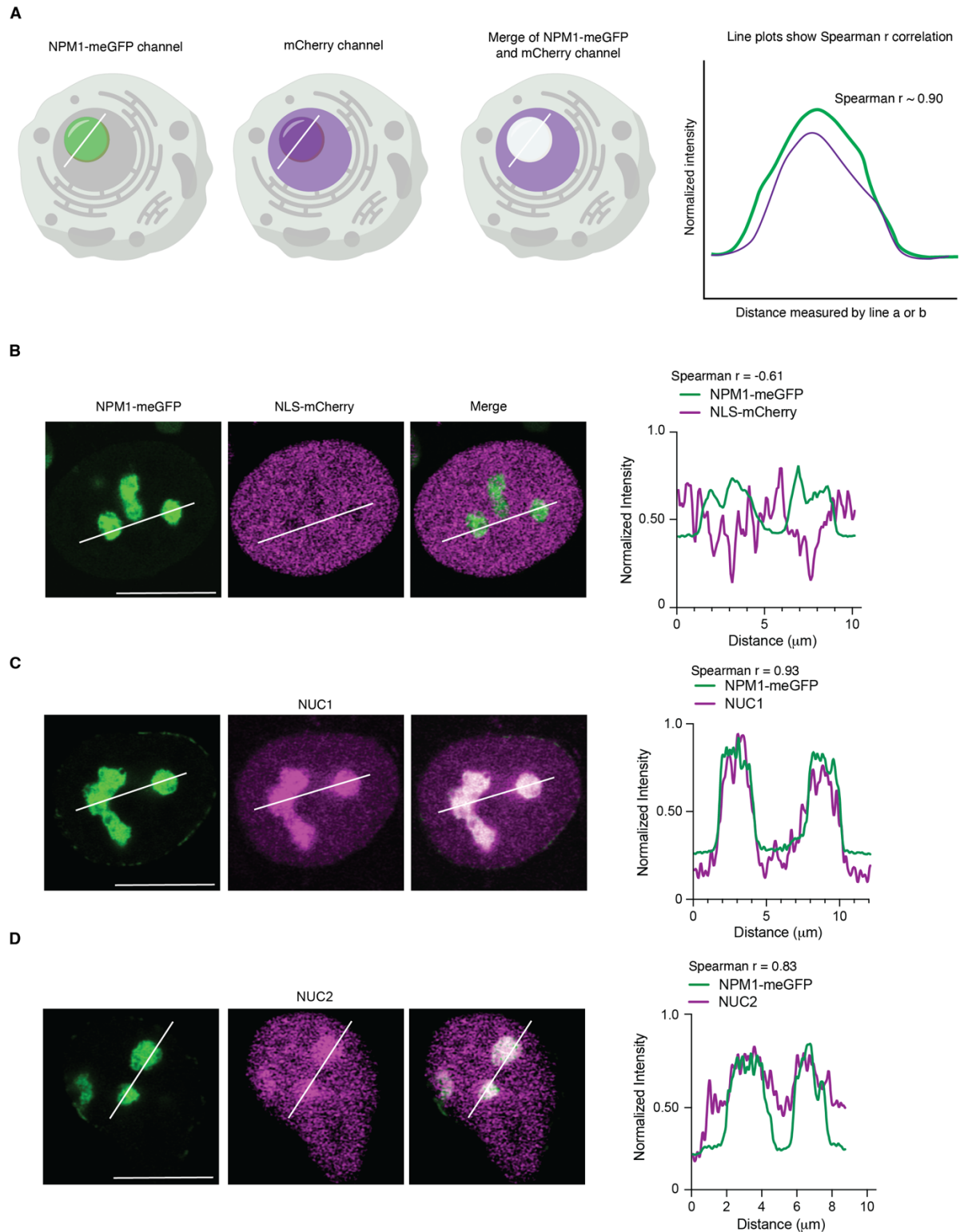
Supplementary Figure 8. Live cell confocal microscopy images of generated proteins targeted to nucleolus. A. Schematic showing how Spearman line-plot analyses were performed in Figures S6-8,12. Live cell images of colorectal cancer (HCT-116) cells expressing NPM1-meGFP (Green) from the endogenous NPM1 locus and induced expression of the indicated nucleolus targeted protein, **B.** Control, NLS-mCherry, **C.** NUC1. **D.** NUC2. Some images are reproduced here from main text Fig. 2C. Scale bar is indicated by white line in bottom left corner, 10 microns.
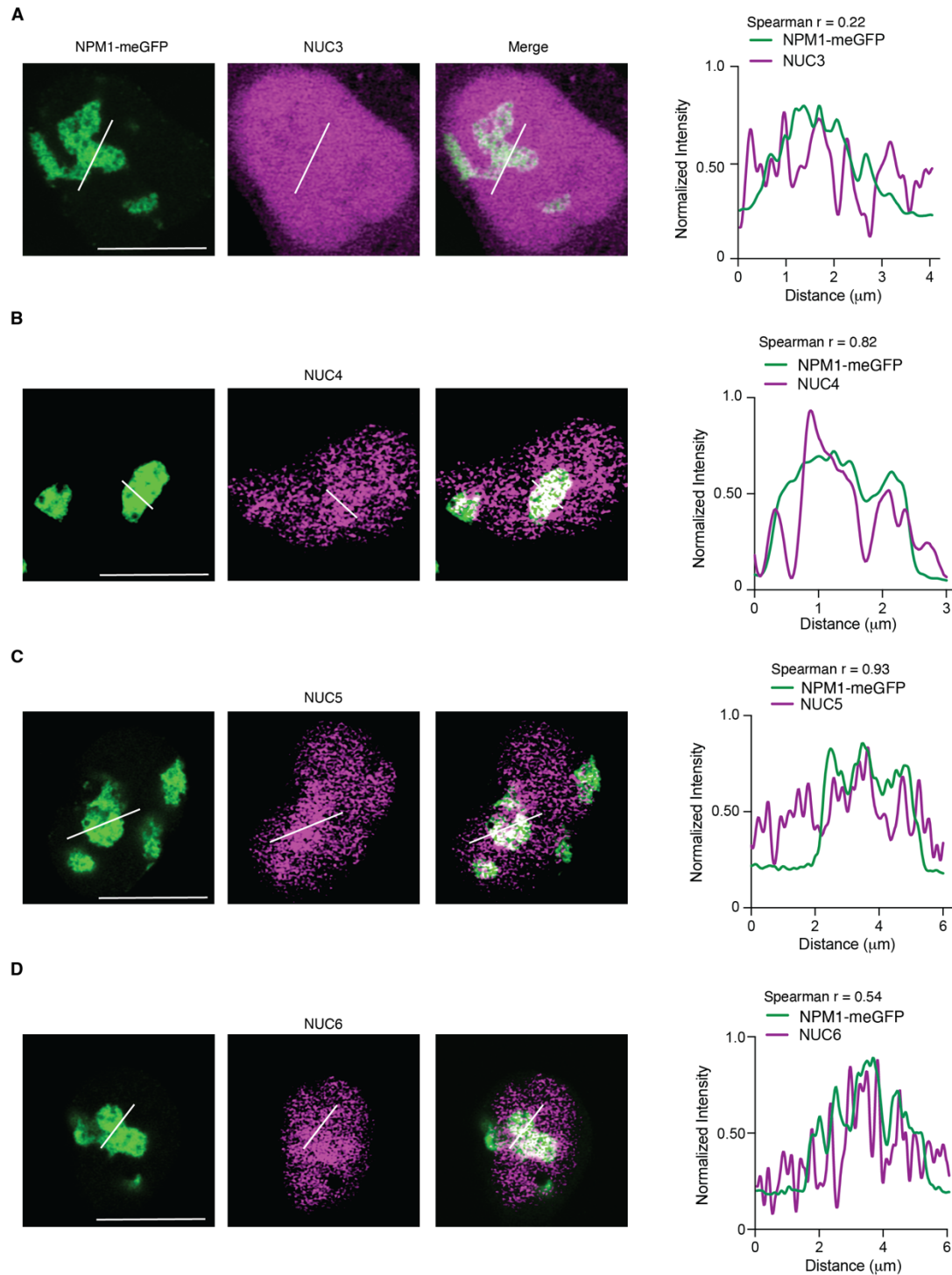
Supplementary Figure 9. Live cell confocal microscopy images of generated proteins targeted to nucleolus. Live cell images of colorectal cancer (HCT-116) cells expressing NPM1-meGFP (green) from the endogenous NPM1 locus and induced expression of the indicated nucleolus targeted protein **A.** NUC3, **B.** NUC4. **C.** NUC5. **D.** NUC6 (magenta). NLS-mCherry control Spearman r correlation = -0.61. Some images are reproduced here from main text Fig. 2C. Scale bar is indicated by white line in bottom left corner, 10 microns.
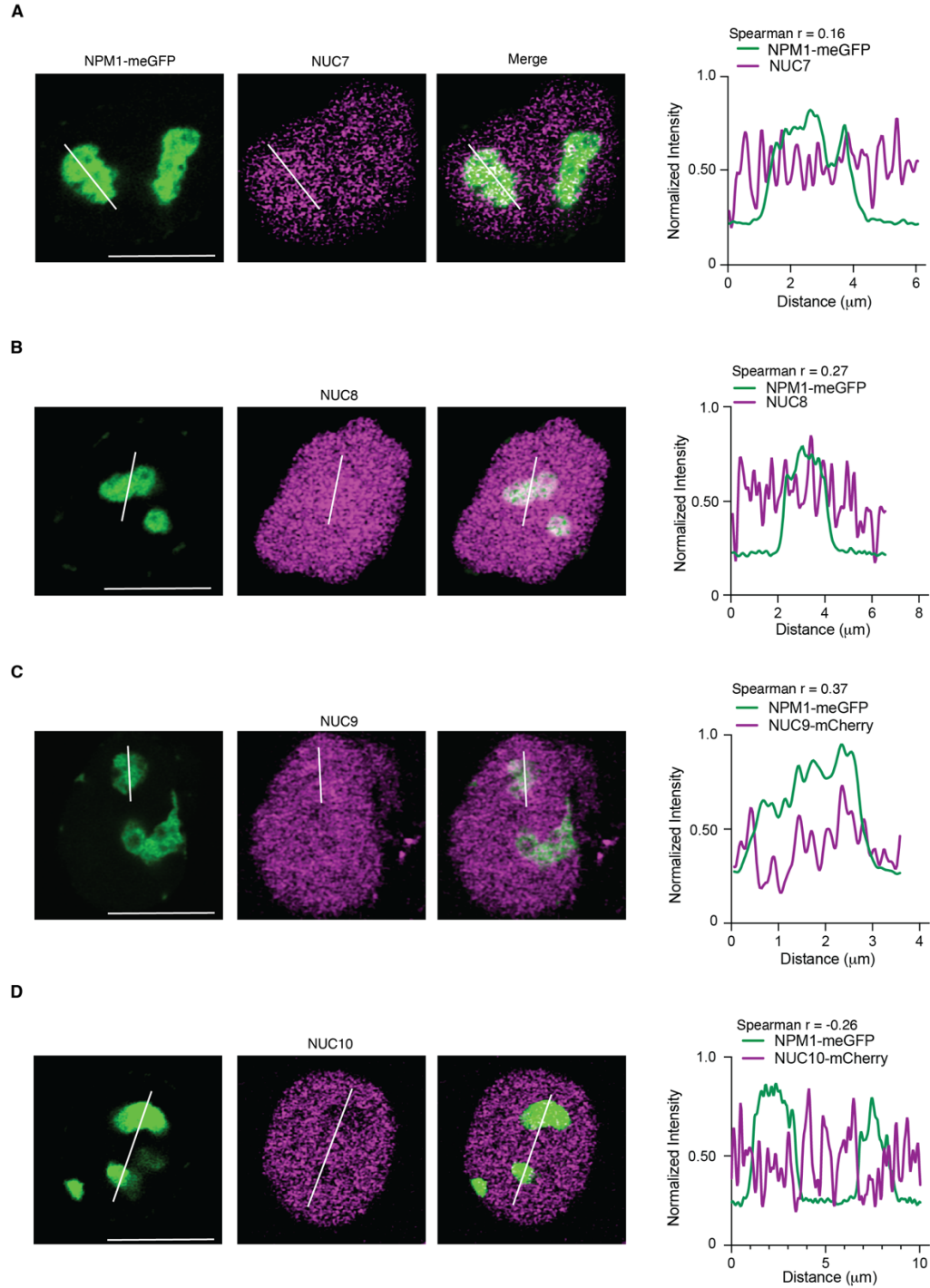
Supplementary Figure 10. Live cell confocal microscopy images of generated proteins targeted to nucleolus. Live cell images of colorectal cancer (HCT-116) cells expressing NPM1-meGFP (green) from the endogenous NPM1 locus and induced expression of the indicated nucleolus targeted protein **A.** NUC7, **B.** NUC8. **C.** NUC9, **D.** NUC10 (magenta). NLS-mCherry control Spearman r correlation = -0.61. Scale bar is indicated by white line in bottom left corner, 10 microns.
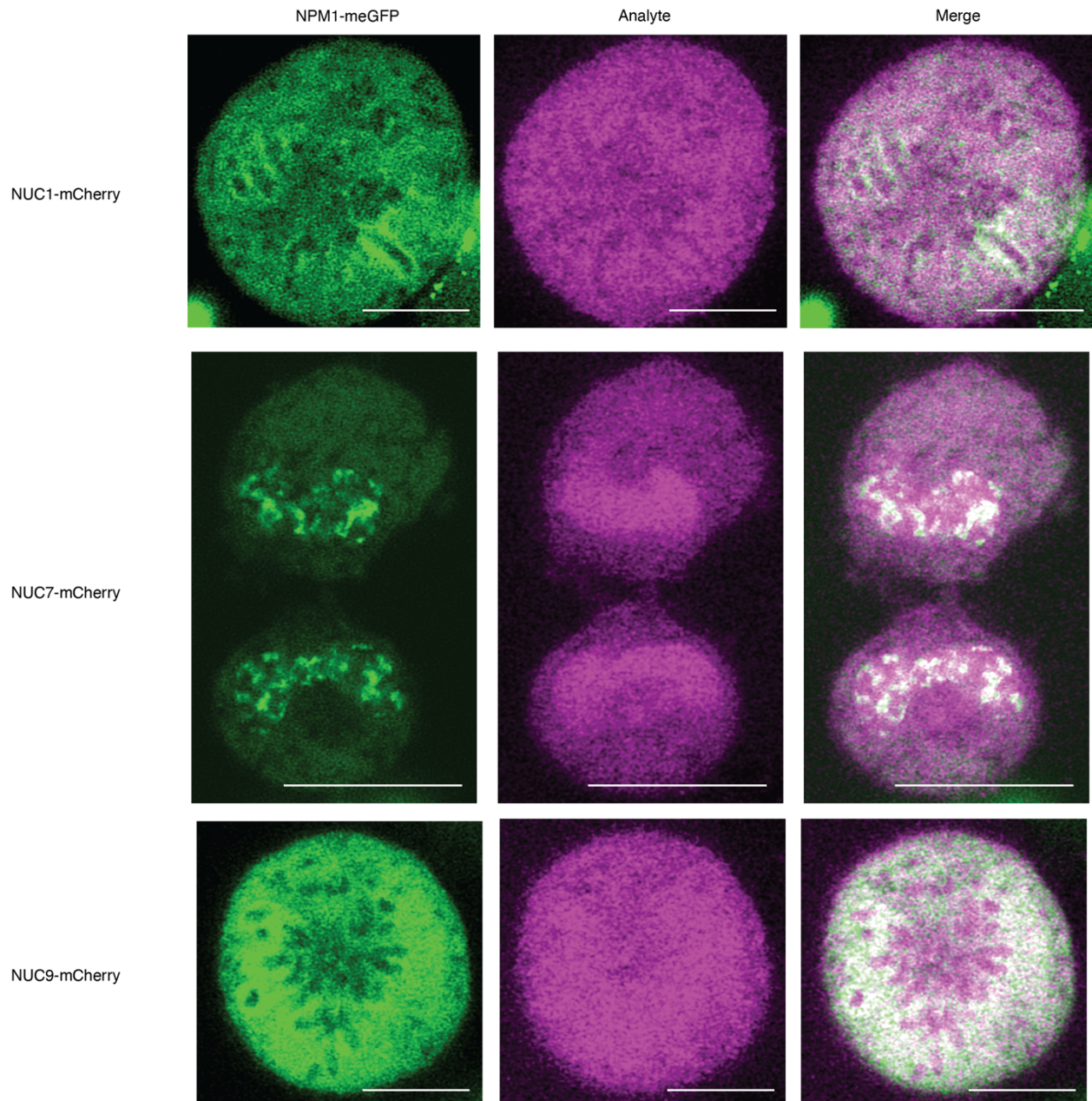
Supplementary Figure 11. Confocal microscopy of a subset of nucleolus targeted proteins at different stages of the cell cycle. Condensate targeted sequences were found to concentrate in puncta defined by NPM1-meGFP during different stages of the cell cycle. Shown are examples of live cell images of colorectal cancer (HCT-116) cells expressing NPM1-meGFP (green) from the endogenous NPM1 locus and induced expression of the indicated nucleolus targeted protein, analyte (magenta), signal overlap at similar signal intensities (white), scale is indicated by white line, 10 microns.
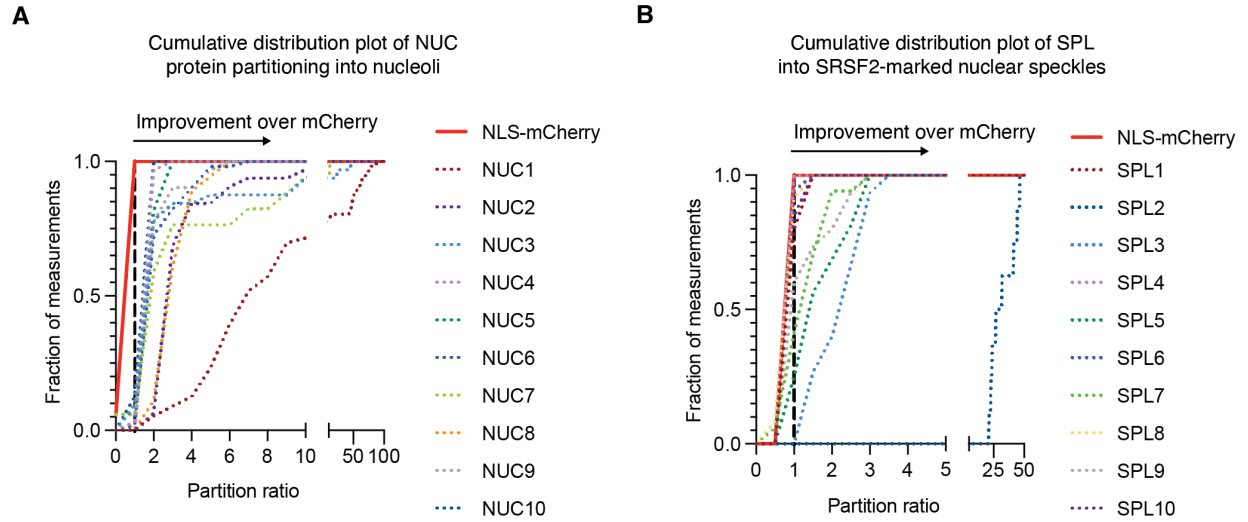
Supplementary Figure 12. Cumulative distribution plots showing protein partitioning into different condensates. Graphs showing the cumulative distribution plots of generated protein sequences partitioning into the target compartments defined by NPM1-meGFP, SRSF2-meGFP. These data show the range of partitioning values found for different foci and the corresponding generative sequences. **A.** NUCX protein partitioning into NPM1-meGFP marked compartments, **B.** SPL protein partitioning into SRSF2-meGFP compartments. Dashed line at partition ratio = 1 indicates lack of enrichment over the nucleoplasm.

A

Partition ratio of SPL proteins in nuclear speckles and
SRSF2-marked foci relative to the nucleoplasm

B

Spearmen r = - 0.49
— SRSF2-meGFP
— NLS-mCherry-a

C

Spearman r = 0.81
— SRSF2-meGFP
— SPL2-mCherry-a

D

Spearman r = 0.97
— SRSF2-meGFP
— SPL3-mCherry-a
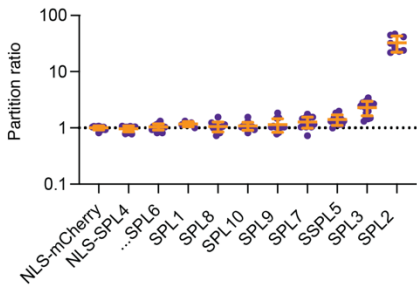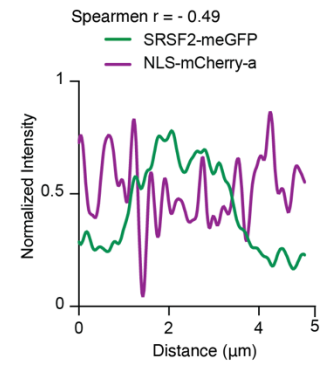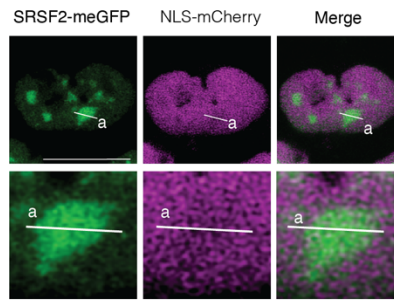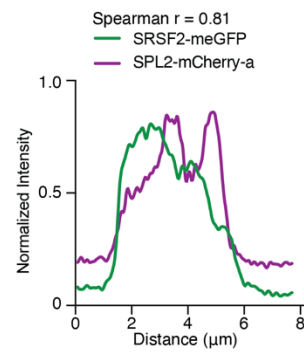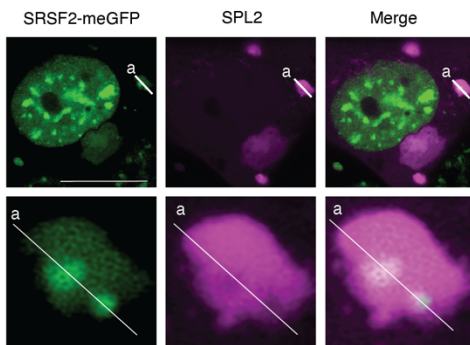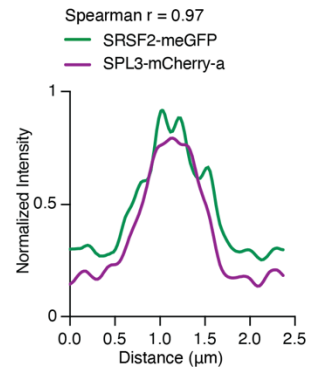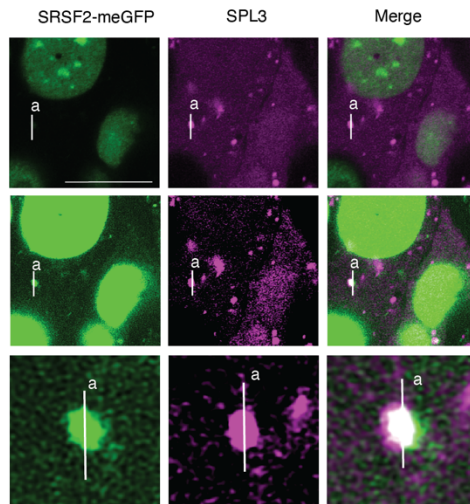
Supplementary Figure 13. Subcellular distribution of SPL proteins. SPL proteins were found to associate with SRSF2-meGFP in cytoplasmic bodies, but lost the capacity to migrate into the nucleus where nuclear speckles are normally formed. This phenotype is analogous to the pathological mutations in the splicing regulator RBM20 that promotes its mislocalization into the cytoplasm and association with splicing proteins, leading to cardiac disease (*33, 34*). **A.** Dot plot showing partition ratio measurements of different SPLX proteins and control NLS-mCherry protein into compartments identified with SRSF2-meGFP signal, dotted line indicates a partition ratio equal to one. Partition ratio measurements compare whole nucleus to puncta (see supplementary materials, *Imaging data analysis* for more details). Representative live cell micrographs and analysis of **B.** NLS-mCherry, **C.** SPL2, **D.** SPL3, constructs (magenta) in colorectal cancer cells (HCT-116) expressing SRSF2-meGFP (green) from the endogenous locus. Figure S11D shows two intensity values: low intensity; top. high intensity; middle, bottom. This was done to clarify that the SRSF2-meGFP is incorporated into SPL3 compartments. Top images show whole nucleus, bottom images show zoom of SRSF2-meGFP marked compartment. Scale: 10 microns.
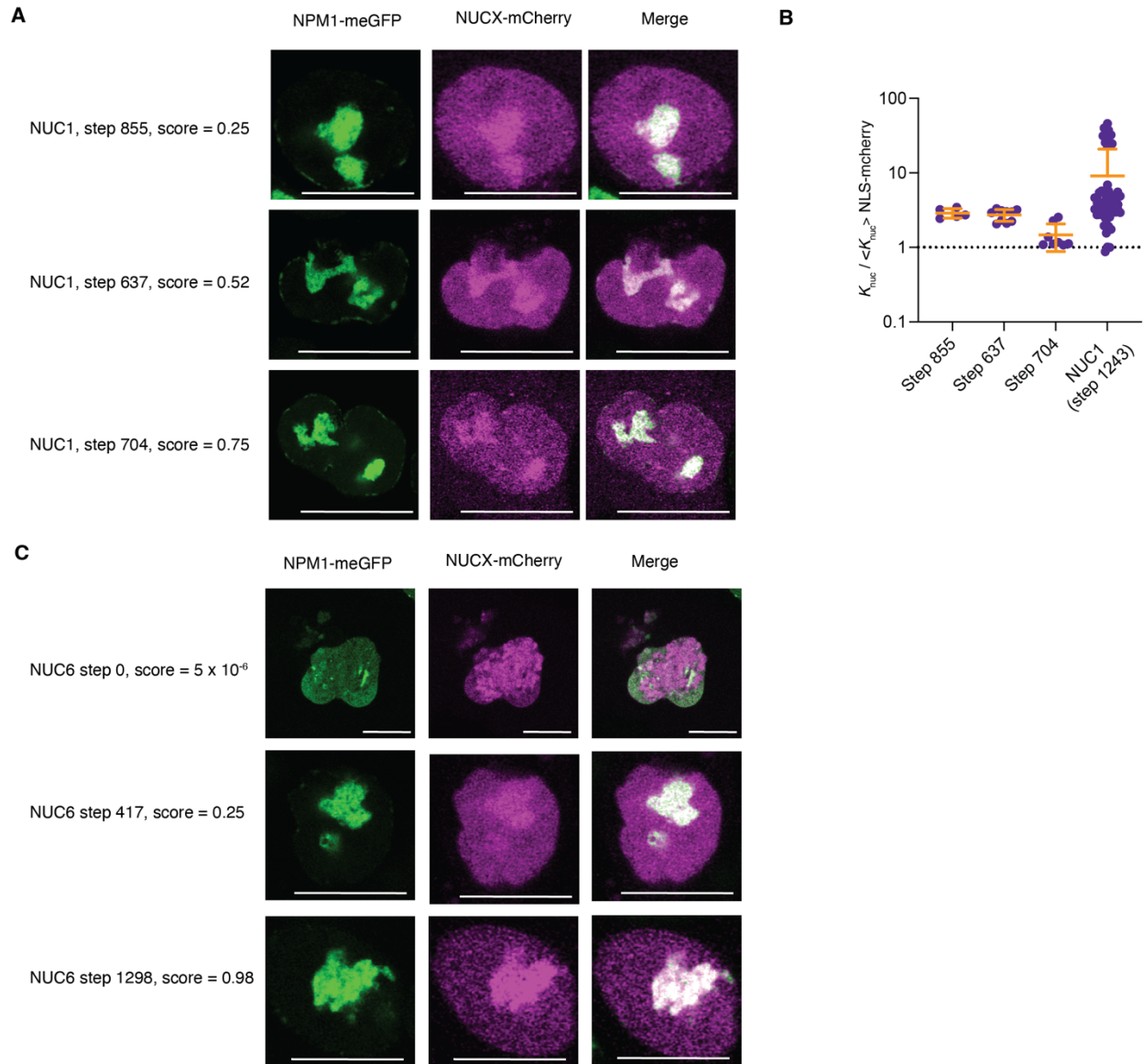
Supplementary Figure 14. Nucleolar partitioning and protein phenotype is sensitive to prediction strength. Sensitivity analysis showing how increased sampling with the MCMC algorithm tends to lead toward improved incorporation into a target compartment. **A.** Live cell images of NUC1 proteins in colon cancer (HCT-116) cells expressing NPM1-meGFP from the endogenous locus. Proteins were generated with MCMC with a range of scores. **B.** Quantification of the partition ratio of each NUC1-X step protein as compared to the average partition ratio of NLS-mCherry. **C.** Live cell images of NUC6 proteins in colon cancer cells expressing NPM1-meGFP from the endogenous locus, merged panels are repeated from Figure 2E. Scale bar is indicated by white line, 10 microns. Quantification given in panel 2E for NUC1 steps.

A Benign and uncertain missense mutations occur with a higher frequency in humans than pathogenic mutations
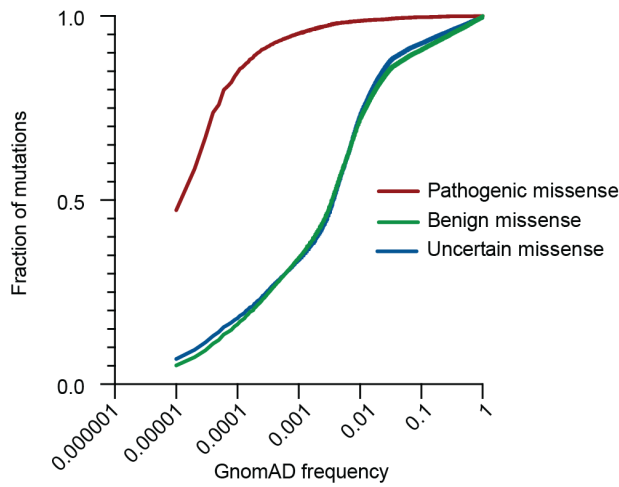
B Pathogenic mutations tend to have a higher Wasserstein distance than benign or uncertain mutations
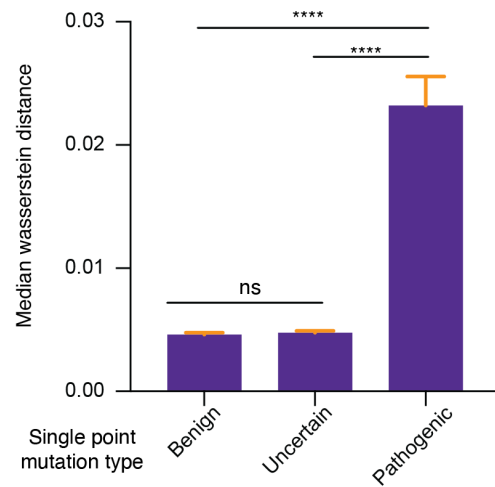
C Pathogenic mutations are more likely than benign or uncertain mutations to occur in structured domains
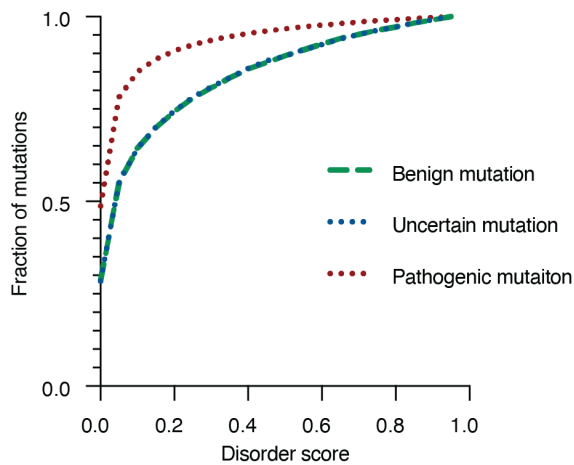
Supplementary Figure 15. Pathogenic human missense mutations occur less frequently in, but have a greater median Wasserstein distance than benign or uncertain mutations. Single point mutations defined as "pathogenic", "benign", or "uncertain" were collected from ClinVar. Wild-type and mutant sequences were analyzed with ProtGPS and the Wasserstein distance between the compartment predictions for wild-type and mutant sequences were computed. Benign mutations are expected to occur more frequently as reflected by a higher GnomAD frequency than pathogenic, but not impact the Wasserstein distance to the same degree as pathogenic mutation if mislocalization was the pathological result of that mutation. **A.** Cumulative distribution showing pathogenic missense mutations tend to occur less frequently in humans than benign and uncertain mutations. **B.** Bar graph showing that pathogenic missense mutations tend to alter the distribution of proteins more than "benign" or "uncertain" mutations. Mann-Whitney test, ****, p-value < 0.001. **C.** Cumulative distribution plot of the fraction of mutation labeled as "pathogenic", "uncertain", or "benign" binned by the predicted protein disorder (DR-BERT) at the mutation site.
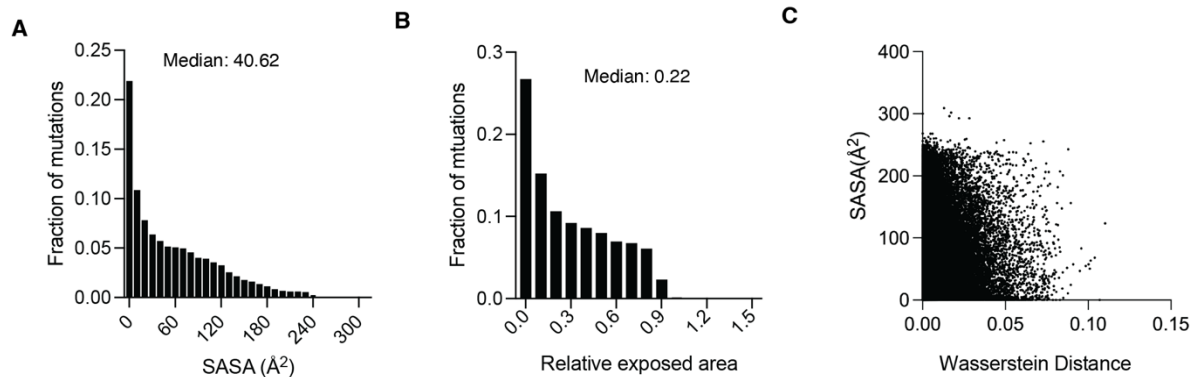
Supplementary Figure 16. Solvent accessible surface area of pathogenic single point mutation sites. Analysis of the solvent accessibility of wild-type amino acids at different mutation sites using the Alphafold2 predicted structures. Mutation sites are defined as the location on a protein where a mutation occurs and the solvent accessibility is computed using the wild-type amino acid. **A.** Histogram of single point mutation sites by solvent accessible surface area. Relative exposed area normalizes the surface area for each amino acid by its hypothetical maximum surface exposed area (*111*). **B.** Histogram of single point mutations binned by relative solvent exposed area. Each amino acid residue was normalized by that amino acid's hypothetical maximum surface area. **C.** Scatter plot of solvent accessible surface area plotted against the Wasserstein distance between wild-type and single point mutant.

A

Pathogenic single point mutations and SLiMs

3484 alter 2 SLiMs   3534 alter >2 SLiMs

10,503 alter 1 slim

100,507 mutations don't alter a SLiM
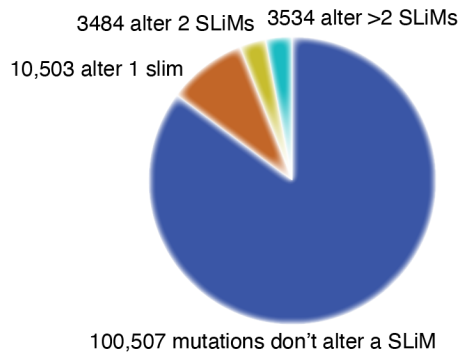
b

ProtGPS is sensitive to single point mutations altering SLiMs

Median Wasserstein distance

# of SLiMs altered by a mutation

C
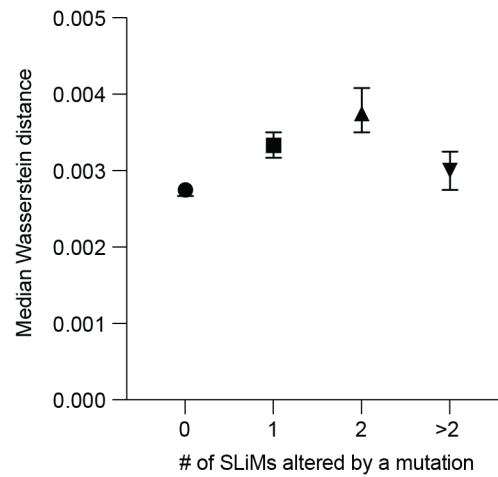
Single point mutations altering SLiMs significantly influence Wasserstein distance

Wasserstein distance

SLiMs altered
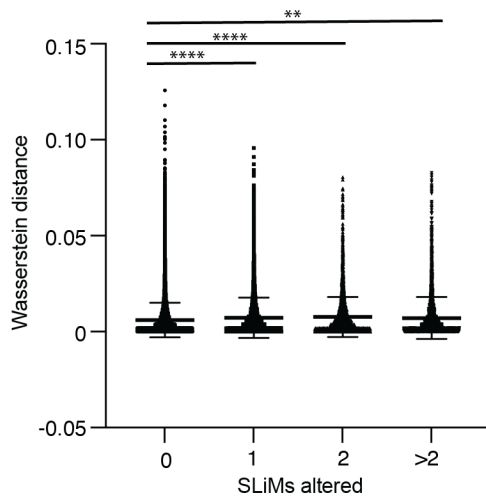
Supplementary Figure 17. Pathogenic mutations within short linear motifs sites. Pathogenic single point mutations within regions defined by a short linear motif were analyzed to determine if they would change the Wasserstein distance computed from ProtGPS predictions of the wild-type and mutant proteins. **A.** Dot plot showing the proportion of proteins with 0, 1, 2, or > 2 short linear motifs (SLiMs) affected by a pathogenic single point mutation in our ClinVar dataset. **B.** Pie-chart showing the number of SLiMs affected by pathogenic single point mutations. **C.** Median Wasserstein distance as a function of SLiMs affected by a single point mutation, error bars show the 95 % confidence interval. **D.** Wasserstein distance between single point mutants and wild-type proteins, showing the range of Wasserstein distances for mutants changing 0, 1, 2, or > 2 short linear motifs. Mann-Whitney U-test comparisons between 0 SLiMs altered and 1, 2, or >2, were found to be significant (****, p-value <0.0001. ***, p-value < 0.001. **, p-value < 0.01).

A

0.97 % of pathogenic mutations affect a known post translational modification site

■ Doesn't effect known PTM
■ Does effect a PTM

B

Single point mutations at post translational modification sites compared to all other single point mutation sites

Fraction of mutations

Comparison to non-PTM site mutations
······ Wasserstein distance, Lipidation sitemutated (n =23), p-value =0.13 , KS = 0.24
······ Wasserstein distance, Glycosylation site mutated (n=232), p-value = 0.15, KS = 0.07
······ Wasserstein distance, PTM mutation site (n=872), p-value = 0.11, KS = 0.04
—— Wasserstein distance, all other mutations (n=77,971)

Wasserstein distance

Supplementary Figure 18. Pathogenic mutation of post translational modifications sites can impact subcellular distribution predictions. Post translational modifications could alter the distribution of proteins in cells by changing their propensity for interaction with other molecules, their solubility within condensate compartments or their ability to associate with other molecules. Single point mutations were assessed if they interfered with known post translational modifications reported in the UniProt database. That information was then compared to the Wasserstein distance computed from compartment predictions made by ProtGPS on wild-type and single point mutant protein sequences. **A**. Pie-chart showing the relative proportion (1130 of 118,069) of pathogenic mutations in our data set at the post translational modification sites reported in the UniProt database. Single point mutations impacting lipidation sites and glycosylation sites would be expected to change the solubility of a protein in opposing directions given their potential size and known opposing aqueous solubilities. **B**. Plot showing cumulative distributions of Wasserstein distances for lipidation (n = 23) (···), glycosylation (n = 233) (···), other post translational modification sites (n = 874) (···), and all other post translationally modified wild-type and mutant pairs (n = 77,971) (—). Kolmogorov-Smirnov statistic and p-value is given for each PTM-type compared to all mutations in panel B.

A

Pathogenic mutations tend to be solvent exposed

Mutation site exposure

- Buried residue mutated (SASA = 0)
- More than half exposed (Rel. SASA=0.5)
- Greater than 25 Å² exposed
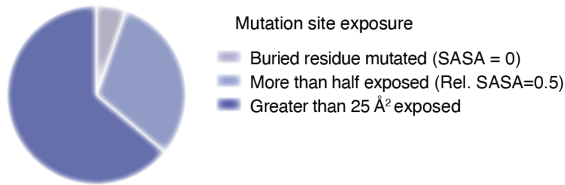
B

Wasserstein distance is independent of site solvent exposure

- Wasserstein SASA = 0 (n=87)
- Wasserstein SASA > 25 (n=495)
- Greater than half exposed (n = 1033)

C
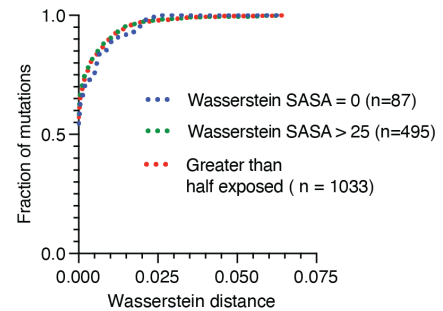
Single point mutations tended to occur in sites predicted to be ordered

Increasing probability of disorder
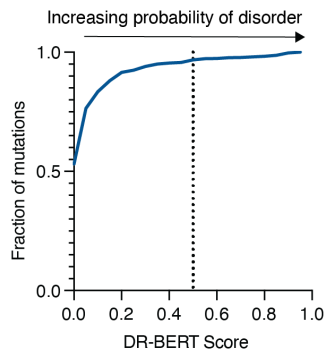
Supplementary Figure 19. Mutation of test set proteins at solvent exposed and buried residues impact the protein distribution code. Protein sequences in the test set were studied to determine if assess if solvent exposure correlated with the Wasserstein distance between the predicted compartmentalization of wild-type and single point mutants. Unexposed residues may impact the stability and subsequent distribution of a protein. **A**. Pie-chart showing the frequency of a pathogenic mutation found in test set proteins to be buried (solvent exposed surface area = 0), solvent exposed (greater than or equal to 50% exposure), or partially buried (> 25 square angstroms). Buried residues or surface residues might influence a protein's subcellular compartmentalization by altering the surface chemistry directly or by reducing the stability of folded conformations. To look for clues, we assessed if test set mutations at buried, partially exposed, or exposed residues tended to have similar Wasserstein distances. **B**. Cumulative distribution plot for Wasserstein distances between pathogenic mutants and wild-type proteins in classes defined in A. Protein structure in test set proteins might influence the outcome of data presented in panel B, therefore a **C.** cumulative distribution plot was computed showing the fraction of mutations occurring in test set pathogenic mutants at amino acid sites as a function of protein region disorder (DR-BERT score).

A

Predicted average disorder of the truncated protein region and the affect of its loss on protein distribution code

B

Predicted disorder at the site of a single point mutation and affect of that mutation on protein distribution code
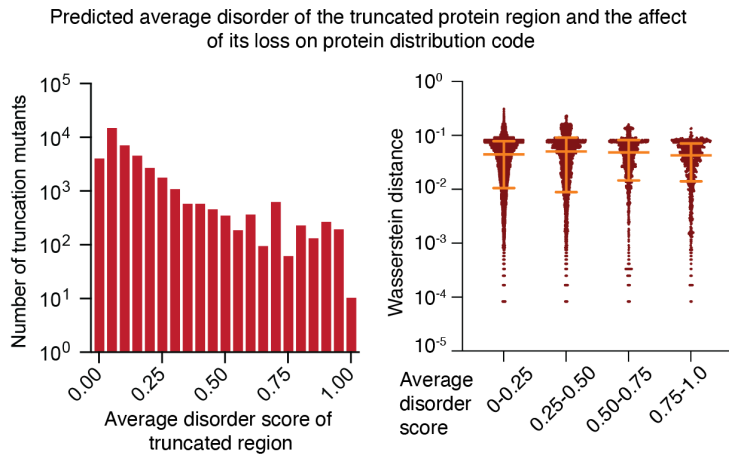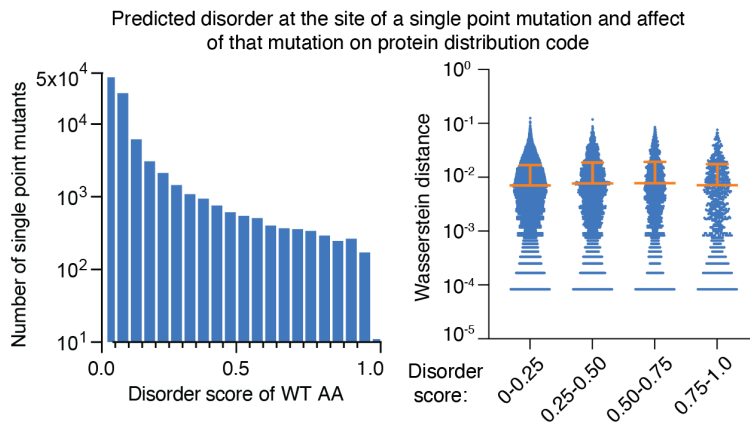
Supplementary Figure 20. ProtGPS and sensitivity toward mutation site structure or disorder. Truncation and single point mutations were contextualized by DR-BERT disorder score predictions. The effect of the loss of a truncated region or influence of a single point mutation on the Wasserstein distance of wild-type and mutant proteins were stratified by the disorder average disorder score of the lost domain or the disorder score at the site of mutation. **A.** (*left*) Histogram of the average disorder score for a pathogenic variant's truncated domain, (*right*) dot plot of Wasserstein distance between wild-type and pathogenic truncation variant for different ranges of disorder scores, mean and standard deviation are shown. All comparisons between disorder score group 0.00-0.25 and other groups were significant, p-value < 0.0001. **B.** (*left*) Histogram of the average disorder score for the wild-type amino acid mutated in each single point variant, (*right*) dot plot of Wasserstein distance between wild-type and pathogenic single point variant for a range of disorder scores. Mann-Whitney U test comparisons between disorder score group 0.00-0.25 and other groups were significant, p-value < 0.0001.
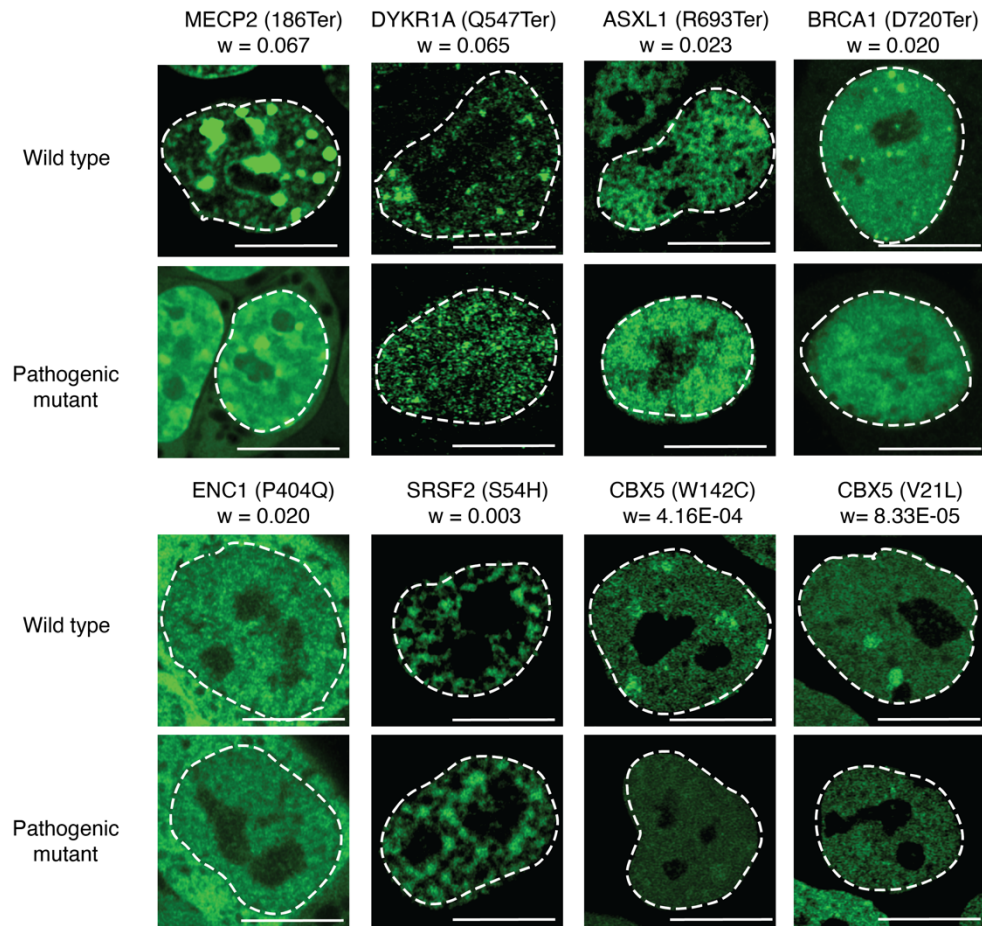
Supplementary Figure 21. Live cell confocal micrographs of wild type and disease variants in mouse embryonic stem cells. Live cell images of mouse embryonic stem cells (v6.5) and MCF7 cells (BRCA1 and BRCA1 D720Ter) expressing wild-type and disease variant proteins listed in Table 6 were fused to meGFP, scale 10 microns.

A

Input image Identify protein distribution

Compute
Co-occurrence matrices

Compute: entropy Compute: homogeneity

Wild-type

$$\begin{bmatrix} p(1,1) & \dots & p(1,Ng) \\ \vdots & & \vdots \\ p(Ng,1) & \dots & p(Ng,Ng) \end{bmatrix}$$

Wild-type entropy
$= -\sum \sum -p(i,J)\log(p(i,j))$

WIld-type homogeneity
$= -\sum \sum -p(i,j)^2$

Mutant

$$\begin{bmatrix} p(1,1) & \dots & p(1,Ng) \\ \vdots & & \vdots \\ p(Ng,1) & \dots & p(Ng,Ng) \end{bmatrix}$$

Mutant entropy
$= -\sum \sum -p(i,J)\log(p(i,j))$

Mutant homogeneity
$= -\sum \sum -p(i,j)^2$
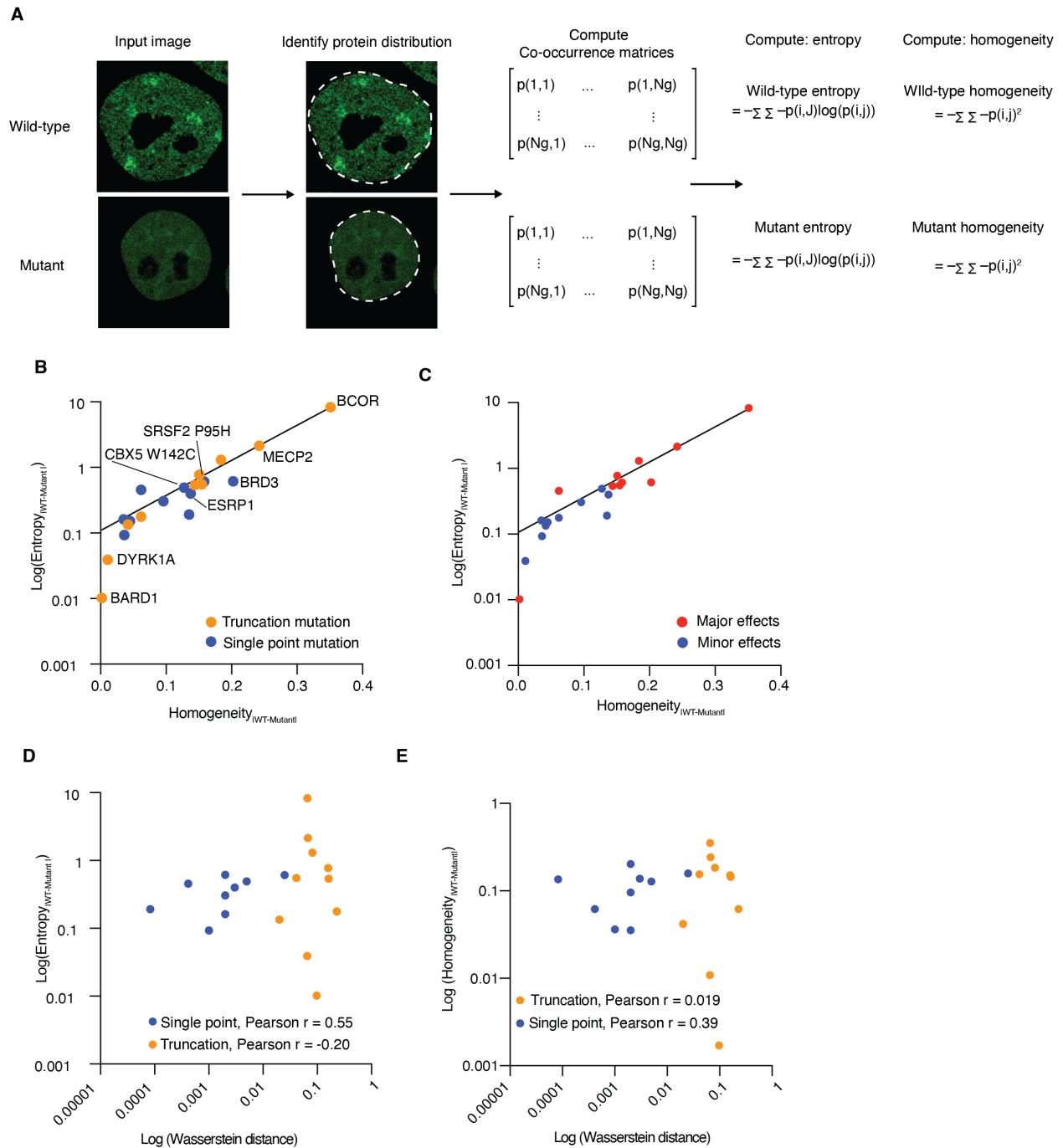
Supplementary Figure 22. Signal homogeneity and entropy changes between wild-type and pathogenic mutant proteins. Changes in the patterns present in an image can be detected with image analysis features. The effect of the mutation on subcellular localization produces a modest correlation between image entropy or homogeneity and Wasserstein distance. Signal homogeneity is measurement of how homogenous a signal is within a defined region of an image, with uniform signal equal to 1. Entropy measurements compute the degree of order within the defined region of an image where higher values indicate a more random distribution of signal. **A**. Schematic showing the approach to calculating entropy and signal homogeneity within the nuclei of pathogenic variant model systems. **B**. Plot of Log_{10} Entropy$_{\text{WT-mutant}}$ against Homogeneity$_{\text{WT-mutant}}$, colored by variant type (single point, blue, truncation, goldenrod). **C**. Plot of Log_{10} Entropy$_{\text{WT-mutant}}$ against Log_{10} Homogeneity$_{\text{WT-mutant}}$, colored by

the "major" or "minor" effects described in Table S10 (minor effect, blue. major-effect red). **D**. Log_{10} Entropy$_{WT-mutant}$ plotted against the Wasserstein distance computed between each wild-type and single point (blue) or truncation mutant (goldenrod) pair, reporting Pearson's r correlation between each data type. **E**. Plot of Log_{10} Homogeneity$_{WT-mutant}$ against Log_{10} Wasserstein distance computed between each wild-type and single point (blue) or truncation mutant (goldenrod) pair, reporting Pearson's r correlation between each data type.

Table S1. Table of area under the receiver operator curve (AUC-ROC) metrics computed for the models studied in this work. See methods for more information on each approach.

| | Random split | | | MmSeqs2 split | | |
	ProtGPS	Random Forest	Logistic Regression	ProtGPS	Random Forest	Logistic Regression
Nuclear speckle	0.905	0.861	0.740	0.759	0.745	0.769
P-body	0.887	0.753	0.645	0.688	0.619	0.543
PML-body	0.768	0.555	0.554	0.614	0.549	0.587
Post synaptic density	0.872	0.767	0.742	0.766	0.718	0.759
Stress granule	0.830	0.658	0.649	0.619	0.668	0.621
Chromatin	0.921	0.794	0.692	0.663	0.674	0.633
Nucleolus	0.920	0.826	0.652	0.796	0.737	0.601
Nuclear pore complex	0.987	0.847	0.863	0.855	0.928	0.776
Cajal body	0.900	0.664	0.822	0.646	0.620	0.50
RNA granule	1.000	1.000	0.994	0.964	0.998	0.968
Cell junction	0.925	0.834	0.713	0.727	0.784	0.731
Transcriptional condensate	0.869	0.709	0.556	0.605	0.708	0.675

Table S2. Autoregressive greedy search generated N-terminal peptides created to target mCherry to the nucleolus.

Sequence ID	Sequence
400	KRIRSIRMMVKYMGEAFEYEGPCHTVGGKFTCHGICSYIHHPRPVMGGNYAWSTRSY WVMMAICTVPKYFNGDICMVSDNGHGCCVGMGLACLCQKHYKMKHMNHDAHTE YNHTAEWHHDWEEWFLECMANAPWMAKMEAQIDFKMKGDT
403	KKRMWDRQRFSTSFCYTGYIESIWGWGLYDAMRTVAPQKWPKVITIMWHAFASPYP KLCLQAHITNCGCHMRTVFIRCPWTFSGEIKHCGWCAWCDRWLQCFAHADMNVACS YSKQKPPPQRNVDRIHCTAVKPCIPRLPYVPHPGPYCC
404	AKKISCHLRILCYDRTSSMRWYAQGRAMRKAKEIRCNNHVQYSCKTIWCRMGIMIM EGNCWETWITYHYTTAHPGKNWHVRRWDNMQKECNHAGWPRPRLIWHTHGLKHA MMSEKSEPHYNQDEACFYNKGMYEYSMMEHHDMDFVTCVS
674	IEEEEIFMHRRRMGDWIQSKCKQCHEWNIKNHCEVWFKWRVSYCRCFNSFNCGQVR NPCQMCDHLVGTTTFMQRKEKDVGEILQMRHPIGRCAVFCSHQPKHNFISHETCRAQ GRWMLNSEEDMEVPIACADCGAWCCIHWEEDHSWPCT
440	GIETNKYRIYGAWDWIVASIIVQGVCDFHYTTTKEALRFKYIFGKMSWKHGCAIERRC NNIGAIIDKHRFHEHIRAEANNAWAWPCAMAIDPIFRCGWFWLKVRPERYVKPKKFW KHKENEDDHIKINTEIHHIWNWMMCWWHDDTKKSV
401	MGVDEKLEMTALFCIGIMGLSWQCCRSEVIMCCDDDTRHIMCMVVHYSIRCSYQRM KYAMRDKKYFIWKMFRSKIGHRKSICLWRLHGIHKPWQTQWTSAAKDHKPYPDKH QNYEHHYYDVRNTCRKIWHPSGYHADMDEWTMNNLEDAE
505	GDKNYAMMKGDNTEAWGGWKRAYSHTHHTKHCIRPRVKCWHYRIKKHSFIHGDW CCYNRVIWGPEKRWWCQHFYGDENHDIKENWEEDRECAPISPGHPILREAVRRRFLR YTRCNIRLSFIRMPEGTSSVQHFTDIQRVKWFDYGFPVH
506	LKHDGNHHHGCAKNVIHRYDEQRCDHTAADHINCVYGGLIKAQDTATWSHMTLTM YMCHFNSHRIPTCKQWWRWYECTVPRLQPMERKDITPRHRGRRLGGKTMYPEWMN RCWNHTNGINHECSGTWDDRSHHGGCPYHCDDGIRNECTS

Table S3. Markov Chain Monte Carlo generated N-terminal peptides created to target mCherry to the nucleolus (NUCX) or nuclear speckles (SPLX).

Name	Sequence
NUC1	FMLVSTLWWKQKRLNNAVRTHTKFLTTINNPWRDFCSHRKKYCQKRKHEHATLKSWGTNN GSRRAAGICSGYGPEHSPDANTVKHCCIDYDSIDPIRCTR
NUC2	HFMRIADRKVMHHGCAKQGNSWNHIGQKPCCSKVKKGEQSQKADAVVWGVKCHMKWE ARSQCNQSFEKMQLHCPMSCRVQESSHNQHNIQPKANHQAMIH
NUC3	ATDYRQEGLKMETQMSVTDAMIPSGPVKWGCNPNSKSKQKPTSVRQATHGTAWTQESHVW NIWGIPCQLHADTHADPFEWKGVAHTADPVNHDRANRNES
NUC4	DWKWRMYGEWDSTGTVMGEWGHRHCDTVQAICWVNRLYRKKEDPQAKHDFRAHQLPM AQNKPKQHQCKKEEGILEPSKGVGGKGIRMWWDPEYRIYQEPL
NUC5	GRPFRRFKKWEELDGPPIGEQLQGRLRETAYPLKEKIHTHHIFGRMVKTDWLPCWQHSGHLI CRRMWSIFPEPTLKKKMDGHSNPHGAEGSQHKDFDPWS
NUC6	NCLETANAEMDEPHDKILHEPRKAVRYQHHGQEYDRLQWPVTHPTFAESEMEKQRHYVHC DRKRWKKCRIEEEKRQKLPHEPLEVSPVKHCPFEAEEYNG
NUC7	HGQNRRRKNIGTLKMHTIRGFFPMFSEIRNNHTFTIHGSKSFNSDFQDQNLHCHDRMMHLQI SDSMNNTGEEWMTEKVNSLPRKGKSGGPPYKPKVWSVQ
NUC8	FMDDVLWQLHARQSFRYAHHFPGPVNSRKHFTHTICSDVDKNTRMGEDNMVPMCMPEAEY ICPIDDLSLARSHKQRDMSTIFQETNLKVSNKQWRRPWLQ
NUC9	HNVHRMNKSKLSLTLKRQPTITAMHFEASVSNHWKGFPSSNVVAHSGYYKEIAPHVTEQAN MDQGVGMGIRSQSHTSQLNEIDNEPPGEAKESSAGCASY
NUC10	LSPTWCDEVANDQQPTGNQAETHICNSIKGQSEAMGEHNNMQHVAGDWEKYMEPVYPHGE EMDNPYAMCDGHFCPEIIKSLGGRANNKTQNGQYQINKHT
SPL1	PGVPHQTLRHHIRPHEHIHAFKDRNWEGKTKGPEFNEYHNAEFHHHGTNESHCSERAKFRF HQQRQTPHREIIESLSEEWNSTQKECHHRTTKEFVKCGK
SPL2	VNDITDVEMAVGRVPREGGNATERCYACFHHLDDYDLHQQMHGRDAPHMRNNSYKKAAH SEHINEVDHQGLQSDVEEYEGVMNEDTFKYMADERDCSPRN
SPL3	TKIKKHRSTPNMIQSPVTYPDEDHTNNHAGWKTTKAAAPKFRCAARQINRTAMMRCENFAI TIDDMPSQDWPHKDDHGAGDDKKDCMPARYDGHTEETND
SPL4	FFDDVLWQLHARQSFRYAHKFPGPVNERKHFTHSRCSDVDKNTRMGEQNKVPMCMPEAEY ICPIDDLSLARSHAQRDMSTSFCETYPKVSNTQWRRPHLQ
SPL5	EKSHMHGLSMHNCHCGGMSCHHYQQPKMHAVSYKKFVNYGPVEDTLGARDEFVYHVRRS EKRREMNNFEPWQFHTKTKTRHHKQSSHEGTWKWPAPQFHP
SPL6	RRFRASIRLVHACGHNHEGKRPFGERWPCEDDKHKPMENQLMKCPFSLMHQQMAYMMEM GDEWHPTMHYHTHMHAPMAEETYKTKVYNSYYGLGWWVDPM
SPL7	THDEYSYHTRNTRNGFAFDRKDTGRSWGEYNQFKQTGADVNTDTRPLHRPAPKNNTRLYA GRGLSRTKCKLERTTSRHQERHTGNNPEFASNCVSEPAFP
SPL8	CEYLHARTFSTRVPHAAISTVSPSKDYEDNGYHPAADDCPADSHCYPTMYDKTQWHEYRWH DTQHPSIDQKGNVSAHSEFHQHTGCNPAFFSKALNVMQY
SPL9	YEFFFPERLVRISQAPKLKELEGTGMRKEPPSTKCTMCFNDLCMLVLHGRIWRIKQQDVKNN PSDAMKVTEGENAADKHDHRKGSRHPMYCCPMCDCDFM
SPL10	NTSTGVEKHKRVTNQKRDDCTKSCCMISQKIAVARDGHDEVTAPPYTRYTHDVPCYGQTSV HKPRLNFKTADVMECDLSGHCSFEKKIDKETQNDEKMLD

Table S4. Number of Eukaryotic short linear motifs found in each of the NUCX series proteins.

Name	NUC1	NUC2	NUC3	NUC4	NUC5	NUC6	NUC7	NUC8	NUC9	NUC10	Count
PDZ domain ligands	0	0	0	1	0	0	0	0	0	0	1
PIKK phosphorylation site	0	1	1	0	1	0	0	0	1	1	5
Cdc14 phosphatase dephosphorylation site	0	0	0	0	0	1	0	0	0	0	1
Polo-like kinase phosphosites	1	0	0	0	0	0	0	0	0	0	1
Caspase cleavage motif	0	0	1	0	1	0	0	0	0	1	3
MAPK Phosphorylation Site	1	0	0	0	0	1	0	0	0	0	2
Casein kinase 2 (CK2) Phosphorylation site	0	0	0	0	0	1	1	0	0	0	2
MATH domain binder of the TRAF6 K63 E3 ligase	0	0	0	0	0	1	0	0	1	0	2
Cyclin N-terminal Domain Docking Motifs	0	0	0	0	0	0	0	1	1	0	2
Casein kinase 1 (CK1) Phosphorylation site	1	0	0	1	0	0	1	0	1	0	4
di Arginine retention/retrieving signal	0	0	0	0	1	1	1	1	0	0	4
PDZ domain ligands	0	0	0	1	0	0	0	0	0	0	1
Apicomplexan export motif	0	0	0	0	0	1	0	0	0	0	1
C-Mannosylation site	0	0	1	0	0	0	0	0	0	0	1
LATS kinase phosphorylation motif	1	0	0	0	0	0	0	0	0	0	1
GSK3 phosphorylation site	1	0	0	0	0	1	1	1	1	0	5
PP1-docking motif RVXF	0	0	0	0	0	1	0	0	0	0	1
Cyclin N-terminal Domain Docking Motifs	0	0	0	0	0	0	0	1	1	0	2
Y-based sorting signal	1	0	0	0	0	1	0	0	1	0	3
PDZ domain ligands	0	0	0	1	0	0	0	0	0	0	1
Cyclin N-terminal Domain Docking Motifs	0	0	0	0	0	0	0	1	1	0	2
CendR Motif Binding to Neuropilin Receptors	1	0	0	0	0	0	0	0	0	0	1
Arc N-lobe binding ligand	1	0	0	1	0	1	1	0	0	0	4
APCC activator-binding ABBA motif	0	0	0	1	0	1	0	0	0	0	2
N-degron	0	0	0	0	0	1	0	0	0	0	1
Glycosaminoglycan attachment site	1	0	1	0	1	0	1	0	1	0	5
FHA phosphopeptide ligands	0	0	1	0	0	1	1	0	0	1	4
FHA phosphopeptide ligands	0	0	1	0	0	1	1	0	0	1	4
Di-Tryptophan motif of Delta-COP MHD domain	0	0	0	1	1	0	0	0	0	0	2
Cyclin N-terminal Domain Docking Motifs	0	0	0	0	0	0	0	1	1	0	2
Apple-PAN domain ligand motif	0	0	0	0	1	0	0	0	0	0	1
Cyclin N-terminal Domain Docking Motifs	0	0	0	0	0	0	0	1	1	0	2
IAP-binding motif (IBM)	0	0	1	0	0	0	0	0	0	0	1
Polo-like kinase phosphosites	1	0	0	0	0	0	0	0	0	0	1
IAP-binding motif (IBM)	0	0	1	0	0	0	0	0	0	0	1
WDR5 WD40 repeat (blade 5,6)-binding ligand	1	1	0	0	1	1	1	1	1	1	8
WDR5 WD40 repeat (blade 5,6)-binding ligand	1	1	0	0	1	1	1	1	1	1	8
NRD cleavage site	1	1	0	1	1	1	1	1	0	0	7
Binding motif for UBA3 adenylation domain	0	0	0	0	1	0	0	0	0	0	1
IAP-binding motif (IBM)	0	0	1	0	0	0	0	0	0	0	1
NEK2 phosphorylation site	0	0	0	0	0	0	0	1	1	0	2
APCC activator-binding ABBA motif	0	0	0	1	0	1	0	0	0	0	2
Cyclin N-terminal Domain Docking Motifs	0	0	0	0	0	0	0	1	1	0	2
PK Phosphorylation site	0	0	1	0	1	0	0	0	0	0	2
PDZ domain ligands	0	0	0	1	0	0	0	0	0	0	1
IAP-binding motif (IBM)	0	0	1	0	0	0	0	0	0	0	1
NEK2 phosphorylation site	0	0	0	0	0	0	0	1	1	0	2
N-degron.1	0	0	0	0	0	1	0	0	0	0	1
N-degron.2	0	0	0	0	0	1	0	0	0	0	1
N-degron.3	0	0	0	0	0	1	0	0	0	0	1
N-degron.4	0	0	0	0	0	1	0	0	0	0	1
C-terminal Imide degron	0	0	0	0	0	0	1	1	0	0	2

Table S5. Statistical testing of generated peptide sequences.

Protein	# of compartments assayed	p-value (KS-test, comparing protein to NLS-mCherry)
NUC1	45	< 0.0001
NUC2	32	< 0.0001
NUC3	27	< 0.0001
NUC4	33	< 0.0001
NUC5	33	< 0.0001
NUC6	51	< 0.0001
NUC7	12	< 0.0001
NUC8	14	< 0.0001
NUC9	28	< 0.0001
NUC10	16	< 0.0001
SPL1	5	0.0193
SPL2	8	< 0.0001
SPL3	16	< 0.0001
SPL4	14	0.7102
SPL5	11	< 0.0001
SPL6	17	0.4129
SPL7	13	0.0013
SPL8	12	0.2343
SPL9	12	0.2923
SPL10	13	0.9242

Table S6. Spearman's r correlation and mean partition ratios measured for each of the generated sequences studied.

Protein	Compartment	Spearman line A	Mean partition ratio	Average Spearman Correlation and standard deviation
NUC1	Nucleolus	0.930	17.353	0.91 ± 0.039
NUC2	Nucleolus	0.830	3.953	0.81 ± 0.033
NUC3	Nucleolus	0.220	4.549	0.19 ± 0.051
NUC4	Nucleolus	0.820	1.985	0.82 ± 0.034
NUC5	Nucleolus	0.930	2.077	0.90 ± 0.053
NUC6	Nucleolus	0.540	2.651	0.46 ± 0.059
NUC7	Nucleolus	0.160	3.888	0.19 ± 0.030
NUC8	Nucleolus	0.270	3.398	0.28 ± 0.025
NUC9	Nucleolus	0.370	2.491	0.36 ± 0.037
NUC10	Nucleolus	-0.260	1.545	-0.22 ± 0.074
NLS-mCherry	Nucleolus	-0.610	1.000	-0.47 ± 0.047
SPL1	Nuclear Speckle	ND	1.173	ND
SPL2	Nuclear Speckle	0.810	32.480	0.81 ± 0.029
SPL3	Nuclear Speckle	0.970	2.301	0.9425 ± 0.031
SPL4	Nuclear Speckle	ND	0.967	ND
SPL5	Nuclear Speckle	ND	1.815	ND
SPL6	Nuclear Speckle	ND	1.046	ND
SPL7	Nuclear Speckle	ND	1.522	ND
SPL8	Nuclear Speckle	ND	1.060	ND
SPL9	Nuclear Speckle	ND	1.436	ND
SPL10	Nuclear Speckle	ND	1.082	ND
NLS-mCherry	Nuclear Speckle	-0.49	1.003	-0.50 ± 0.038

Table S7. Top forty SLiMS most frequently mutated by pathogenic single point mutations.

SLIM (ELM database annotation functional site name)	Count	Median Wasserstein distance
di Arginine retention/retrieving signal	546	0.00458333
PP1-docking motif RVXF	525	0.00416667
Cyclin N-terminal Domain Docking Motifs	499	0.00391667
Cyclin N-terminal Domain Docking Motifs.1	499	0.00391667
Cyclin N-terminal Domain Docking Motifs.2	499	0.00391667
Cyclin N-terminal Domain Docking Motifs.3	499	0.00391667
Cyclin N-terminal Domain Docking Motifs.4	499	0.00391667
Cyclin N-terminal Domain Docking Motifs.5	499	0.00391667
NEK2 phosphorylation site	467	0.00375
NEK2 phosphorylation site.1	467	0.00375
Di-Tryptophan targeting motif to the Delta-COP MHD domain	463	0.00483334
SUMO interaction site	452	0.00341667
SUMO interaction site.1	452	0.00341667
Oomycete secretory protein processing motif permissive variant	435	0.00483334
PKA Phosphorylation site	434	0.00420833
PKA Phosphorylation site.1	434	0.00420833
Apicomplexan export motif	422	0.00445833
Calcineurin (PP2B) PxIxIT docking motif	419	0.00316667
Peptide Amidation Site	408	0.00341667
N-glycosylation site	395	0.00441667
N-glycosylation site.1	395	0.00441667
PKB Phosphorylation site	393	0.00308333
PK Phosphorylation site	378	0.00291667
Cks1 ligand	374	0.00270833
PCSK cleavage site	361	0.00291667
PCSK cleavage site.1	361	0.00291667
PCSK cleavage site.2	361	0.00291667
PCSK cleavage site.3	361	0.00291667
PCSK cleavage site.4	361	0.00291667
NES Nuclear Export Signals	352	0.00370833
NES Nuclear Export Signals.1	352	0.00370833
WAVE regulatory complex (WRC) binding site motif	347	0.00258333
Caspase cleavage motif	340	0.00341667
MSH2 lever 1 domain ligand	334	0.00350000
Binding motif for UBA3 adenylation domain	330	0.00412500
CDK Phosphorylation Site	330	0.00233333
CDK Phosphorylation Site.1	330	0.00233333
CDK Phosphorylation Site.2	330	0.00233333
PP2A holoenzyme B56-docking site	314	0.002875000
Polo-like kinase phospho sites	294	0.00358334

Table S8. Top 40 SLiMS by median Wasserstein distance between their wild-type and single point mutant.

SLIM	Count	Median Wasserstein distance
Tyrosine-based sorting signal	108	0.00945833
Helical calmodulin binding motifs	41	0.00766667
Helical calmodulin binding motifs.1	41	0.00766667
UHM domain Ligand Motif	52	0.007
HCF-1 binding motif	63	0.0065
KEAP1 binding degron	22	0.00616667
KEAP1 binding degron.1	22	0.00616667
F and H motif	46	0.00579167
ASX EGF hydroxylation	25	0.00566667
WRxxL motif	45	0.00541666
NRD cleavage site	215	0.00525
MAD2 binding motif	122	0.004875
Di-Tryptophan targeting motif to the Delta-COP MHD domain	463	0.00483334
Oomycete secretory protein processing motif permissive variant	435	0.00483334
MDM2 binding motif	116	0.00475
AAK1 and BIKe phosphorylation site motif	27	0.00458334
di Arginine retention/retrieving signal	546	0.00458333
Casein kinase 1 (CK1) Phosphorylation site	99	0.00450001
Apicomplexan export motif	422	0.00445833
N-glycosylation site	395	0.00441667
N-glycosylation site.1	395	0.00441667
Extracellular side LRP5 and -6 binding motif	50	0.004375
C-Mannosylation site	150	0.00433333
Cyclin D-specific Helical docking motif	29	0.00433333
Ligand motif binding the CSL BTD	136	0.00429167
Tankyrase-binding motif	215	0.00425
Glycosaminoglycan attachment site	98	0.00420833
PKA Phosphorylation site	434	0.00420833
PKA Phosphorylation site.1	434	0.00420833
PP1-docking motif RVXF	525	0.00416667
Arc N-lobe binding ligand	271	0.00416667
PTB ligand	219	0.00416666
PTB ligand.1	219	0.00416666
Binding motif for UBA3 adenylation domain	330	0.004125
O-Fucosylation site	230	0.00408333
Docking motif binding to N-terminal kinase domain of RSK family kinases	35	0.004
APCC activator-binding ABBA motif	244	0.00395833
APCC activator-binding ABBA motif.1	244	0.00395833
EH ligand	49	0.00391667
Cyclin N-terminal Domain Docking Motifs	499	0.00391667

Table S9. Post translation modification site categories, count in test set sequences, and their median Wasserstein distances.

PTM-category	Count	Median Wasserstein distance
Glycosylation site	234	0.00225
Lipidation site	23	0.00275
Other PTM-site types (Acetlyation, phosphorylation etc.)	834	0.00316
Site does not contain a known PTM	7131	0.00602

Table S10. Table of wild-type and pathogenic protein variants studied in mouse embryonic stem cells.

Protein	Mutation	Magnitude of effect on distribution	Wasserstein Distance (WT, Mutant)	Fraction of predicted NES and NLS signals mutated
DAXX	R318Ter	Major	0.162	NES 3/6 NLS 11/14
TCOF	Q55Ter	Major	0.159	NES 2/4 NLS 50/53
BARD1	R406Ter	Major	0.098	NES 1/4 NLS 2/6
BCL11A	Q177Ter	Major	0.081	NES 2/4 NLS 1/3
BCOR	Y657Ter	Major	0.066	NES 3/4 NLS 13/13
SALL1	S372Ter	Major	0.041	NES 3/5 NLS
SRSF2	P95H, S54H	Major, Minor	0.025, 0.003	NES 0/0, 0/0 NLS 0/5, 0/5
ESRP1	L259V	Minor	0.005	NES 0/6 NLS 0/0
BRD3	F334S	Major	0.002	NLS 0/24
TERT	T567M	Minor	0.002	NES 0/3 NLS 0/5
BCL6	R594Q	Minor	0.002	NES 0/3 NLS 0/1
RBM10	V354M	Major	0.000	NES 0/3 NLS 0/17
MECP2	R186Ter	Major	0.067	NES 1/3 NLS 11/16
DYRK1A	Q547Ter	Minor	0.065	NES 0/3 NLS 0/8
ASXL1	R693Ter	Minor	0.023	NES 1/2 NLS 0/13
BRCA1	D720Ter	Minor	0.020	NES 0/2 NLS 0/2
ENC1	P404Q	Minor	0.001	NES 0/3 NLS 0/2
CBX5/HP1a	V21L, W142C	Minor, Major	8.33E-05, 4.16E-04	NES 0/1, 0/1 NLS 0/6, 0/6