

## Previews

## Putting proteins in context

Mengzhou Hu<sup>1</sup> and Trey Ideker<sup>1,\*</sup><sup>1</sup>Department of Medicine, University of California, San Diego, La Jolla, CA, USA\*Correspondence: [tideker@health.ucsd.edu](mailto:tideker@health.ucsd.edu)<https://doi.org/10.1016/j.cels.2024.09.009>

Proteins exhibit cell-type-specific functions and interactions, yet most ways of representing proteins lack any biological or environmental context. To address this gap, recent work by Li et al.<sup>1</sup> introduces PINNACLE, a geometric deep learning approach that generates contextualized representations of proteins by combined analysis of protein interactions and multiorgan single-cell transcriptomics.

While the genome is largely constant across cell types, the proteins it encodes, and their resulting functions, can display substantial variation depending on cellular and tissue context.<sup>2,3</sup> This contextual variability poses challenges in understanding protein behavior across diverse biological settings.

Recently, researchers have turned to advanced computational methods, such as deep learning, to translate the complex set of features and data describing a protein into compact numerical formats, also known as object embeddings or “representations”<sup>4</sup> (Figure 1). Such representations have shown significant promise in their ability to comprehensively predict the functions and/or interactions of proteins of interest.<sup>5,6</sup> Thus far, however, proteins have been represented and analyzed largely without consideration of the relevant cell type or environmental context. This practice is, in fact, quite usual in genomics and proteomics at large: for example, structures of proteins and protein complexes in the Protein Data Bank<sup>7</sup> are not generally annotated with specific environments or cell types for which those structures are relevant. Nevertheless, omitting such information ignores the intricate relationship between protein function and the surrounding cellular and tissue organization. Hence, there is a clear need for standard protein representations that reflect the specific cellular context(s) of interest, leading to a more complete view of protein function in complex living systems.

Li et al.<sup>1</sup> address this need through PINNACLE, a geometric deep learning model for context-aware protein representations. To capture context, PINNACLE integrates a network of cell-type-specific

protein interactions with a second network of cell-cell interactions interlinking cells of different types and tissues. The first of these inputs—cell-type-specific protein interactions—is inferred by overlaying single-cell mRNA expression data for different cell types of interest onto a common reference human protein interactome. The second required input—a network of interconnections among different cell types and tissues—includes documented ligand-receptor binding interactions as well as hierarchical relations between cell types and the tissues and organs they comprise. Both of these inputs are modeled using graph neural networks, which serve to focus attention on certain protein interactions within cell types and key cell-cell interactions that underlie tissue organization. The result is to place proteins (as well as cell types and tissues) into an embedded “latent space”—a compressed, multidimensional representation of the data where similar entities are positioned closer together (Figure 1). Each protein representation is conditioned on a distinct context, such that a protein can take on a different representation (coordinates in the latent space) in each of the 156 cell types considered by the authors.

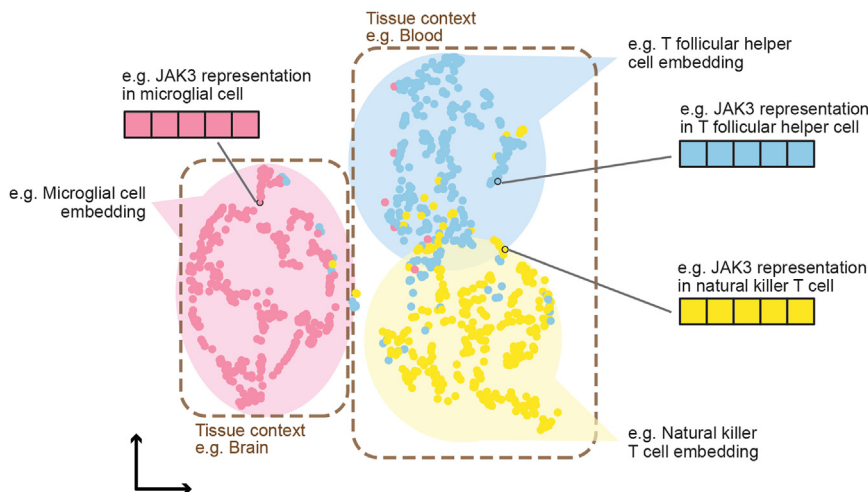
To validate the contextualized protein representations, the authors present several complementary analyses. They first show that protein representations from the same tissue tend to lie in close proximity to each other in the latent space, suggesting that this space is indeed substantially influenced by context. Furthermore, the authors compare the embeddings for different tissues to show that these are more similar for tissues from the same organ. Together, these results underscore PINNACLE’s capacity to learn contextualized representations that reflect

the underlying organization of cells and tissues.

As a proof-of-concept application, Li et al. use the PINNACLE latent space to distinguish proteins that are known therapeutic targets in rheumatoid arthritis (RA) and inflammatory bowel disease (IBD). Both of these diseases involve cell-type-specific mechanisms,<sup>8,9</sup> and as a result, the authors show that PINNACLE is able to recover known drug targets as well as the most relevant cell types for disease progression and drug action. For instance, PINNACLE recovers JAK3 (a protein kinase) and IL6R (an interleukin receptor), known drug targets and cellular context for RA<sup>8</sup>; these targets are identified predominantly in T follicular helper cells and classical monocytes, respectively. In the case of IBD, PINNACLE recovers the known drug targets ITGA4 (integrin subunit  $\alpha$ 4) and PPARG (peroxisome proliferator-activated receptor  $\gamma$ ) as highly relevant proteins in regulatory T cells and paneth cells. These nominated cell types align well with the current understanding of the pathophysiology of IBD,<sup>9</sup> showcasing PINNACLE’s ability to provide cell-type-specific insights that offer valuable direction for understanding disease mechanisms and drug action in the relevant cellular environments.

Li et al. further demonstrate the utility of PINNACLE’s context-aware protein representations by integrating them with 3D protein structural information. They focus this analysis on PD-1 and B7-1, important proteins functioning at the immune checkpoint that are targeted by cancer immunotherapies. By concatenating PINNACLE’s context-specific protein representations with separate embeddings based on 3D structure, they generate contextualized





**Figure 1. Conceptual overview of PINNACLE embedding space**

Schematic illustration of the unified embedding space of PINNACLE, where each point represents a protein representation colored by its associated cell type. PINNACLE generates distinct representation vectors (denoted by rows of colored boxes) for the same protein (e.g., JAK3) across different cellular contexts (e.g., microglial cells in pink, T follicular helper cells in blue, and natural killer T cells in yellow). Proteins from similar cell types (colored ovals) or tissues (dashed rectangles) are proximal to one another, reflecting these distinct cellular contexts.

structural protein representations. They find that the model assigns higher protein-protein similarity to known binding pairs (e.g., PD-1 and PD-L1) and lower similarity to nonbinding pairs (e.g., PD-1 and other proteins), with a larger gap between these scores than previous methods. This improvement highlights the potential of combining contextual information with structural data to improve our understanding of protein-protein interactions in specific cellular environments, which could have significant implications for drug design and understanding of disease mechanisms.

Beyond these case studies, PINNACLE is of potential interest in precision medicine, where cell-type-specific representations could enhance drug response predictions and facilitate the identification of therapeutic targets. While the authors focused on recovering known drug targets for RA and IBD in the present study (see above), a clear implication is that PINNACLE might also be used to nominate novel drug targets for diseases with cell-type-specific manifestations. We also foresee opportunities for integrating protein embeddings from other data modalities (e.g., protein localization images, post-translational modification profiles, and protein-DNA interactions) with broad applications in molecular biology.

A potential limitation of PINNACLE is that it is fundamentally based on access to curated cell-type-specific protein interaction networks. In the present study, these networks are not measured directly but inferred from single-cell mRNA transcriptomes. This step is somewhat problematic in that mRNA expression levels are only loose surrogates of protein expression and abundance.<sup>10</sup> One also wonders about the extent to which the inferred input networks are incomplete or biased, with low coverage of less-studied cell types or proteins. Thus, an important future direction may be to enhance the quality and coverage of experimental protein interaction data measured specifically for different cell types, perhaps by incorporating multiple layers of omics data. A complementary direction would be to develop and deploy downstream experimental validation methods for PINNACLE's predictions. Addressing these challenges could further PINNACLE's impact on deciphering molecular biology complexities and advancing precision medicine.

#### ACKNOWLEDGMENTS

We gratefully acknowledge support from the Bridge2AI program of the National Institutes of Health Common Fund (OT2 OD032742) and the Multi-Scale Integrated Cell (MuSIC) program from Schmidt Futures.

#### DECLARATION OF INTERESTS

T.I. is a cofounder and member of the advisory board of and has an equity interest in Data4Cure and Serinus Biosciences. T.I. is a consultant for and has an equity interest in IDEAYA Biosciences. The terms of these arrangements have been reviewed and approved by the University of California, San Diego, in accordance with its conflict-of-interest policies.

#### REFERENCES

- Li, M.M., Huang, Y., Sumathipala, M., Liang, M.Q., Valdeolivas, A., Ananthakrishnan, A.N., Liao, K., Marbach, D., and Zitnik, M. (2024). Contextual AI models for single-cell protein biology. *Nat. Methods* 21, 1546–1557. <https://doi.org/10.1038/s41592-024-02341-3>.
- Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. <https://doi.org/10.1038/ng.3259>.
- Ittisoponpisan, S., Alhuzimi, E., Sternberg, M.J.E., and David, A. (2017). Landscape of pleiotropic proteins causing human disease: Structural and system biology insights. *Hum. Mutat.* 38, 289–296. <https://doi.org/10.1002/humu.23155>.
- Yang, K.K., Wu, Z., Bedbrook, C.N., and Arnold, F.H. (2018). Learned protein embeddings for machine learning. *Bioinformatics* 34, 2642–2648. <https://doi.org/10.1093/bioinformatics/bty178>.
- Forster, D.T., Li, S.C., Yashiroda, Y., Yoshimura, M., Li, Z., Isuhuaylas, L.A.V., Itto-Nakama, K., Yamanaka, D., Ohya, Y., Osada, H., et al. (2022). BIONIC: biological network integration using convolutions. *Nat. Methods* 19, 1250–1261. <https://doi.org/10.1038/s41592-022-01616-x>.
- Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M.M., and Correia, B.E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* 17, 184–192. <https://doi.org/10.1038/s41592-019-0666-6>.
- wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 47, D520–D528. <https://doi.org/10.1093/nar/gky949>.
- Yap, H.-Y., Tee, S.Z.-Y., Wong, M.M.-T., Chow, S.-K., Peh, S.-C., and Teow, S.-Y. (2018). Pathogenic role of immune cells in rheumatoid arthritis: Implications in clinical treatment and biomarker development. *Cells* 7, 161. <https://doi.org/10.3390/cells7100161>.
- Chang, J.T. (2020). Pathophysiology of inflammatory bowel diseases. *J. Med.* 383, 2652–2664. <https://doi.org/10.1056/NEJMra2002697>.
- Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell* 165, 535–550. <https://doi.org/10.1016/j.cell.2016.03.014>.