Contextual AI models for context-specific prediction in biology

We developed PINNACLE, a graph-based AI model for learning protein representations across cell-type contexts. These contextualized protein representations enable the integration of 3D protein structure with single-cell genomic-based representations to enhance protein–protein interaction prediction, analysis of drug effects across cell-type contexts, and prediction of therapeutic targets in a cell type-specific manner.

This is a summary of:
Li, M. M. et al. Contextual AI models for single-cell protein biology. *Nat. Methods* https://doi.org/10.1038/s41592-024-02341-3 (2024)

Published online: 22 July 2024

The problem

Proteins perform a myriad of functions that often vary across cellular environments[1]. Despite each cell having a nearly identical genome, gene expression and the encoded protein functionality in a cell are influenced by its specific context, including its type, health status and disease condition. This underscores the need to consider cellular contexts to develop comprehensive models of protein functionality[2]. Computational methods that incorporate cellular contexts could offer precise context-specific insights, enhancing protein characterization.

However, most artificial intelligence (AI) methods lack this contextual specificity. Context-agnostic AI methods for protein characterization build an integrated summary signature (or an embedding) for each protein. These context-free embeddings are used for prediction of protein function, yet they fail to account for the contextual nuances of their environments, such as variations among cell types or disease states. This limitation hampers the ability to precisely predict protein functions across cellular contexts. To generate protein representations that incorporate biological context, contextual AI models must capture and distinguish context-specific patterns, thereby enabling the models to dynamically adjust their outputs to the contexts in which they operate.

The solution

We have developed PINNACLE, a contextual AI model that produces cell type context-specific protein representations. At its core, PINNACLE uses a contextual graph neural network to map the relationships among proteins, cell types and tissues (Fig. 1a). By integrating single-cell datasets with a protein–protein interaction network (PPIN), cell-type communication network and tissue hierarchy network, PINNACLE navigates PPINs across cell types from diverse tissues. Unlike models that offer context-agnostic protein representations, PINNACLE's approach produces protein representations that capture relationships in cell type-specific PPINs (here, constructed by combining single-cell transcriptomic data with a reference PPIN). Training PINNACLE involves self-supervised learning at the protein, cell-type and tissue levels to produce an embedding space where every protein gets multiple embeddings and each embedding is tailored to a different cell type (Fig. 1a). The resulting embedding space captures context-specific PPINs, alongside cellular and tissue structures.

PINNACLE learns an embedding space of cell type-specific PPINs for which the proximity between protein embeddings reflects cell-type communication and tissue organization. This enables PINNACLE to support a wide range of applications, from understanding protein–protein interactions, which are influenced by both 3D structural conformations and cell-type contexts[3], to enhancing protein binding affinity predictions (Fig. 1b). For instance, we show that combining contextualized representations with context-free 3D structure representations of PD-1–PD-L1 and B7-1–CTLA-4 better differentiates between binding and non-binding proteins without additional training. We also adapt PINNACLE to nominate therapeutic targets for rheumatoid arthritis and inflammatory bowel diseases, demonstrating improved predictive performance over context-free models while pinpointing the cell types with the highest predictive relevance.

Future directions

PINNACLE is an AI model for multipurpose contextualized prediction in user-defined biological contexts. Because it adjusts its outputs on the basis of contexts in which it operates, PINNACLE is well-suited for a wide spectrum of applications where cell-type dependencies and cell type-specific mechanisms influence protein function.

As PINNACLE is trained on single-cell data from healthy individuals, it would need to be retrained on disease-specific single-cell and proteomic data to specialize on certain disease contexts. To model more diverse cell types than those provided by Tabula Sapiens (a human multi-organ single-cell transcriptomic atlas), PINNACLE will need to be scaled to single-cell datasets that characterize millions of cells and hundreds of cell types. One such data resource is CELLxGENE, with over 80 million cells representing over 800 cell types[4]. PINNACLE is currently trained on PPINs, yet it can easily be applied to networks generated from other data modalities like ATAC-seq to enable advanced contextualized predictions.

Looking ahead, understanding protein functions and developing molecular therapies will require a comprehensive understanding of proteins' roles in different cell types and the interactions between proteins across diverse cell-type contexts[5]. Beyond nominating therapeutic targets, we envision a family of models, such as PINNACLE, tailored to diverse biological contexts and used for context-specific tasks, from identifying disease biomarkers to predicting therapeutic effects.

Michelle M. Li & Marinka Zitnik
Harvard University, Cambridge, MA, USA.

EXPERT OPINION

"This manuscript describes a method to obtain contextual protein embedding by considering not only protein–protein interactions but also between-cell and tissue interaction as a meta-graph using graph attention networks for each context. The paper is very timely, is technically interesting and addresses an important and challenging problem." **An anonymous reviewer.**

FIGURE

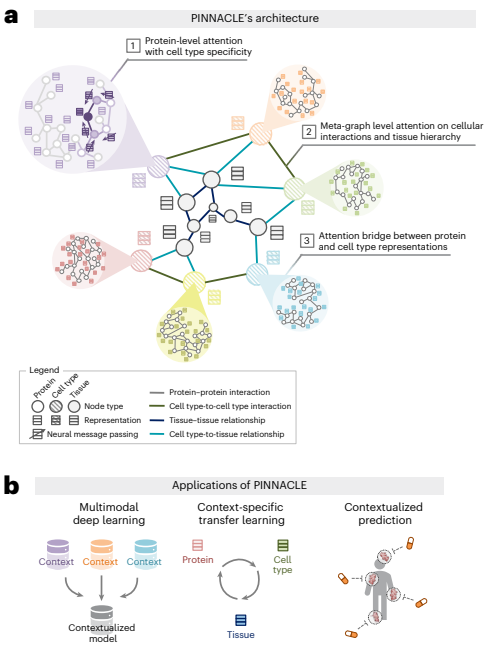

Fig. 1 | PINNACLE, a model for learning contextual protein representations. a, PINNACLE takes as input context-specific protein networks (periphery; in purple, orange, green, blue, yellow, red) and a network of cell-type interactions and tissue hierarchy (center). Leveraging attention-based graph neural network architecture, PINNACLE performs message passing (arrows) on the networks of proteins, cell types and tissues (via modules 1–3, which attend to protein, cell type and tissue nodes with highest predictive value) and generates cell type-specific protein representations that capture cellular and tissue relationships. PINNACLE produces multiple representations for every protein, each tailored for a different context. **b**, PINNACLE supports multimodal deep learning, context-specific transfer learning and contextualized prediction.

BEHIND THE PAPER

The most challenging aspect of this study is the lack of large-scale single-cell proteomic data to enable context-specific prediction tasks of proteins across cell types. We instead construct approximate cell type-specific PPINs by integrating single-cell transcriptomic data with a context-free PPIN. To robustly demonstrate the capabilities and impact of contextual learning on biology and medicine, we nominate therapeutic targets for two therapeutic areas with cell type-specific mechanisms of action. We show that PINNACLE both outperforms context-free methods and generates meaningful cell type-specific hypotheses. Such strong performance gains by PINNACLE indicate the importance of contextualized predictions, as well as the need for single-cell proteomic atlases and accurate cell type-specific protein networks.

Alongside PINNACLE, the contextual learning algorithm, we share the cell type-specific PPINs and two context-specific benchmark therapeutic datasets. We hope they will facilitate further advances in contextual learning in biology and medicine. **M.M.L. & M.Z.**

REFERENCES

1. Pan, J. et al. Sparse dictionary learning recovers pleiotropy from human cell fitness screens. *Cell Syst.* **13**, 286–303 (2022).
   This paper describes a framework to disentangle gene functions that vary across cell contexts defined by fitness screens.
2. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
   This paper introduces a deep learning model using sequence and structural contexts for pathogenicity prediction of missense variants.
3. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
   This paper describes the AlphaFold model for 3D structure protein folding prediction, which incorporates physical and biological priors into protein modeling.
4. CZI Single-Cell Biology Program et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at https://doi.org/10.1101/2023.10.30.563174 (2023).
   This paper describes a data platform with curated and interoperable single-cell datasets across healthy and disease states.
5. Sheridan, C. Can single-cell biology realize the promise of precision medicine? *Nat. Biotechnol.* **42**, 159–162 (2024).
   This news article outlines the current state of the single-cell biology field and its potential for precision medicine.

FROM THE EDITOR

"PINNACLE stands out for its ability to generate contextualized protein representations that allow looking at protein interactions within the context of different cell types and tissue hierarchies. This is a technically interesting attempt to understand protein functional characteristics." **Arunima Singh, Senior Editor, *Nature Methods*.**