Contextual AI models for single-cell protein biology

Michelle M. Li[1], Yepeng Huang[1], Marissa Sumathipala[1], Man Qing Liang[1], Alberto Valdeolivas[2], Ashwin N. Ananthakrishnan[1,3], Katherine Liao[1,4], Daniel Marbach[2], and Marinka Zitnik[1,5,6,7‡]

[1]Harvard Medical School, Boston, MA, USA
[2]Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland
[3]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA
[4]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA, USA
[5]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA
[6]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[7]Harvard Data Science Initiative, Cambridge, MA, USA

‡Corresponding author. Email: marinka@hms.harvard.edu

Understanding protein function and developing molecular therapies require deciphering the cell types in which proteins act as well as the interactions between proteins. However, modeling protein interactions across biological contexts remains challenging for existing algorithms. Here, we introduce PINNACLE, a geometric deep learning approach that generates context-aware protein representations. Leveraging a multi-organ single-cell atlas, PINNACLE learns on contextualized protein interaction networks to produce 394,760 protein representations from 156 cell type contexts across 24 tissues. PINNACLE's embedding space reflects cellular and tissue organization, enabling zero-shot retrieval of the tissue hierarchy. Pretrained protein representations can be adapted for downstream tasks: enhancing 3D structure-based representations for resolving immuno-oncological protein interactions, and investigating drugs' effects across cell types. PINNACLE outperforms state-of-the-art models in nominating therapeutic targets for rheumatoid arthritis and inflammatory bowel diseases, and pinpoints cell type contexts with higher predictive capability than context-free models. PINNACLE's ability to adjust its outputs based on the context in which it operates paves way for diverse large-scale context-specific predictions in biology.

Introduction

Proteins are the functional units of cells, and their interactions enable different biological functions. The development of high-throughput methods has facilitated the characterization of large maps of protein interactions. Leveraging these protein interaction networks, computational methods [1, 2] have been developed to improve the understanding of protein structure [3], accurately predict functional annotations [4, 5], and inform the design of therapeutic targets [6, 7]. Among them, representation learning methods have emerged as a leading strategy to model proteins [8–10]. These approaches can resolve protein interaction networks across tissues [11–13] and cell types by integrating molecular cell atlases [14] and extending our understanding of the relationship between protein and function [15]. Protein representation learning methods can predict multicellular functions across human tissues [12], design target-binding proteins [16] and novel protein interactions [17], and predict interactions between transcription factors and genes [15].

Proteins can have distinct roles in different biological contexts [18, 19]. While nearly every cell contains the same genome, the expression of genes and the function of proteins encoded by these genes depend on cellular and tissue contexts [11, 20, 21]. Gene expression and the function of proteins can also differ significantly between healthy and disease states [21, 22]. Methods incorporating biological contexts can improve the characterization of proteins and provide precise, context-specific insights. However, deep learning methods produce protein representations (or embeddings) that are *context-free:* each protein has only one representation learned from either a single context or an integrated view across many contexts [15, 23]. These methods generate one representation for each protein, providing an integrated summary. Context-free protein representations are not tailored to specific biological contexts, such as cell types and disease states. These representations cannot identify protein functions that vary across different cell types, which in turn hamper the prediction of pleiotropy and protein roles in a cell type-specific manner.

Sequencing technologies that measure gene expression with single-cell resolution pave the way toward addressing this challenge. Single-cell transcriptomic atlases [20, 24–27] measure activated genes across many cellular contexts. Through attention-based deep learning [28, 29], which specify models that can pay attention to large inputs and learn the most important elements to focus on in each context, single-cell atlases can be leveraged to boost the mapping of gene regulatory networks that drive disease progression and reveal treatment targets [30]. However, incorporating the expression of protein-coding genes into protein interaction networks remains a challenge. Existing algorithms, including protein representation learning, cannot contextualize protein representations.

We introduce PINNACLE (Protein Network-based Algorithm for Contextual Learning), a context-specific model for comprehensive protein understanding. PINNACLE is a geometric deep learning model adept at generating protein representations through the analysis of protein interactions within various cellular contexts. Leveraging single-cell transcriptomics combined with networks of protein-protein interactions, cell-type-to-cell-type interactions, and a tissue hierarchy, PINNACLE generates high-resolution protein representations tailored to each cell type. In contrast to existing methods that provide a single representation for each protein, PINNACLE generates a distinct representation for each cell type in which a protein-coding gene is activated. With 394,760 contextualized protein representations produced by PINNACLE, where each protein representation is imbued with cell type specificity, we demonstrate PINNACLE's capability to integrate protein interactions with the underlying protein-coding gene transcriptomes of 156 cell type contexts. PINNACLE models support a broad array of tasks; they can enhance three-dimensional (3D) structural protein representations, analyze the effects of drugs across cell type contexts, nominate therapeutic targets in a cell type specific manner, retrieve tissue hierarchy in a zero-shot manner, and perform context-specific transfer learning. PINNACLE models dynamically adjust their outputs based on the context in which they operate and can pave the way for the broad use of foundation models tailored to diverse biological contexts.

Results

Constructing context-specific networks.

Generating protein representations embedded with cell type context calls for protein interaction networks that consider the same context. We assembled a dataset of context-sensitive protein interactomes, beginning with a multi-organ single-cell transcriptomic atlas [20] that encompasses 24 tissue and organ samples sourced from 15 human donors (Figure 1a). We compile activated genes for every expert-annotated cell type in this dataset by evaluating the average gene expression in cells from that cell type relative to a designated reference set of cells (Figure 1a; Methods 2). Here, 'activated genes' are defined as those demonstrating a higher average expression in cells annotated as a particular type than the remaining cells documented in the dataset. Based on these activated gene lists, we extracted the corresponding proteins from the comprehensive reference protein interaction network and retained the largest connected component (Figure 1a). As a result, we have 156 context-aware protein interaction networks, each with $2,530 \pm 677$ proteins, that are maximally similar to the global reference protein interaction network and still highly cell type spe-

cific (Supplementary Figures S1-S2). Our context-aware protein interaction networks from 156 cell type contexts span 62 tissues of varying biological scales.

Further, we constructed a network of cell types and tissues (metagraph) to model cellular interactions and the tissue hierarchy (Methods 2). Given the cell type annotations designated by the multi-organ transcriptomic atlas [20], the network consists of 156 cell type nodes. We incorporated edges between pairs of cell types based on the existence of significant ligand-receptor interactions and validated that the proteins correlating to these interactions are enriched in the context-aware protein interaction networks in comparison to a null distribution (Methods 2; Supplementary Figure S1c-d). Leveraging information on tissues in which the cell types were measured, we began with 24 tissue nodes and established edges between cell type nodes and tissue nodes if the cell type was derived from the corresponding tissue. We then identified all ancestor nodes, including the root, of the 24 tissue nodes within the tissue hierarchy (Methods 2) to feature 62 tissue nodes interconnected by parent-child relationships. Our dataset thus comprises 156 context-aware protein interaction networks and a metagraph reflecting cell type and tissue organization.

Overview of PINNACLE model.

PINNACLE is a geometric deep learning model capable of generating protein representations predicated on protein interactions within a spectrum of cell type contexts. Trained on an integrated set of context-aware protein interaction networks, complemented by a network capturing cellular interactions and tissue hierarchy (Figure 1b-c), PINNACLE generates contextualized protein representations that are tailored to cell types in which protein-coding genes are activated (Figure 1d). Unlike context-free models, PINNACLE produces multiple representations for every protein, each contingent on its specific cell type context. Additionally, PINNACLE produces representations of the cell type contexts and representations of the tissue hierarchy (Figure 1d-e). This approach ensures a multifaceted understanding of protein interaction networks, taking into account the myriad of contexts in which proteins function.

Given multi-scale model inputs, PINNACLE learns the topology of proteins, cell types, and tissues by optimizing a unified latent representation space. PINNACLE integrates different context-specific data into one context-aware model (Figure 1f) and transfers knowledge between protein-, cell type- and tissue-level data to contextualize representations (Figure 1g). To infuse cellular and tissue organization into this embedding space, PINNACLE employs protein-, cell type-, and tissue-level attention along with respective objective functions (Figure 1b-c; Methods 3). Conceptually, pairs of proteins that physically interact (i.e., are connected by edges in input networks) are closely

embedded. Similarly, proteins are embedded near their respective cell type contexts while maintaining a substantial distance from unrelated ones. This ensures that interacting proteins within the same cell type context are situated proximally within the embedding space yet are separated from proteins from other cell type contexts. This approach yields an embedding space that accurately represents the intricacies of relationships between proteins, cell types, and tissues.

PINNACLE disseminates graph neural network messages between proteins, cell types, and tissues using a series of attention mechanisms tailored to each specific node and edge type (Methods 3). The protein-level pretraining tasks consider self-supervised link prediction on protein interactions and cell type classification on protein nodes. These tasks enable PINNACLE to sculpt an embedding space that encapsulates the topology of the context-aware protein interaction networks and the cell type identity of the proteins. PINNACLE's cell type- and tissue-specific pretraining tasks rely exclusively on self-supervised link prediction, facilitating the learning of cellular and tissue organization. The topology of cell types and tissues is imparted to the protein representations through an attention bridge mechanism, effectively enforcing tissue and cellular organization onto the protein representations. PINNACLE's contextualized protein representations capture the structure of context-aware protein interaction networks. The regional arrangement of these contextualized protein representations in the latent space reflects the cellular and tissue organization represented by the metagraph. This leads to a comprehensive and context-specific representation of proteins within a unified cell type- and tissue-specific framework.

PINNACLE embeddings capture cellular and tissue organization.

PINNACLE generates protein representations for each of the 156 cell type contexts spanning 62 tissues of varying hierarchical scales. In total, PINNACLE's unified multi-scale embedding space comprises 394,760 protein representations, 156 cell type representations, and 62 tissue representations (Figure 1a). We show that PINNACLE learns an embedding space where proteins are positioned based on cell type context. We first quantify the spatial enrichment of PINNACLE's protein embedding regions using a systematic method, SAFE [31] (Methods 6.3). PINNACLE's contextualized protein representations self-organize in PINNACLE's embedding space as evidenced by the enrichment of spatial embedding regions for protein representations that originate from the same cell type context (significance cutoff $\alpha = 0.05$; Figure 2; Supplementary Figures S3-S4).

Next, we evaluate embedding regions to confirm that they are separated by cell type and tissue identity by calculating the similarities between protein representations across cell type contexts. Protein representations from the same cell type are more similar than those from different

cell types (Figure 3a). In contrast, a model without cellular or tissue context fails to capture any differences between protein representations across cell type contexts (Figure 3b). Further, we expect the representations of proteins that act on multiple cell types to be highly dissimilar, reflecting specialized cell type-specific protein functions (Supplementary Note S1). We calculate the similarities of protein representations (i.e., cosine similarities of a protein's representations across cell type contexts) based on the number of cell types in which the protein is active (Supplementary Figure S5a-b). Representational similarities of proteins negatively correlate with the number of cell types in which they act (Spearman's $\rho = -0.9798$; p-value < 0.001), and the correlation is weaker in the ablated model with cellular and tissue metagraph turned off (Spearman's $\rho = -0.6334$; p-value < 0.001).

We additionally examine whether protein embedding regions are organized by the tissue hierarchy. We leverage PINNACLE's tissue representations to perform zero-shot retrieval of the tissue hierarchy and then compare tissue ontology distance to tissue embedding distance. Tissue ontology distance is defined as the sum of the shortest path lengths from two tissue nodes to the lowest common ancestor node in the tissue hierarchy, and tissue embedding distance is the cosine distance between the corresponding tissue representations. We expect a positive correlation: the farther apart the nodes are according to the tissue hierarchy, the more dissimilar the tissue representations are. As hypothesized, embedding distances in the latent space and the corresponding distances in the tissue ontology of the same tissues are positively correlated (Spearman's $\rho = 0.36$; p-value $= 1.85 \times 10^{-119}$; Figure 3c), and the distribution of tissue embedding distances cannot be attributed to random effects (Kolmogorov-Smirnov two-sided test $= 0.50$; p-value < 0.001). When the tissue ontology is randomly shuffled, the correlation with distances in the embedding space diminishes significantly (Spearman's $\rho = 0.005$; p-value $= 0.349$; Figure 3c). Since PINNACLE uses the metagraph to systematically integrate tissue organization into both cell type and protein representations, it follows that all of PINNACLE's representations inherently reflect this tissue organization (Methods 3; Supplementary Figure S6).

PINNACLE enhances 3D structural representations of PPIs.

Protein-protein interactions (PPIs) depend on both 3D structure conformations of the proteins [32, 33] and cell type contexts within which the proteins act [34]. However, protein representations produced by existing artificial intelligence (AI) models based on 3D molecular structures lack cell type context information. We hypothesize that incorporating cellular context information can better differentiate binding from non-binding proteins (Figure 3d). Because 3D structures of

molecules (containing precise atom or residue level contact information) provide complementary knowledge to protein-protein interaction networks (summarizing binary interactions between proteins), we expect that context-aware protein interaction networks can improve the ability to differentiate between binding and non-binding proteins across different cell types [35]. As no large-scale dataset with matched structural biology and genomic readouts currently exists to perform systematic analyses, we focus on PD-1/PD-L1 and B7-1/CTLA-4 interacting proteins, important immune checkpoint protein interactors involved in cancer immunotherapies [36].

We compare contextualized and context-free protein representations for binding proteins (i.e., PD-1/PD-L1 and B7-1/CTLA-4) and non-binding proteins (i.e., one of the four binding proteins paired with RalB, RalBP1, EPO, EPOR, C3, or CFH). Cell type context is incorporated into 3D structure-based protein representations [3, 17] by concatenating them with PINNACLE's protein representation (Figure 3e; Methods 4). Context-free protein representations are generated by concatenating 3D structure-based representations [3, 17] with an average of PINNACLE's protein representations across all cell type contexts (Methods 4). Contextualized representations, resulting from a combination of protein representations based on 3D structure and context-aware PPI networks, give scores (via cosine similarity) for binding and non-binding proteins of 0.9690 ± 0.0049 and 0.9571 ± 0.0127, respectively. Using PINNACLE's context-specific protein representations, which have no 3D structure information, binding and non-binding proteins are scored 0.0385 ± 0.1531 and 0.0218 ± 0.1081, respectively. In contrast, using context-free representations, binding and non-binding proteins are scored at 0.9789 ± 0.0004 and 0.9742 ± 0.0078, respectively. Further, comparative analysis of the gap in scores between interacting vs. non-interacting proteins yields gaps of 0.011 (PD-1/PD-L1) and 0.015 (B7-1/CTLA-4) for PINNACLE's contextualized representations (p-value $= 0.0299$; Supplementary Figure S7), yet only 0.003 (PD-1/PD-L1) and 0.006 (B7-1/CTLA-4) for context-free representations (Figure 3f and Supplementary Figure S7). Incorporating information about biological contexts can help better distinguish protein interactions from non-interacting proteins in specific cell types, suggesting that PINNACLE's contextualized representations can enhance protein representations derived from 3D protein structure modality. Modeling context-dependent interactions involving immune checkpoint proteins can deepen our understanding of how these proteins are used in cancer immunotherapies. Our benchmarking results further suggest that incorporating context can improve 3D structure prediction of protein interactions (Supplementary Note S2).

Contextual models outperform context-free target prediction.

With the representations from PINNACLE infused with cellular and tissue context, we can fine-tune them for downstream tasks (Figure 1f-h). We hypothesize that PINNACLE's contextualized latent space can better differentiate between therapeutic targets and proteins with no therapeutic potential than a context-free latent space. Here, we focus on modeling the therapeutic potential of proteins across cell types for therapeutic areas with cell type-specific mechanisms of action (MoA) (Figure 4). Certain cell types are known to play crucial and distinct roles in the disease pathogenesis of rheumatoid arthritis (RA) and inflammatory bowel disease (IBD) therapeutic areas [24, 37–40]. There is currently no cure for either type of condition, and the medications prescribed to mitigate the symptoms can lead to undesired side effects [41]. The new generation of therapeutics in development for RA and IBD conditions is designed to target specific cell types so that the drugs maximize efficacy and minimize adverse events (e.g., by directly impacting the affected/responsible cells and avoiding off-target effects on other cells) [41, 42]. We adopt PINNACLE models to predict the therapeutic potential of proteins in a cell type-specific manner.

We fine-tune PINNACLE to predict therapeutic targets for RA and IBD diseases. Specifically, we perform binary classification on each contextualized protein representation, where $y = 1$ indicates that the protein is a therapeutic candidate for the given therapeutic area and $y = 0$ otherwise. The ground truth positive examples (where $y = 1$) are proteins targeted by drugs that have at least completed one clinical trial of phase two or higher for indications under the therapeutic area of interest, indicating that the drugs are safe and potentially efficacious in an initial cohort of humans (Figure 4a-b). The negative examples (where $y = 0$) are druggable proteins that have not been studied for the therapeutic area (Figure 4b; Methods 5). The binary classification model can be of any architecture; our results for nominating RA and IBD therapeutic targets are generated by a multi-layer perceptron trained for each therapeutic area (Figure 4c).

To evaluate PINNACLE's contextualized protein representations, we compare PINNACLE's fine-tuned models against three context-free models. We apply a random walk algorithm [43] and a graph attention network (GAT) [44] on the context-free reference protein interaction network. The BIONIC model is a graph convolutional neural network designed for (context-free) multi-modal network integration [15].

We find that PINNACLE's protein representations for all cell type contexts outperform the random walk model for both RA (Figure 4d) and IBD (Figure 4e) diseases. Protein representations from 44.9% (70 out of 156) and 37.5% (57 out of 152) cell types outperform the GAT model for RA (Figure 4d) and IBD (Figure 4e) diseases, respectively. Although both PINNACLE and

BIONIC can integrate the 156 cell type-specific protein interaction networks, PINNACLE's protein representations outperform BIONIC [15] in 18.6% of cell types (29 out of 156) and 8.6% of cell types (13 out of 152) for RA (Figure 4d) and IBD diseases (Figure 4e), respectively, highlighting the utility of contextualizing protein representations. PINNACLE outperforms these three context-free models via other metrics for both RA and IBD therapeutic areas (Supplementary Figure S8). We have confirmed no significant correlation between the node degree of proteins in cell type-specific PPI networks and performance in RA and IBD models (Supplementary Figure S9a). Additionally, there is only a moderate correlation between PINNACLE's performance and the enrichment of positive targets in these cell type-specific PPI networks (Supplementary Figure S9b-c). These findings underscore that PINNACLE's predictions cannot be solely ascribed to the characteristics of the cell type-specific PPI networks. Benchmarking results indicate combining global reference networks with advanced deep graph representation learning techniques, such as GAT, can yield better predictors than network-based random walk methods alone. Integrative approaches, exemplified by methods like BIONIC, enhance performance, a finding consistent with the established benefits of data integration. Contextualized learning approaches, like PINNACLE, have the potential to enhance model performance and enable predictions tailored to specific contexts.

PINNACLE can nominate targets across cell type contexts.

There is existing evidence that drug effects vary with cell type depending on where therapeutic targets are expressed and where proteins act [45–49]. For instance, CD19-targeting chimeric antigen receptor T (CAR-T) cell therapy has been highly effective in treating B cell malignancies, yet causes a high incidence of neurotoxicity [47]. A recent study shows that CAR-T cells induce off-target effects by targeting the CD19 expressed in brain mural cells, likely causing the brain barrier leakiness responsible for neurotoxicity [47]. We hypothesize that the predicted protein druggability varies across cell types, and such variations can provide insights into the cell types' relevance for a therapeutic area.

Among the 156 biological contexts modeled by PINNACLE's protein representations, we examine the most predictive cell type contexts for nominating therapeutic targets of RA. We find that the most predictive contexts consist of CD4+ helper T cells, CD4+ $\alpha\beta$ memory T cells, CD1c+ myeloid dendritic cells, gut endothelial cells, and pancreatic acinar cells (Figure 5a). Immune cells play a significant role in the disease pathogenesis of RA [37, 38]. Since CD4+ helper T cells (PINNACLE-predicted rank $= 1$), CD4+ $\alpha\beta$ memory T cells (PINNACLE-predicted rank $= 2$), and CD1c+ myeloid dendritic cells (PINNACLE-predicted rank $= 3$) are immune cells, it is ex-

pected that PINNACLE's protein representations in these contexts achieve high performance in our prediction task. Also, patients with RA often have gastrointestinal (GI) manifestations, whether concomitant GI autoimmune diseases or GI side effects of RA treatment [50]. Pancreatic acinar cells (PINNACLE-predicted rank $= 5$) can behave like inflammatory cells during acute pancreatitis [51], one of the accompanying GI manifestations of RA [50]. In addition to GI manifestations, endothelial dysfunction is commonly detected in patients with RA [52]. While rare, rheumatoid vasculitis, which affects endothelial cells and is a serious complication of RA, has been found to manifest in the large and small intestines (gut endothelial cell context has PINNACLE-predicted rank $= 4$), liver, and gallbladder [50, 53]. Further, many of the implicated cell types for RA patients (e.g., T cells, B cells, natural killer cells, monocytes, myeloid cells, and dendritic cells) are highly ranked by PINNACLE [24, 25, 39] (Supplementary Table S1). Our results suggest that injecting cell type context to protein representations can significantly improve performance in nominating therapeutic targets for RA diseases while potentially revealing the cell types underlying disease processes.

The most predictive cell type contexts for nominating therapeutic targets of IBD are CD4+ $\alpha\beta$ memory T cells, enterocytes of epithelium of large intestine, T follicular helper cells, plasmablasts, and myeloid dendritic cells (Figure 5d). The intestinal barrier comprises a thick mucus layer with antimicrobial products, a layer of intestinal epithelial cells, and a layer of mesenchymal cells, dendritic cells, lymphocytes, and macrophages [54]. As such, these five cell types are expected to yield high predictive ability. Moreover, many of the implicated cell types for IBD (e.g., T cells, fibroblasts, goblet cells, enterocytes, monocytes, natural killer cells, B cells, and glial cells) are highly ranked by PINNACLE [26, 27, 55] (Supplementary Table S2). For example, CD4+ T cells are known to be the main drivers of IBD [56]. They have been found in the peripheral blood and intestinal mucosa of adult and pediatric IBD patients [57]. Patients with IBD tend to develop uncontrolled inflammatory CD4+ T cell responses, resulting in tissue damage and chronic intestinal inflammation [58, 59]. Due to the heterogeneity of CD4+ T cells in patients, treatment efficacy can depend on the patient's subtype of CD4+ T cells [58, 59]. Thus, the highly predictive cell type contexts according to PINNACLE should be further investigated to design safe and efficacious therapies for RA and IBD diseases.

Conversely, we hypothesize that the cell type contexts of protein representations that yield worse performance than the cell type-agnostic protein representations may not have the predictive power (given the current list of targets from drugs that have at least completed phase 2 of clinical

trials) for studying the therapeutic effects of candidate targets for RA and IBD therapeutic areas.

In the context-aware model trained to nominate therapeutic targets for RA diseases, the protein representations of duodenum glandular cells, endothelial cells of hepatic sinusoid, myometrial cells, and hepatocytes performance worse than the cell type-agnostic protein representations (Figure 5a). The RA therapeutic area is a group of inflammatory diseases in which immune cells attack the synovial lining cells of joints [37]. Since duodenum glandular cells (PINNACLE-predicted rank $= 153$), endothelial cells of hepatic sinusoid (PINNACLE-predicted rank $= 126$), myometrial cells (PINNACLE-predicted rank $= 119$), and hepatocytes (PINNACLE-predicted rank $= 116$) are neither immune cells nor found in the synovium, these cell type contexts' protein representations expectedly perform poorly. For IBD diseases, the protein representations of the limbal stem cells, melanocytes, fibroblasts of cardiac tissue, and radial glial cells have worse performance than the cell type-agnostic protein representations (Figure 5d). The IBD therapeutic area is a group of inflammatory diseases in which immune cells attack tissues in the digestive tract [40]. As limbal stem cells (PINNACLE-predicted rank $= 152$), melanocytes (PINNACLE-predicted rank $= 147$), fibroblasts of cardiac tissue (PINNACLE-predicted rank $= 135$), and radial glial cells (PINNACLE-predicted rank $= 107$) are neither immune cells nor found in the digestive tract, these cell type contexts' protein representations should also perform worse than context-free representations.

The least predictive cellular contexts in PINNACLE's models for RA and IBD have no known role in disease, indicating that protein representations from these cell type contexts are poor predictors of RA and IBD therapeutic targets. PINNACLE's overall improved predictive ability compared to context-free models indicates the importance of understanding cell type contexts where therapeutic targets are expressed and act.

Predictive cell type contexts reflect MoAs in RA therapies.

Recognizing and leveraging the most predictive cell type context for examining a therapeutic area can be beneficial for predicting candidate therapeutic targets [45–49]. We find that considering only the most predictive cell type contexts can yield significant performance improvements compared to context-free models (Supplementary Figure S10). We examine cell type contexts selected by PINNACLE as the most predictive for JAK3 and IL6R, two protein targets of RA drugs.

Disease-modifying anti-rheumatic drugs (DMARDs), such as Janus kinase (JAK) inhibitors (i.e., tofacitinib, upadacitinib, and baricitinib), are commonly prescribed to patients with RA [60, 61]. For JAK3, PINNACLE's five most predictive cell type contexts are T follicular helper cells, microglial cells, DN3 thymocytes, CD4+ $\alpha\beta$ memory T cells, and hematopoietic stem cells (Fig-

ure 5b). Since the expression of JAK3 is limited to hematopoietic cells, mutations or deletions in JAK3 tend to cause defects in T cells, B cells, and NK cells [62–65]. For instance, patients with JAK3 mutations tend to be depleted of T cells [63], and the abundance of T follicular helper cells is highly correlated with RA severity and progression [66]. JAK3 is also highly expressed in double negative (DN) T cells (early stage of thymocyte differentiation) [67], and the levels of DN T cells are higher in synovial fluid than peripheral blood, suggesting a possible role of DN T cell subsets in RA pathogenesis [68]. Lastly, dysregulation of the JAK/STAT pathway, which JAK3 participates in, has pathological implications for neuroinflammatory diseases, a significant component of disease pathophysiology in RA [69, 70].

Tocilizumab and sarilumab are FDA-approved for treating RA, and target the interleukin six receptor, IL6R [61]. For IL6R, PINNACLE's five most predictive cellular contexts are classical monocytes, NAMPT neutrophils, intermediate monocytes, mesenchymal stem cells, and regulatory T cells (Figure 5c). IL6R is predominantly expressed on neutrophils, monocytes, hepatocytes, macrophages, and some lymphocytes [71]. IL6R simulates the movement of T cells and other immune cells to the site of infection or inflammation [72] and affects T cell and B cell differentiation [71, 73]. IL6 acts directly on neutrophils, essential mediators of inflammation and joint destruction in RA, through membrane-bound IL6R [71]. Experiments on fibroblasts isolated from the synovium of RA patients show that anti-IL6 antibodies prevented neutrophil adhesion, indicating a promising therapeutic direction for IL6R on neutrophils [71]. Lastly, mice studies have shown that pre-treatment of mesenchymal stem/stromal cells with soluble IL6R can enhance the therapeutic effects of mesenchymal stem/stromal cells in arthritis inflammation [74].

PINNACLE's hypotheses to examine JAK3 and IL6R in the highly predictive cell type contexts, according to PINNACLE, to maximize therapeutic efficacy seem to be consistent with their roles in the cell types. It seems that targeting these proteins may directly impact the pathways contributing to the pathophysiology of RA therapeutic areas. Further, our results for IL6R suggest that PINNACLE's contextualized representations could be leveraged to evaluate potential enhancement in efficacy (e.g., targeting multiple points in a pathway of interest).

Predictive cell type contexts elucidate MoAs in IBD therapies.

Like RA, we must understand the cells in which therapeutic targets are expressed and act to maximize the efficacy of targeted IBD therapies [75]. To support our hypothesis, we evaluate PINNACLE's predictions for two protein targets of commonly prescribed treatments for IBD diseases: ITGA4 and PPARG.

Vedolizumab and natalizumab target the integrin subunit alpha 4, ITGA4, to treat the symptoms of IBD therapeutic area [61]. PINNACLE's five most predictive cell type contexts for ITGA4 are regulatory T cells, dendritic cells, myeloid dendritic cells, granulocytes, and CD8+ $\alpha\beta$ cytotoxic T cells (Figure 5e). Integrins mediate the trafficking and retention of immune cells to the gastrointestinal tract; immune activation of integrin genes increases the risk of IBD [76]. For instance, ITGA4 is involved in homing memory and effector T cells to inflamed tissues, including intestinal and non-intestinal tissues, and imbalances in regulatory and effector T cells may lead to inflammation [77]. Circulating dendritic cells express the gut homing marker encoded by ITGA4; the migration of blood dendritic cells to the intestine allows these dendritic cells to become mature, which leads to gut inflammation and tissue damage, indicating that future studies are warranted to elucidate the functional properties of blood dendritic cells in IBD [78].

Balsalazide and mesalamine are aminosalicylate drugs (DMARDs) commonly used to treat ulcerative colitis by targeting peroxisome proliferator activated receptor gamma (PPARG) [61, 79]. PINNACLE's five most predictive cell types for PPARG are paneth cells of the epithelium of large intestines, endothelial cells of the vascular tree, classic monocytes, goblet cells of small intestines, and serous cells of epithelium of bronchus (Figure 5f). PPARG is highly expressed in the gastrointestinal tract, higher in the large intestine (e.g., colonic epithelial cells) than the small intestine [80–82]. In patients with ulcerative colitis, PPARG is often substantially downregulated in their colonic epithelial cells [82]. PPARG promotes enterocyte development [83] and intestinal mucus integrity by increasing the abundance of goblet cells [82]. Further, PPARG activation can inhibit endothelial inflammation in vascular endothelial cells [84, 85], which is significant due to the importance of vascular involvement in IBD [86]. Additionally, PPARG agonists have been shown to act as negative regulators of monocytes and macrophages, which can inhibit the production of proinflammatory cytokines [87]. Intestinal mononuclear phagocytes, such as monocytes, play a major role in maintaining epithelial barrier integrity and fine-tuning mucosal immune system responsiveness [88]. Studies show that newly recruited monocytes in inflamed intestinal mucosa drive the immunopathogenesis of IBD, suggesting that blocking monocyte recruitment to the intestine could be one avenue for therapeutic development [88]. Lastly, PPARG is found to regulate mucin and inflammatory factors in bronchial epithelial cells [89]. Given the pulmonary complications of IBD, PPARG could be a promising target to investigate for treating IBD and pulmonary symptoms [90]. The predictive power of cell type contexts to examine ITGA4 and PPARG, according to PINNACLE, for IBD therapeutic development are thus well-supported.

Discussion

PINNACLE is a flexible geometric deep learning approach for contextualized prediction in user-defined biological contexts. Integrating single-cell transcriptomic atlases with the protein interactome, cell type interactions, and tissue hierarchy, PINNACLE produces latent protein representations specialized to each biological context. PINNACLE's protein representations capture cellular and tissue organization spanning 156 cell types and 62 tissues of varying hierarchical scales. In addition to multi-modal data integration, a pretrained PINNACLE model generates protein representations that can be used for downstream prediction on tasks where cell type dependencies and cell type-specific mechanisms are relevant.

One limitation of the study is the use of the human protein interactome, which is not measured in a cell type-specific manner [91]. No systematic measurements of protein interactions across cell types exist. So, we create cell type-specific protein interaction networks by overlaying single-cell measurements on the reference protein interaction network, leveraging previously validated techniques for the reconstruction of cell-type-specific interactomes at single-cell resolution [14] and conducting sensitivity network analyses to confirm the validity of the networks used to train PINNACLE models (Supplementary Figures S2-S3). This approach enriches networks for cell type-relevant proteins (Supplementary Figure S2). The resulting networks may contain false-positive protein interactions (e.g., proteins that interact in the reference protein interaction network but do not interact in a specific cell type) and false-negative protein interactions (e.g., proteins that interact only within a particular cell type context that may be unknown, or even undiscovered). Further, PINNACLE does not currently model proteins that may play a role in the cell type yet are unaffected by cell type specificity. Nevertheless, strong performance gains of PINNACLE over context-free models indicate the importance of contextualized prediction and suggest a direction to enhance existing analyses on protein interaction networks [4, 6, 7].

We can leverage and extend PINNACLE in many ways. PINNACLE can accommodate and supplement diverse data modalities. We developed PINNACLE models using Tabula Sapiens [20], a molecular reference atlas comprising almost 500,000 cells from 24 distinct tissues and organs. However, since the tissues and cell types associated with specific diseases may not be entirely represented in the atlas of healthy human subjects, we anticipate that our predictive power may be limited. Tabula Sapiens does not include synovial tissues associated with RA disease progression [25, 39], but these can be found in synovial RA atlases [92] and stromal cells obtained from individuals with chronic inflammatory diseases [93]. To enhance the predictive ability of PINNACLE models,

they can be trained on disease-specific or perturbation-specific networks. In this study, PINNACLE representations capture physical interactions between proteins at the cell type level (Supplementary Note S3); PINNACLE can also be applied to cell type-specific protein networks created from other modalities, such as cell type-specific gene expression networks [94]. We show that PINNACLE's representations can supplement protein representations generated from other data modalities, including protein molecular surfaces [3, 17]. While this study is concentrated on protein-coding genes, information on protein isoforms and differential information, such as alternative splicing or allosteric changes, can be used with PINNACLE when such data becomes available. In addition to prioritizing candidate therapeutic targets, PINNACLE's representations can be fine-tuned to identify populations of cells with specific characteristics, such as drug resistance [95], adverse drug events [96], or disease progression biomarkers [97]. Lastly, to move towards a "lab-in-the-loop" framework, where computational and experimental scientists can iteratively refine the machine learning model and validate hypotheses via experiments, recent techniques on conformal prediction [98] and evidential layers can be integrated with PINNACLE to quantify the uncertainty of model outputs.

Existing protein representation learning models are context-free and are limited in directly analyzing protein phenotypes that are resolved by contexts and vary with cell types and tissues. To address this limitation, we introduce PINNACLE that produces specialized protein representations tailored to cell type contexts. We demonstrate that contextual learning can provide a more comprehensive understanding of protein roles across cell type contexts [99]. As experimental technologies advance, it is becoming feasible to generate adaptive protein representations across cell type contexts and leverage contextualized representations to predict cell type specific protein functions and nominate therapeutic candidates at the cell type level. Looking to the future, understanding protein functions and developing molecular therapies will require a comprehensive understanding of the roles that proteins have in different cell types as well as the interactions between proteins across diverse cell type contexts [100]. Approaches, such as PINNACLE, can pave the way to realizing this future potential by generating contextualized protein representations and using them to predict cell type-specific protein functions and nominate therapeutic targets at the cell type level.

Acknowledgements. We thank Alexandros Xenos for his valuable feedback on analyses of cell-type-specific and tissue-agnostic protein functions. M.M.L. is supported by T32HG002295 from the National Human Genome Research Institute and a National Science Foundation Graduate Research Fellowship. M.M.L. and M.Z. gratefully acknowledge the support of NIH R01HD108794, US DoD FA8702-15-D-0001, awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, Bayer Early Excellence in Science, AstraZeneca Research, Roche Alliance with Distinguished Scientists, and Kempner Institute for the Study of Natural and Artificial Intelligence. A.N.A. gratefully acknowledges the support of NIH R01DK127171. K.L. gratefully acknowledges the support of NIH P30 AR072577. The content is solely the responsibility of the authors. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors contributions. M.M.L. retrieved and processed Tabula Sapiens, the global reference protein interaction network, CellPhoneDB repository of ligand-receptor interactions, and the tissue hierarchy to construct the cell type specific protein interaction networks and metagraph of cell type and tissue relationships. M.M.L. and M.S. performed the network analysis. M.M.L. and Y.H. retrieved and processed the OpenTargets data. M.M.L. developed, implemented, and benchmarked PINNACLE, Y.H. improved the scalability of PINNACLE, and M.M.L. and Y.H. performed detailed analyses of PINNACLE's algorithm. A.V. and D.M. advised the network construction and the analysis of PINNACLE's outputs. M.Q.L., K.L., and A.N.A. provided clinical expertise on using PINNACLE for predicting therapeutic targets in a cell type specific manner and interpreting the resulting cell type contexts for rheumatoid arthritis and inflammatory bowel disease. M.M.L. and M.Z. designed the study. All authors contributed to writing the manuscript.

Competing interests. D.M. and A.V. are currently employed by F. Hoffmann-La Roche Ltd. The remaining authors declare no competing interests.

Figure Captions

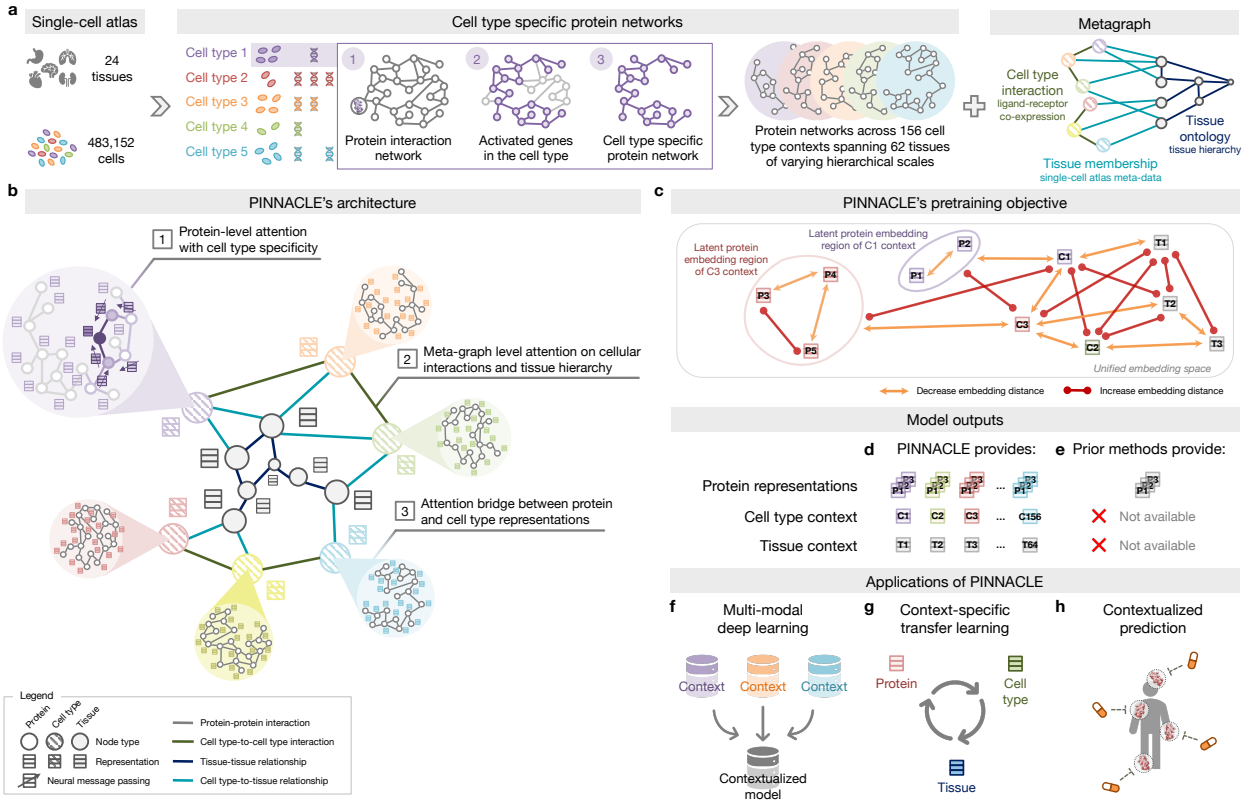

Figure 1: Overview of PINNACLE. (a) Cell type-specific protein interaction networks and metagraph of cell type and tissue organization are constructed from a multi-organ single-cell transcriptomic atlas of humans, a human reference protein interaction network, and a tissue ontology. **(b)** PINNACLE has protein-, cell type-, and tissue-level attention mechanisms that enable the algorithm to generate contextualized representations of proteins, cell types, and tissues in a single unified embedding space. **(c)** PINNACLE is designed such that the nodes (i.e., proteins, cell types, and tissues) that share an edge are embedded closer (decreased embedding distance) to each other than nodes that do not share an edge (increased embedding distance); proteins activated in the same cell type are embedded more closely (decreased embedding distance) than proteins activated in different cell types (increased embedding distance); and cell types are embedded closer to their activated proteins (decreased embedding distance) than other proteins (increased embedding distance). **(d)** As a result, PINNACLE generates protein representations injected with cell type and tissue context; a unique representation is produced for each protein activated in each cell type. PINNACLE simultaneously generates representations for cell types and tissues. **(e)** Existing methods, however, are context-free. They generate a single embedding per protein, representing only one condition or context for each protein, without any notion of cell type or tissue context. **(f-h)** The PINNACLE algorithm and its outputs enable **(f)** multi-modal deep learning (e.g., single-cell transcriptomic data with interactomes), **(g)** context-specific transfer learning (e.g., between proteins, cell types, and tissues), and **(h)** contextualized predictions (e.g., efficacy and safety of therapeutics).

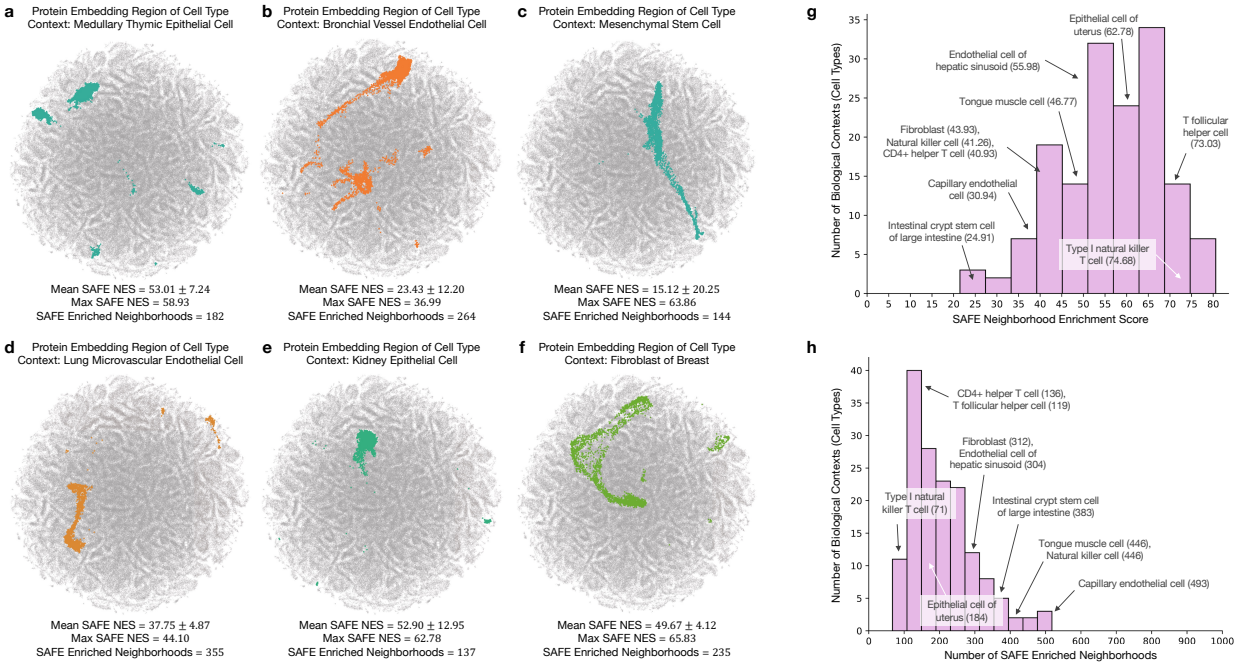

Figure 2: Enrichment of PINNACLE's protein embedding regions. (a-f) Shown are two-dimensional UMAP plots of contextualized protein representations generated by PINNACLE. Each dot is a protein representation, and non-gray colors indicate the cell type context. Gray dots are representations of proteins from other cell types. Each protein embedding region is expected to be enriched neighborhoods that are spatially localized according to cell type context. To quantify this, we compute spatial enrichment of each protein embedding region using SAFE [31], and provide the mean and max neighborhood enrichment scores (NES) and the number of enriched neighborhoods output by the tool (Methods 6 and Supplementary Figure S3-S4). **(g-h)** Distribution of **(g)** the maximum SAFE NES and **(h)** the number of enriched neighborhoods for 156 cell type contexts (each context has p-value < 0.05; hypergeometric test, adjusted using the Benjamin-Hochberg false discovery rate correction with significance cutoff $\alpha = 0.05$). 10 randomly sampled cell type contexts are annotated, with their maximum SAFE NES or number of enriched neighborhoods in parentheses.

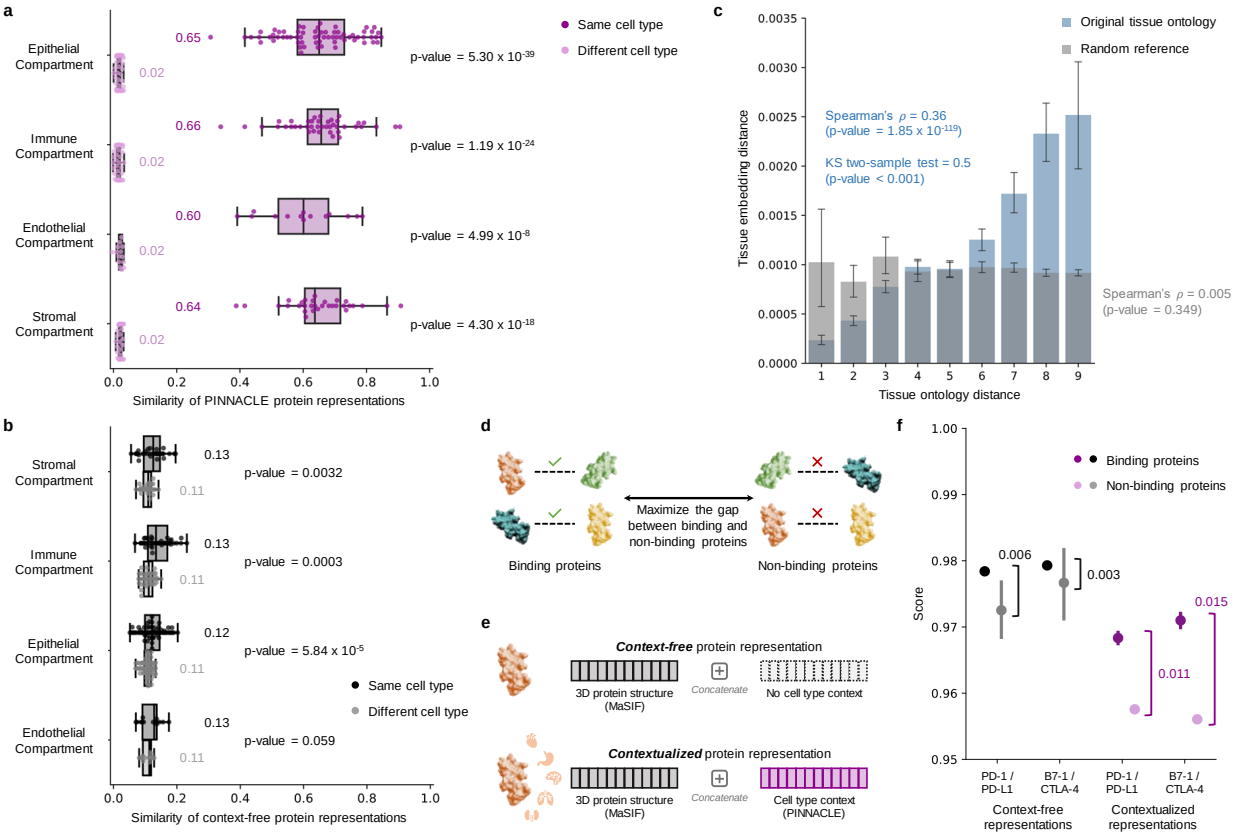

Figure 3: Evaluation of PINNACLE's contextual representations. (a-b) Gap between embedding similarities using **(a)** PINNACLE's protein representations and **(b)** a non-contextualized model's protein representations on $n = 394,760$ samples (i.e., cell type specific protein representations). Similarities are calculated between pairs of proteins in the same cell type (dark shade of color) or different cell types (light shade of color), and stratified by the compartment from which the cell types are derived. We use the two-sided two-sample Kolmogorov-Smirnov test for goodness of fit. Annotations indicate median values. The non-contextualized model is an ablated version of PINNACLE without any notion of tissue or cell type organization (i.e., remove cell type and tissue network and all cell type- and tissue-related components of PINNACLE's architecture and objective function). The bounds of the box show the quartiles of the data, the center indicates the median value of the data, and the whiskers represent the farthest data point within $1.5 \times$ IQR. **(c)** Embedding distance of PINNACLE's 62 tissue representations as a function of tissue ontology distance. Gray bars indicate a null distribution (refer to Methods 6 for more details). Both the Spearman correlation (p-value $= 1.85 \times 10^{-119}$) and Kolmogorov-Smirnov (p-value < 0.001) statistical tests are two-sided. Data are represented as mean values with error bars indicating a 95% confidence interval. **(d)** Prediction task in which protein representations are optimized to maximize the gap between binding and non-binding proteins. **(e)** Cell type context (provided by PINNACLE) is injected into context-free structure-based protein representations (provided by MaSIF [3], which learns a protein representation from the protein's 3D structure) via concatenation to generate contextualized protein representations. Lack of cell type context is defined by an average of PINNACLE's protein representations. **(f)** Comparison of context-free and contextualized representations in differentiating between binding and non-binding proteins. Scores are computed using cosine similarity on $n = 22$ unique protein pairs (2 binding and 20 non-binding); since PINNACLE generates multiple representations per protein based on context, there are n =7,956 pairwise computations (180 binding and 7,776 non-binding) for the contextualized representations. The binding proteins evaluated are PD-1/PD-L1 and B7-1/CTLA-4. Pairwise scores also are calculated for each of these four proteins and proteins that they do not bind with (i.e., RalB, RalBP1, EPO, EPOR, C3, and CFH). The gap between the average scores of binding and non-binding proteins is annotated for context-free and contextualized representations. The significance of the score gaps between binding and non-binding proteins is measured using a one-sided non-parametric permutation test. Data are represented as mean values with error bars indicating a 95% confidence interval.

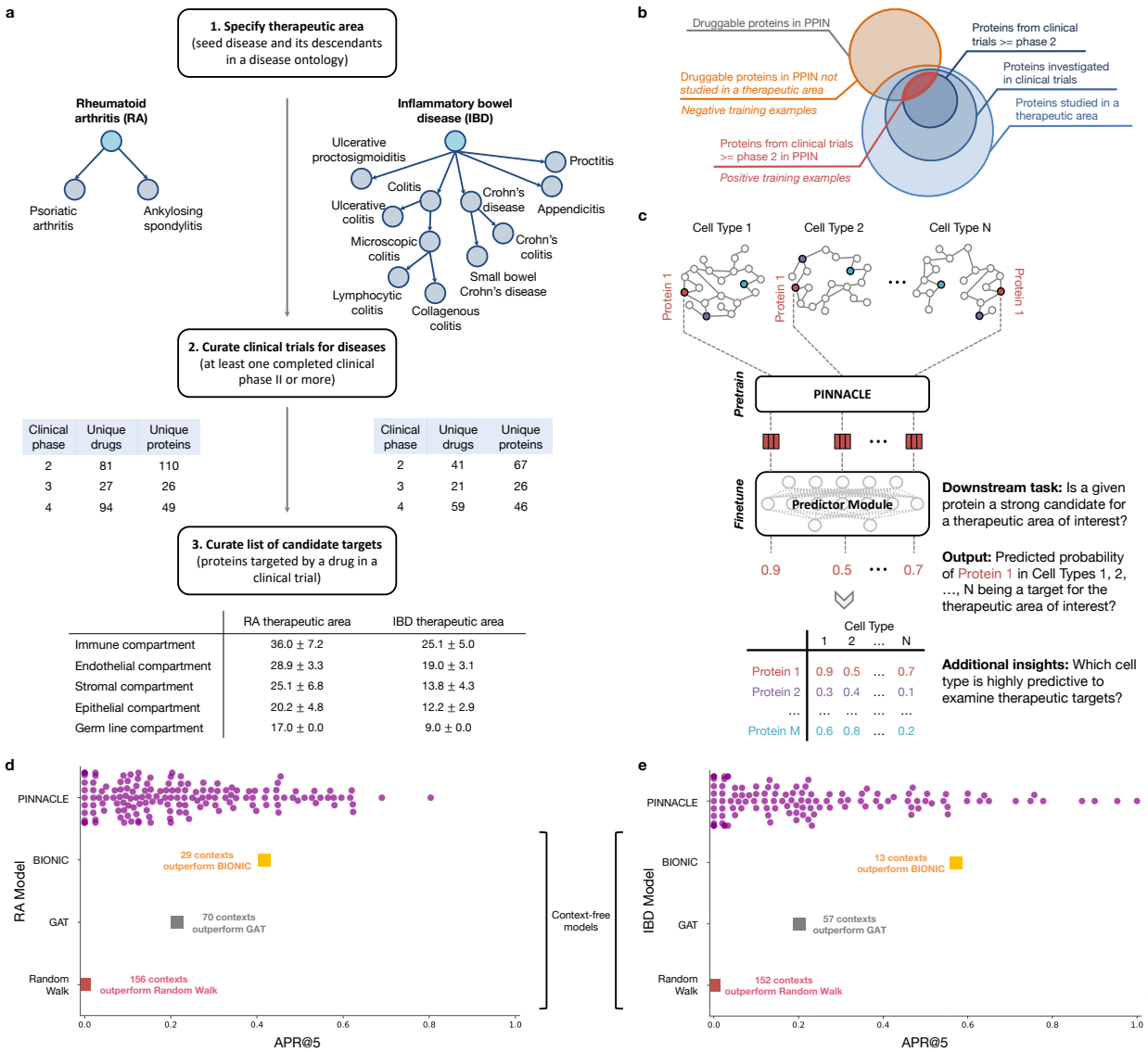

Figure 4: Fine-tuning contextualized protein representations for therapeutic target prioritization. **(a)** Workflow to curate positive training examples for rheumatoid arthritis (left) and inflammatory bowel disease (right) therapeutic areas. **(b)** We construct positive examples by selecting proteins from our protein-protein interaction network (PPIN) that are targeted by compounds that have at least completed phase 2 for treating the therapeutic area of interest. These proteins are deemed safe and potentially efficacious for humans with the disease. We construct negative examples by selecting proteins from our PPIN that do not have associations with the therapeutic area yet have been targeted by at least one existing drug/compound. **(c)** Cell type-specific protein interaction networks are embedded by PINNACLE, and finetuned for a downstream task. Here, the predictor module (i.e., multi-layer perceptron) finetunes the (pretrained) contextualized protein representations for predicting whether a given protein is a strong candidate for the therapeutic area of interest. Additional insights of our setup include hypothesizing highly predictive cell types for examining candidate therapeutic targets. **(d-e)** Benchmarking of context-aware and context-free approaches for **(d)** RA and **(e)** IBD therapeutic areas. Each dot is the performance (averaged across 10 random seeds) of protein representations from a given context (i.e., cell type context for PINNACLE, context-free global reference protein interaction network for GAT and random walk, and context-free multi-modal protein interaction network for BIONIC).

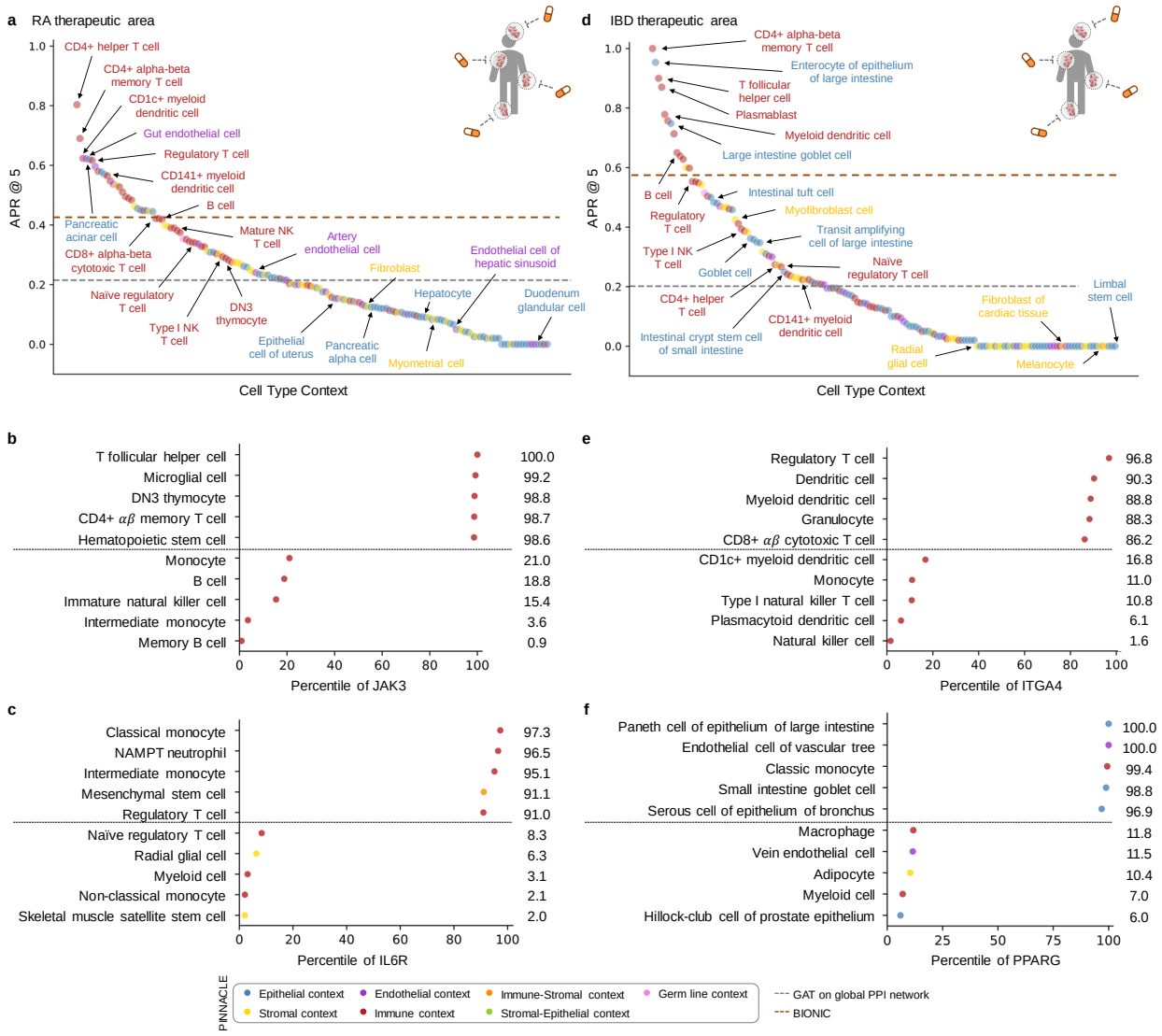

Figure 5: Performance of contextualized target prioritization for RA and IBD therapeutic areas. (a,d) Model performance (measured by APR@5) for RA and IBD therapeutic areas, respectively. APR@K (or Average Precision and Recall at K) is a combination of Precision@K and Recall@K (refer to Methods 6 for more details). Each dot is the performance (averaged across 10 random seeds) of PINNACLE's protein representations from a specific cell type context. The gray and dark orange lines are the performance of the GAT and BIONIC models, respectively. For each therapeutic area, 22 cell types are annotated and colored by their compartment category. Supplementary Figure S8 contains model performance measured by APR@10, APR@15, and APR@20 for RA and IBD therapeutic areas. **(b-c, e-f)** Selected proteins for RA and IBD therapeutic areas. Dotted line separates the top and bottom 5 cell types. **(b-c)** Two selected proteins, JAK3 and IL6R, that are targeted by drugs that have completed Phase IV of clinical trials for treating RA therapeutic area. **(e-f)** Two selected proteins, ITGA4 and PPARG, that are targeted by drugs that have completed Phase IV for treating IBD therapeutic area.

Online Methods

The Methods describe (1) the curation of datasets, (2) the construction and representation of multi-scale single-cell networks, (3) our multi-scale graph neural network, PINNACLE, (4) the finetuning of PINNACLE for target prioritization, and (5) the metrics and statistical analyses used.

1 Datasets

Reference human physical protein interaction network. Our global reference protein-protein interaction (PPI) network is the union of physical multi-validated interactions from BioGRID [101, 102], the Human Reference Interactome (HuRI) [91], and Menche *et al.* [103] with 15,461 nodes and 207,641 edges. Different sources of PPI have their own methods of curating and validating physical interactions between proteins. BioGRID, HuRI, and Menche *et al.* are PPI networks from three well-cited publications and databases regarding human protein interactions. By joining the three networks, we construct a comprehensive global PPI network for our analysis.

Multi-organ, single-cell transcriptomic atlas of humans. We leverage Tabula Sapiens [20] data source as our multi-organ, single-cell transcriptomic atlas of humans. The data consists of 15 donors, with 59 specimens total. There are 483,152 cells after quality control, of which 264,824 are immune cells, 104,148 are epithelial cells, 31,691 are endothelial cells, and 82,478 are stromal cells. The cells correspond to 177 unique cell ontology classes.

2 Construction of multi-scale networks

Our multi-scale networks comprises protein-protein physical interactions, cell type to cell type communication, cell type to tissue relationships, and tissue-tissue hierarchy.

Cell type-specific protein interaction networks. For each cell type, we create a cell type specific network that represents the physical interactions between proteins (or genes) that are likely expressed in the cell type. Intuitively, our approach identifies genes significantly expressed in a given cell type with respect to the rest of the cells in the dataset. Concretely, we use the processed Tabula Sapiens count matrix to calculate the average expression of each gene in a cell type of interest and the average expression of the corresponding gene in all other cells. Then, we use the Wilcoxon rank-sum test on the two sets of average gene expression. From the resulting ranked list of genes based on activation, we filter for the top K most activated genes. We repeat these two steps N times and filter for genes that appear in at least 90% of iterations. Finally, we extract these genes' corre-

sponding proteins from the global protein interaction network and take only the largest connected component. To ensure high-quality representations of cell types in our networks, we keep networks with at least 1,000 proteins. We do not perform subsampling of cells (i.e., sample the same number of cells per cell type) to minimize information loss for constructing protein interaction networks (Supplementary Figure S2). Limitations are described in the Discussion section.

Cell type and tissue relationships in the metagraph. We identify interactions between cell types based on ligand-receptor (LR) expression using the CellPhoneDB [104] tool and database. An edge between a pair of cell types indicates that CellphoneDB predicts at least one significantly expressed LR pair (with a p-value of less than 0.001) between them. As recommended by CellPhoneDB, cells are subsampled prior to running the algorithm, which uses geometric sketching [105] to efficiently sample a small representative subset of cells from massive datasets while preserving biological complexity. We choose to subsample 25% of cells and run CellPhoneDB for 100 iterations. We determine cell type-tissue relationships and extract tissue-tissue relationships using Tabula Sapiens meta-data. For relationships between cell types and tissues, we draw edges between cell types and the tissue that the cells were taken from. For tissue-tissue relationships, we select the nodes corresponding to the tissues where samples were taken from and all parent nodes up to the root of the BRENDA tissue ontology [106]. We perform sensitivity and ablation analyses on different components of the metagraph (Supplementary Table S3-S5).

Final dataset. We have 156 cell type specific protein interaction networks, which have, on average, $2,530 \pm 677$ proteins per network. The number of unique proteins across all cell type specific protein interaction networks is $13,643$ of the $15,461$ proteins in the global reference protein interaction network. In the metagraph, we have 62 tissues (nodes), and 24 are directly connected to cell types. There are 3,567 cell-cell interactions, 372 cell-tissue edges, and 79 tissue-tissue edges.

3 Multi-scale graph neural network

Overview. PINNACLE performs biologically-informed message passing through proteins, cell types, and tissues to learn cell type specific protein representations, cell type representations, and tissue representations in a unified multi-scale embedding space. Specifically, PINNACLE traverses through protein-protein physical interactions in each cell type specific PPI network, cell type-cell type communication, cell type-tissue relationships, and tissue-tissue hierarchy with an attention mechanism over individual nodes and edge types. Its objective function is designed and opti-

mized for learning the topology across biological scales, from proteins to cell types to tissues. The resulting embeddings from PINNACLE can be visualized and manipulated for hypothesis-driven interrogation and finetuned for diverse downstream biomedical prediction tasks.

Problem formulation. Let $\mathcal{G} = \{G_1, \ldots, G_{|\mathcal{C}|}\}$ be a set of cell type specific protein-protein interaction networks, where $\mathcal{C}$ is a set of unique cell types. Each $G_{c_i} = (V_{c_i}, E_{c_i})$ consists of a set of nodes V_{c_i} and edges E_{c_i} for a given cell type $c_i \in \mathcal{C}$ specific protein-protein interaction network. Their nodes $u, v \in V_{c_i}$ are proteins, and edges $e^{\text{PP}}_{u,v} \in E_{c_i}$ are physical protein-protein interactions (denoted with PP in the superscript). Cell types and tissues form a network, referred to as a metagraph. The metagraph's set of nodes comprises cell types $c_i \in \mathcal{C}$ and tissues $t_i \in \mathcal{T}$. The types of edges are cell type-cell type interactions (denoted with CC in the superscript) $e^{\text{CC}}_{c_i,c_j}$ between any pair of cell types $c_i, c_j \in \mathcal{C}$; cell type-tissue associations (denoted with CT in the superscript) $e^{\text{CT}}_{c_i,t_i}$ between any pair of cell type $c_i \in \mathcal{C}$ and tissue $t_i \in \mathcal{T}$; and tissue-tissue relationships (denoted with TT in the superscript) $e^{\text{TT}}_{t_i,t_j}$ between any pair of tissues $t_i, t_j \in \mathcal{T}$.

3.1 Protein-level attention with cell type specificity

For each cell type specific protein-protein interaction network $\mathcal{G}_{c_i}$, we leverage protein-level attention to learn cell type specific embeddings of proteins. Intuitively, protein-level attention learns which neighboring nodes are likely most important for characterizing a particular cell type's protein. As such, each cell type specific protein interaction network has its own cell type specific set of learnable parameters. Concretely, at each layer-wise update of layer l, the node-level attention learns the importance $\alpha_{u,v}$ of protein u to its neighboring protein v in a given cell type $c_i \in \mathcal{C}$:

$$\mathbf{h}^{\text{PP}}_u \leftarrow \text{AGG}\big(\sigma(\sum_{v \in \mathcal{N}_u} \alpha_{u,v} \mathbf{W}^{\text{PP}} \mathbf{h}^{\text{PP}}_v)\big) \tag{1}$$

where AGG is an aggregation function (i.e., concatenation across K attention heads), σ is the nonlinear activation function (i.e., ReLU), $\mathcal{N}_u$ is the set of neighbors for u (including itself via self-attention), $\alpha_{u,v}$ is an attention mechanism defined as $\alpha_{u,v} = \frac{\exp(\sigma(\mathbf{a}^T \cdot [\mathbf{h}_u \| \mathbf{h}_v]))}{\sum_{v \in \mathcal{N}_u} \exp(\sigma(\mathbf{a}^T \cdot [\mathbf{h}_u \| \mathbf{h}_v]))}$ between a pair of interacting proteins from a specific cell type, $\mathbf{W}^{\text{PP}}$ is a PP-specific transformation matrix to project the features of protein u in its cell type specific protein interaction network, and $\mathbf{h}^{\text{PP}}_v$ is the previous layer's cell type specific embedding for protein v. Practically, we leverage the attention function in graph attention neural networks (i.e., GATv2) [44]. Proteins of the same identity are initialized with the same random Gaussian vector to maintain their identity during training.

3.2 Metagraph-level attention on cellular interactions and tissue hierarchy

For the metagraph, we use node-level and edge-level attention to learn which neighboring nodes and edge types are likely most important for characterizing the target node (i.e., the node of interest). Intuitively, to learn an embedding for a specific cell type or tissue, we evaluate the informativeness of each direct cell type or tissue neighbor, as well as the relationship type between the cell type or tissue and their neighbors (e.g., parent-child tissue relationship, tissue from which a cell type is found, cell type with which the cell type of interest communicates with).

Concretely, at each layer l of PINNACLE, the embeddings of a cell type $c_i \in \mathcal{C}$ are the result of aggregating (via function AGG) the embeddings ($\mathbf{h}_c^{\text{CC}}$ and $\mathbf{h}_t^{\text{CT}}$) of its direct cell type neighbor c and tissue neighbor t that are projected via edge-type-specific transformation matrices ($\mathbf{W}^{\text{CC}}$ and $\mathbf{W}^{\text{CT}}$) and weighted by learned attention weights ($\alpha_{c_i,c}$ and $\alpha_{c_i,t}$ respectively):

$$\mathbf{h}_{c_i}^{\text{CC}} \leftarrow \text{AGG}\big(\sigma(\sum_{c \in \mathcal{N}_{c_i}} \alpha_{c_i,c} \mathbf{W}^{\text{CC}} \mathbf{h}_c^{\text{CC}})\big) \tag{2}$$

$$\mathbf{h}_{c_i}^{\text{CT}} \leftarrow \text{AGG}\big(\sigma(\sum_{t \in \mathcal{N}_{c_i}} \alpha_{c_i,t} \mathbf{W}^{\text{CT}} \mathbf{h}_t^{\text{CT}})\big) \tag{3}$$

The embeddings generated from separately propagating messages through cell type-cell type edges or cell type-tissue edges are combined using learned attention weights β^{CC} and β^{CT}, respectively:

$$\mathbf{h}_{c_i} = \beta^{\text{CC}} \mathbf{h}_{c_i}^{\text{CC}} + \beta^{\text{CT}} \mathbf{h}_{c_i}^{\text{CT}} \tag{4}$$

Similarly, the embeddings of a tissue $t_i \in \mathcal{T}$ are the result of aggregating (via function AGG) the embeddings ($\mathbf{h}_t^{\text{TT}}$ and $\mathbf{h}_c^{\text{TC}}$) of its direct tissue neighbor t and cell type neighbor c that are projected via edge-type-specific transformation matrices ($\mathbf{W}^{\text{TT}}$ and $\mathbf{W}^{\text{TC}}$) and weighted by learned attention weights ($\alpha_{t_i,t}$ and $\alpha_{t_i,c}$ respectively):

$$\mathbf{h}_{t_i}^{\text{TT}} \leftarrow \text{AGG}\big(\sigma(\sum_{t \in \mathcal{N}_{t_i}} \alpha_{t_i,t} \mathbf{W}^{\text{TT}} \mathbf{h}_t^{\text{TT}})\big) \tag{5}$$

$$\mathbf{h}_{t_i}^{\text{TC}} \leftarrow \text{AGG}\big(\sigma(\sum_{c \in \mathcal{N}_{t_i}} \alpha_{t_i,c} \mathbf{W}^{\text{TC}} \mathbf{h}_c^{\text{TC}})\big) \tag{6}$$

The embeddings generated from separately propagating messages through tissue-tissue edges or

tissue-cell type edges are combined using learned attention weights β^{TT} and β^{TC}, respectively:

$$\mathbf{h}_{t_i} = \beta^{\text{TT}}\mathbf{h}_{t_i}^{\text{TT}} + \beta^{\text{TC}}\mathbf{h}_{t_i}^{\text{TC}} \tag{7}$$

For the node-level attention mechanisms (Equations 2, 3, 5, and 6), AGG is an aggregation function (i.e., concatenation across K attention heads), σ is the nonlinear activation function (i.e., ReLU), $\mathcal{N}_{c_i}$ and $\mathcal{N}_{t_i}$ are the sets of neighbors for c_i and t_i respectively (includes itself via self-attention), $\mathbf{W}^{\text{CC}}$, $\mathbf{W}^{\text{CT}}$, $\mathbf{W}^{\text{TC}}$, and $\mathbf{W}^{\text{TT}}$ are edge-type-specific transformation matrices to project the features of a given target node, $\mathbf{h}_c^{\text{CC}}$, $\mathbf{h}_t^{\text{CT}}$, $\mathbf{h}_t^{\text{TT}}$, and $\mathbf{h}_c^{\text{TC}}$ are the previous layer's embedding for c given the edge type CC, t given the edge type CT, t given the edge type TT, and c given the edge type TC, respectively. Practically, we leverage the attention function in graph attention neural networks (i.e., GATv2) [44]. Finally, the node-level attention mechanism for a given source node u and edge type r is $\alpha_{u,v}^r = \frac{\exp(\sigma(\mathbf{a}_r^T \cdot [\mathbf{h}_u \| \mathbf{h}_v]))}{\sum_{v \in \mathcal{N}_u} \exp(\sigma(\mathbf{a}_r^T \cdot [\mathbf{h}_u \| \mathbf{h}_v]))}$. For the attention mechanisms over edge types (Equations 4 and 7), $\beta^r = \frac{\exp(m_r)}{\sum_{r \in R} \exp(m_r)}$ such that $m_r = \sum_{u \in V_q} \mathbf{s}^T \cdot \tanh(\mathbf{M} \cdot \mathbf{h}_u^r + \mathbf{b})$ where V_q is the set of nodes in the metagraph, $\mathbf{s}$ is the attention vector, $\mathbf{M}$ is the weight matrix, and $\mathbf{b}$ is the bias vector. These parameters are shared for all edge types in the metagraph.

3.3 Bridge between protein and cell type embeddings

Using a pooling mechanism, we bridge cell type specific protein embeddings with their corresponding cell type embeddings. We initialize cell type embeddings by taking the average of their proteins' embeddings: $\mathbf{h}_{c_i} = \frac{1}{|V_{c_i}|} \sum_{u \in V_{c_i}} \mathbf{h}_u$ where $\mathbf{h}_u$ is the embedding of protein node $u \in V_{c_i}$ in the PPI subnetwork for cell type c_i. Similarly, we initialize tissue embeddings by taking the average of their neighbors: $\mathbf{h}_{t_i} = \frac{1}{|\mathcal{N}_{t_i}|}\big(\sum_{t \in \mathcal{N}_{t_i}} \mathbf{h}_t + \sum_{c \in \mathcal{N}_{t_i}} \mathbf{h}_c\big)$, where h_t and h_c are the embeddings of tissue node t and cell type node c, respectively, in the immediate neighborhood of source tissue node t_i. At each layer $l > 0$, we learn the importance $\gamma_{c_i,u}$ of node $u \in V_{c_i}$ to cell type c_i such that

$$\mathbf{h}_{c_i} \leftarrow \mathbf{h}_{c_i} + \text{AGG}\big(\sigma(\sum_{u \in V_{c_i}} \gamma_{c_i,u}\mathbf{h}_u)\big) \tag{8}$$

After propagating cell type and tissue information in the metagraph (namely, Equations 2-6), we apply $\gamma_{c_i,u}$ to the cell type embedding of c_i such that

$$\mathbf{h}_u \leftarrow \mathbf{h}_u + \gamma_{c_i,u}\mathbf{h}_{c_i} \tag{9}$$

Intuitively, we are imposing the structure of the metagraph onto the PPI subnetworks based on a protein's importance to its corresponding cell type's identity.

3.4 PINNACLE: Overall objective function

PINNACLE is optimized for three biological scales: protein-, cell type-, and tissue-level. Concretely, the loss function $\mathcal{L}$ has three components corresponding to each biological scale:

$$\mathcal{L} = \mathcal{L}_{protein} + (1-\theta)(\mathcal{L}_{celltype} + \mathcal{L}_{tissue}) \tag{10}$$

where $\mathcal{L}_{protein}$, $\mathcal{L}_{celltype}$, and $\mathcal{L}_{tissue}$ minimize the loss from protein-level predictions, cell type-level predictions, and tissue-level predictions, respectively. θ is a tunable parameter with a range of 0 and 1 that scales the contribution of the link prediction loss of the metagraph relative to that of the protein-protein interactions. At the protein level, we consider two aspects: prediction of protein-protein interactions at each cell type specific PPI network ($\mathcal{L}_{ppi}$) and prediction of cell type identity of each protein ($\mathcal{L}_{celltypeid}$). The contribution of the latter is scaled by λ, which is a tunable parameter with a range of 0 and 1:

$$\mathcal{L}_{protein} = \theta\mathcal{L}_{ppi} + \lambda\mathcal{L}_{celltypeid} \tag{11}$$

Intuitively, we aim to simultaneously learn the topology of each cell type specific PPI network (i.e., $\mathcal{L}_{ppi}$) and the nuanced differences between proteins activated in different cell types. Specifically, we use binary cross-entropy to minimize the error of predicting positive and negative protein-protein interactions in each cell type specific PPI network:

$$\mathcal{L}_{ppi} = \sum_{c_i \in \mathcal{C}} \sum_{u,v \in V_{c_i}} y_{u,v} \log(\hat{y}_{u,v}) + (1 - y_{u,v}) \log(1 - \hat{y}_{u,v}) \tag{12}$$

and center loss [107] for discriminating between protein embeddings $\mathbf{z}_u$ from different cell types, represented by embeddings denoted as $\mathbf{z}_{c_i}$:

$$\mathcal{L}_{celltypeid} = \sum_{c_i \in \mathcal{C}} \sum_{u \in V_{c_i}} ||\mathbf{z}_u - \mathbf{z}_{c_i}||_2^2 \tag{13}$$

At the cell type level, we use binary cross-entropy to minimize the error of predicting cell type-cell type interactions and cell type-tissue relationships:

$$\mathcal{L}_{celltype} = \mathcal{L}^{\text{CC}}_{celltype} + \mathcal{L}^{\text{CT}}_{celltype} \tag{14}$$

such that

$$\mathcal{L}^{\text{CC}}_{celltype} = \sum_{c_i, c_j \in \mathcal{C}} y_{c_i,c_j} \log(\hat{y}_{c_i,c_j}) + (1 - y_{c_i,c_j}) \log(1 - \hat{y}_{c_i,c_j}) \tag{15}$$

$$\mathcal{L}^{\text{CT}}_{celltype} = \sum_{c_i \in \mathcal{C}} \sum_{t_k \in \mathcal{T}} y_{c_i,t_k} \log(\hat{y}_{c_i,t_k}) + (1 - y_{c_i,t_k}) \log(1 - \hat{y}_{c_i,t_k}) \tag{16}$$

Similarly, at the tissue level, we use binary cross-entropy to minimize the error of predicting tissue-tissue and tissue-cell type relationships:

$$\mathcal{L}_{tissue} = \mathcal{L}^{\text{TT}}_{tissue} + \mathcal{L}^{\text{TC}}_{tissue} \tag{17}$$

such that

$$\mathcal{L}^{\text{TT}}_{tissue} = \sum_{t_k, t_q \in \mathcal{T}} y_{t_k,t_q} \log(\hat{y}_{t_k,t_q}) + (1 - y_{t_k,t_q}) \log(1 - \hat{y}_{t_k,t_q}) \tag{18}$$

$$\mathcal{L}^{\text{TC}}_{tissue} = \sum_{t_k \in \mathcal{T}} \sum_{c_i \in \mathcal{C}} y_{t_k,c_i} \log(\hat{y}_{t_k,c_i}) + (1 - y_{t_k,c_i}) \log(1 - \hat{y}_{t_k,c_i}) \tag{19}$$

The probability of an edge of type i between nodes u and v is calculated using a bilinear decoder:

$$y_{u,v} = \mathbf{z}_u \cdot \mathbf{r}_i \cdot \mathbf{z}_v \tag{20}$$

where $\mathbf{z}_u$ and $\mathbf{z}_v$ are embeddings of nodes u and v, and $\mathbf{r}_i$ is the embedding for edge type i. Note that any decoder can be used for link prediction in PINNACLE.

3.5 Training details for PINNACLE

Overview. PINNACLE is trained using the cell type identity of the protein interaction networks and the graph connectivity of the cell type specific protein interaction networks and metagraph. To learn cell type identity, PINNACLE predicts the cell type(s) that the node(s) corresponding to each

protein are activated in. For capturing graph connectivity, PINNACLE performs self-supervised link prediction; it predicts whether an edge (and its type) exists between a pair of nodes. For link prediction, a randomly selected subset of edges is masked (or hidden) from the model, and the model must be able to predict that such edges exist (and that the randomly generated false edges do not exist). Practically, this means that the graphs being fed as input into PINNACLE during train, validation, or test do not contain the masked edges.

Data split. Protein-protein edges are randomly split into train (80%), validation (10%), and test (10%) sets. The metagraph edges are not split into train, validation, and test sets because there are relatively few of them, and they are all critical for injecting cell type and tissue organization to the model. The proteins involved in the train edges are considered in the cell type identification term of the loss function ($\mathcal{L}_{celltypeid}$).

Sampling negative edges. For link prediction, false (or negative) edges have the label of 0 and are randomly generated (via `structured_negative_sampling` function in Pytorch Geometric). The ratio of positive to negative edges is 1:1.

Hyperparameter tuning. We leverage Weights and Biases [108] to select optimal hyperparameters via a random search over the hyperparameter space. The best-performing hyperparameters for PINNACLE are selected by optimizing the ROC and Calinski-Harabasz score [109] on the validation set. The hyperparameter space on which we perform a random search to choose the optimal set of hyperparameters is: the dimension of the nodes' feature matrix $\in [1024, 2048]$, dimension of the output layer $\in [4, 8, 16, 32]$, lambda $\in [0.1, 0.01, 0.001]$, learning rate for link prediction task $\in [0.01, 0.001]$, learning rate for protein's cell type classification task $\in [0.1, 0.01, 0.001]$, number of attention heads $\in [4, 8]$, weight decay rate $\in [0.0001, 0.00001]$, dropout rate $\in [0.3, 0.4, 0.5, 0.6, 0.7]$, and normalization layer $\in [\texttt{layernorm}, \texttt{batchnorm}, \texttt{graphnorm}, \texttt{none}]$. The best hyperparameters are as follows: the dimension of the nodes' feature matrix $= 1024$, dimension of the output layer $= 16$, lambda $= 0.1$, learning rate for link prediction task $= 0.01$, learning rate for protein's cell type classification task $= 0.1$, number of attention heads $= 8$, weight decay rate $= 0.00001$, dropout rate $= 0.6$, and normalization layers are `layernorm` and `batchnorm`. Further, PINNACLE consists of two custom graph attention neural network layers (as specified in Section 3) per cell type specific PPI network and metagraph, and is trained for 250 epochs.

Implementation. We implement PINNACLE using Pytorch (Version 1.12.1) [110] and Pytorch Geometric (Version 2.1.0) [111]. We leverage Weights and Biases [108] for hyperparameter tuning

and model training visualization, and we create interactive demos of the model using Gradio [112]. Models are trained on a single NVIDIA Tesla V100-SXM2-16GB GPU.

4 Generating contextualized 3D protein representations

After pre-training PINNACLE, we can leverage the output protein representations for diverse downstream tasks. Here, we demonstrate PINNACLE's ability to improve the prediction of protein-protein interactions by injecting context into 3D molecular structures of proteins.

Overview. Given a protein of interest, we generate both the context-free structure-based representation via MaSIF [3, 17] and a contextualized PPI network-based representation via PINNACLE. We calculate the binding score of a pair of proteins based on either context-free representations or contextualized representations of the proteins. To quantify the added value, if any, provided by contextualizing protein representations with cell type context, we compare the size of the gap between the average binding scores of binding and non-binding proteins in the two approaches.

Dataset. The proteins being compared are PD-1, PD-L1, B7-1, CTLA-4, RalB, RalBP1, EPO, EPOR, C3, and CFH. The pairs of binding proteins are PD-1/PD-L1 (PDB ID: 4ZQK) and B7-1/CTLA-4 (PDB ID: 1I8L). The non-binding proteins are any of the four proteins paired with any of the remaining six proteins (e.g., PD-1/RalB, PD-1/RalBP1, PD-L1/RalBP1). The PDB IDs for the other six proteins are 2KWI for RalB/RalBP1, 1CN4 for EPO/EPOR, and 3OXU for C3/CFH.

Structure-based protein representation learning model. We directly apply the pretrained model for MaSIF [3, 17] to generate the 3D structure-based protein representations. We use the model pretrained for MaSIF-site task, named `all_feat_3l_seed_benchmark`. The output of the pretrained model for a given protein is $P \times d$, where P is the number of patches (precomputed by the authors of MaSIF [3, 17]) and $d = 4$ is the dimension of the pretrained model's output layer. As proteins vary in size (i.e., the number of patches to cover the surface of the protein), we select a fixed k number of patches that are most likely to be part of the binding site (according to the pretrained MaSIF model). For each protein, we select $k = 200$ patches, which is the average number of patches for PD-1, PD-L1, B7-1, and CTLA-4, resulting in a matrix of size 200×4. Finally, we take the element-wise median on the 200×4 matrix to transform it into a vector of length 200. This vector becomes the structure-based protein representation for a given protein.

Experimental setup. For each cell type context of a given protein, we concatenate the 3D structure-based protein representation (from MaSIF) with the cell type specific protein represen-

tation (from PINNACLE) to generate a contextualized structure-based protein representation. To create the context-free protein representation, we concatenate the structure-based protein representation with an element-wise average of PINNACLE's protein representations. This is to maintain consistent dimensionality and latent space between context-free and contextualized protein representations. Given a pair of proteins, we calculate a score via cosine similarity (a function provided by `sklearn` [113]) using the context-free or contextualized protein representations. Lastly, we quantify the gap between the scores of binding and non-binding proteins using context-free or contextualized protein representations to evaluate the added value (if any) of contextual AI.

5 Fine-tuning PINNACLE for target prioritization

After pre-training PINNACLE, we can fine-tune the output protein representations for diverse biomedical downstream tasks. Here, we demonstrate PINNACLE's ability to enhance the performance of predicting a protein's therapeutic potential for a specific therapeutic area.

Overview. For each protein of interest, we feed its PINNACLE-generated embedding into a multi-layer perceptron (MLP). The model outputs a score between 0 and 1, where 1 indicates strong candidacy to target (i.e., by a compound/drug) for treating the therapeutic area and 0 otherwise. Since a protein has multiple representations corresponding to the cell types it is activated in, the MLP model generates a score for each of the protein's cell type-specific representations (Figure 4a). For example, Protein 1's representation from Cell type 1 is scored independently of its representation from Cell type 2. The output scores can be examined to identify the most predictive cell types and the strongest candidates for therapeutic targets in any specific cell type.

5.1 Therapeutic targets dataset

We obtain labels for therapeutic targets from the Open Targets Platform [61].

Therapeutic area selection. To curate target information for a therapeutic area, we examine the drugs indicated for the therapeutic area of interest and its descendants. The two therapeutic areas examined are rheumatoid arthritis (RA) and inflammatory bowel disease. For rheumatoid arthritis, we collected therapeutic data (i.e., targets of drugs indicated for the therapeutic area) from Open-Targets [61] for rheumatoid arthritis (EFO_0000685), ankylosing spondylitis (EFO_0003898), and psoriatic arthritis (EFO_0003778). For inflammatory bowel disease, we collected therapeutic data for ulcerative colitis (EFO_0000729), collagenous colitis (EFO_1001293), colitis (EFO_0003872), proctitis (EFO_0005628), Crohn's colitis (EFO_0005622), lymphocytic colitis (EFO_1001294),

Crohn's disease (EFO_0000384), microscopic colitis (EFO_1001295), inflammatory bowel disease (EFO_0003767), appendicitis (EFO_0007149), ulcerative proctosigmoiditis (EFO_1001223), and small bowel Crohn's disease (EFO_0005629).

Positive training examples. We define positive examples (i.e., where the label $y = 1$) as proteins targeted by drugs that have at least completed phase 2 of clinical trials for treating a certain therapeutic area. As such, a protein is a promising candidate if a compound that targets the protein is safe for humans and effective for treating the disease. We retain positive training examples that are activated in at least one cell type specific protein interaction network. The final number of positive training examples for RA and IBD are 152 and 114, respectively.

Negative training examples. We define negative examples (i.e., where the label $y = 0$) as druggable proteins that do not have any known association with the therapeutic area of interest according to OpenTargets. A protein is deemed druggable if it is targeted by at least one existing drug [114]. We extract drugs and their nominal targets from DrugBank [79]. We retain negative training examples that are activated in at least one cell type specific protein interaction network. The final number of negative training examples for RA and IBD are 1,465 and 1,377, respectively.

Data processing workflow. For a therapeutic area of interest, we identify its descendants. With the list of disease terms for the therapeutic area, we curate its positive and negative training examples. We split the dataset such that about 60%, 20%, and 20% of the proteins are in the train, validation, and test sets, respectively. We additionally apply two criteria to avoid data leakage and ensure that all cell types are represented during training/inference: Proteins are assigned to train (60%), validation (20%), and test (20%) datasets based on their identity; this is to prevent data leakage where cell type specific representations of a single protein are observed in multiple data splits. We also ensure that there are sufficient numbers of train, validation, and test positive samples per cell type; proteins may be reassigned to a different data split so that each cell type is represented during training, validating, and testing stages. With these criteria, the train, validation, and test dataset splits may not necessarily consist of approximately 60%, 20%, and 20% of the total protein representations (Supplementary Table S6).

5.2 Finetuning model details

Model architecture. Our multi-layer perceptron (MLP) comprises an input feedforward neural network, one hidden feedforward neural network layer, and an output feedforward neural network

layer. In between each layer, we have a non-linear activation layer. In addition, we use dropout and normalization layers between the input and hidden layer (see the Implementation section for more information). Our objective function is binary cross-entropy loss.

Hyperparameter tuning. We leverage Weights and Biases [108] to select optimal hyperparameters via a random search over the hyperparameter space. The best-performing hyperparameters are selected by optimizing the AUPRC on the validation set. The hyperparameter space on which we perform a random search to choose the optimal set of hyperparameters is the dimension of the first hidden layer $\in [8, 16, 32]$, dimension of the second hidden layer $\in [8, 16, 32]$, learning rate $\in [0.01, 0.001, 0.0001]$, weight decay rate $\in [0.001, 0.0001, 0.00001, 0.000001]$, dropout rate $\in [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$, normalization layer $\in [\texttt{layernorm}, \texttt{batchnorm}, \texttt{none}]$, and the ordering of dropout and normalization layer (i.e., normalization before dropout, or vice versa).

Implementation. We implement the MLP using Pytorch (Version 1.12.1) [110]. In addition, we use Weights and Biases [108] for hyperparameter tuning and model training visualization. Models are trained on a single NVIDIA Tesla M40 GPU.

6 Metrics and statistical analyses

Here, we describe metrics, visualization methods, and statistical tests used in our analysis.

6.1 Visualization of embeddings

We visualize PINNACLE's embeddings using a uniform manifold approximation and projection for dimension reduction (UMAP) [115] and `seaborn`. Using the Python package, `umap`, we transform PINNACLE's embeddings to two-dimensional vectors via the parameters: $\texttt{n_neighbors} = 10$, $\texttt{min_dist} = 0.9$, $\texttt{n_components} = 2$, and the euclidean distance metric. The plots are created using the `seaborn` package's `scatterplot` function.

6.2 Visualization of cell type embedding similarity

The pairwise similarity of PINNACLE's cell type embeddings is calculated using cosine similarity (a function provided by `sklearn` [113]). Then, these similarity scores are visualized using the `seaborn` package's `clustermap` function. For visualization purposes, similarity scores are mapped to colors after being raised to the 20th power.

6.3 Spatial enrichment analysis of PINNACLE's protein embeddings

To quantify the spatial enrichment for PINNACLE's protein embedding regions, we apply a systematic approach, SAFE [31], that identifies regions that are over-represented for a feature of interest (Supplementary Figures S3-S4). The required input data for SAFE are networks and annotations of each node. We first construct an unweighted similarity network on PINNACLE protein embeddings: (1) calculate pairwise cosine similarity, (2) apply a similarity threshold on the similarity matrix to generate an adjacency matrix, and (3) extract the largest connected component. The protein nodes are labeled as 1 if they belong to a given cell type context and 0 otherwise. We then apply SAFE to each network using the recommended settings: neighborhoods are defined using the `shortpath_weighted_layout` metric for node distance and neighborhood radius of 0.15, and p-values are computed using the hypergeometric test, adjusted using the Benjamin-Hochberg false discovery rate correction (significance cutoff $\alpha = 0.05$).

Due to computation and memory constraints, we sample 50 protein embeddings from a cell type context of interest and 10 protein embeddings from each of the other 155 cell type contexts. We use a threshold of 0.3 in our evaluation of PINNACLE's protein embedding regions (Figure 2; Supplementary Figures S3). We also evaluate the spatial enrichment analysis on networks constructed from different thresholds to ensure that the enrichment is not sensitive to our network construction method: $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ (Supplementary Figures S4). We use the Python implementation of SAFE: https://github.com/baryshnikova-lab/safepy.

6.4 Statistical significance of tissue embedding distance

Tissue embedding distance between a given pair of tissue nodes is calculated using cosine distance (a function provided by `sklearn` [113]). Tissue ontology distance between a given pair of tissue nodes is calculated by taking the sum of the nodes' shortest path lengths to the lowest common ancestor (functions provided by `networkx` [116]. We use the two-sample Kolmogorov-Smirnov test (a function provided by `scipy` [117]) to compare PINNACLE embedding distances against randomly generated vectors (via the `randn` function in `numpy` to sample an equal number of vectors from a standard normal distribution). We also use the Spearman correlation (a function provided by `scipy` [117]) to correlate PINNACLE embedding distance to tissue ontology distance. We additionally generate a null distribution of tissue ontology distance by calculating tissue ontology distance on a shuffled tissue hierarchy (repeated 10 times). Concretely, we shuffle the node identities of the Brenda Tissue Ontology [106] and compute the pairwise tissue ontology distances.

6.5 Statistical significance of binding and non-binding proteins' score gaps

We perform a one-sided non-parametric permutation test. First, we concatenate the scores for the N binding pairs and M non-binding pairs. Next, for 100,000 iterations, we randomly sample N scores as the new set of binding protein scores and M scores as the new set of non-binding protein scores, calculate the mean μ_N of the N binding protein scores and the mean μ_M of the M non-binding protein scores, calculate the score gap by taking the difference of the means as $\mu_N - \mu_M$, and keep track of the score gaps that are greater than or equal to the true score gap calculated from the real data. Lastly, we calculate the p-value, defined as the fraction of 100,000 iterations in which the permuted score gap is greater than or equal to the true score gap (i.e., one-sided non-parametric permutation test).

6.6 Performance metric for therapeutic target prioritization

For our downstream therapeutic target prioritization task (Methods 5), we use a metric called Average Precision and Recall at K (APR@K) to evaluate model performance. APR@K leverages a combination of Precision@K and Recall@K to measure the ability to rank the most relevant items (in our case, proteins) among the top K predictions. In essence, APR@K calculates Precision@K for each $k \in [1, \ldots, K]$, multiplying each Precision@k by whether the kth item is relevant, and divides by the total number of relevant items r at K:

$$\text{APR@}K = \frac{1}{r} \sum_{k=1}^{K} \text{Precision@}k \cdot \text{rel}(k)$$

where

$$rel(k) = \begin{cases} 1, & \text{if item at } k \text{ is relevant} \\ 0, & \text{otherwise} \end{cases}$$

Given the nature of our target prioritization task, some key advantages of using APR@K include robustness to (1) varied numbers of protein targets activated across cell type-specific protein interaction networks and (2) varied sizes of cell type specific protein interaction networks.

Data availability. All data used in the paper, including the cell type specific protein interaction networks, the metagraph of cell type and tissue relationships, PINNACLE's contextualized representations, the therapeutic targets of rheumatoid arthritis and inflammatory bowel diseases, and the final and intermediate results of the analyses, are shared via the project website at https://zitniklab.hms.harvard.edu/projects/PINNACLE. Datasets are available via Figshare at https://doi.org/10.6084/m9.figshare.22708126.

Code availability. Python implementation of the methodology developed and used in the study is available via the project website at https://zitniklab.hms.harvard.edu/projects/PINNACLE. The code to reproduce results, together with documentation and examples of usage, are available on GitHub at https://github.com/mims-harvard/PINNACLE. We provide an interactive demo via HuggingFace to explore PINNACLE's contextualized protein representations.

References

1. Lund-Johansen, F., Tran, T. & Mehta, A. Towards reproducibility in large-scale analysis of protein–protein interactions. *Nature Methods* **18,** 720–721 (2021).
2. Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods* **19,** 774–779 (2022).
3. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **17,** 184–192 (2019).
4. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12,** 56–68 (2010).
5. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377,** 387–394 (2022).
6. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences* **118,** e2025581118 (2021).
7. Loscalzo, J. Molecular interaction networks and drug development: Novel approach to drug target identification and drug repositioning. *The FASEB Journal* **37** (2023).
8. Somnath, V. R., Bunne, C. & Krause, A. *Multi-Scale Representation Learning on Proteins* in *Advances in Neural Information Processing Systems* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W.) **34** (Curran Associates, Inc., 2021), 25244–25255.
9. Aykent, S. & Xia, T. *GBPNet: Universal Geometric Representation Learning on Protein Structures* in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ACM, Washington DC USA, 2022), 4–14.
10. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. en. *Proceedings of the National Academy of Sciences* **118,** e2016239118 (2021).
11. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics* **47,** 569–576 (2015).
12. Zitnik, M. & Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* **33,** i190–i198 (2017).
13. Ziv, M., Gruber, G., Sharon, M., Vinogradov, E. & Yeger-Lotem, E. The TissueNet v.3 Database: Protein-protein Interactions in Adult and Embryonic Human Tissue contexts. *Journal of Molecular Biology* **434,** 167532 (2022).
14. Mohammadi, S., Davila-Velderrain, J. & Kellis, M. Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution. *Cell Systems* **9,** 559–568.e4 (2019).
15. Forster, D. T. *et al.* BIONIC: biological network integration using convolutions. *Nature Methods* **19,** 1250–1261 (2022).
16. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R. & Jaakkola, T. *Equibind: Geometric deep learning for drug binding structure prediction* in *International Conference on Machine Learning* (2022).
17. Gainza, P. *et al.* De novo design of protein interactions with learned surface fingerprints. *Nature* **617,** 176–184 (2023).
18. Ittisoponpisan, S., Alhuzimi, E., Sternberg, M. J. E. & David, A. Landscape of Pleiotropic Proteins Causing Human Disease: Structural and System Biology Insights: HUMAN MUTATION. *Human Mutation* **38,** 289–296 (2017).

19. Pan, J. *et al.* Sparse dictionary learning recovers pleiotropy from human cell fitness screens. *Cell Systems* **13,** 286–303.e10 (2022).

20. Tabula Sapiens Consortium *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376,** eabl4896 (2022).

21. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Science Advances* **6,** eaba2083 (2020).

22. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nature Reviews Genetics* **21,** 137–150 (2020).

23. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* **19,** 41–50 (2022).

24. Lewis, M. J. *et al.* Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes. *Cell Reports* **28,** 2455–2470.e5 (2019).

25. Zhang, F. *et al.* Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature* **623,** 616–624 (2023).

26. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178,** 714–730.e22 (2019).

27. Kong, L. *et al.* The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity* **56,** 2855 (2023).

28. Vaswani, A. *et al. Attention is All You Need* in *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017).

29. Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M. & Zitnik, M. Multimodal learning with graphs. *Nature Machine Intelligence* **5,** 340–350 (2023).

30. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618,** 616–624 (2023).

31. Baryshnikova, A. Systematic Functional Annotation and Visualization of Biological Networks. *Cell Systems* **2,** 412–421 (2016).

32. Halakou, F., Kilic, E. S., Cukuroglu, E., Keskin, O. & Gursoy, A. Enriching Traditional Protein-protein Interaction Networks with Alternative Conformations of Proteins. *Scientific Reports* **7,** 7180 (2017).

33. Chakrabarti, K. S. *et al.* Conformational Selection in a Protein-Protein Interaction Revealed by Dynamic Pathway Analysis. *Cell Reports* **14,** 32–42 (2016).

34. Federico, A. & Monti, S. Contextualized Protein-Protein Interactions. *Patterns* **2,** 100153 (2021).

35. Braberg, H., Echeverria, I., Kaake, R. M., Sali, A. & Krogan, N. J. From systems to structure — using genetic data to model protein structures. *Nature Reviews Genetics* **23,** 342–354 (2022).

36. Robert, C. A decade of immune-checkpoint inhibitors in cancer therapy. *Nature Communications* **11,** 3801 (2020).

37. Yap, H.-Y. *et al.* Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. *Cells* **7,** 161 (2018).

38. Chang, M. H. *et al.* Arthritis flares mediated by tissue-resident memory T cells in the joint. *Cell Reports* **37** (2021).

39. Vickovic, S. *et al.* Three-dimensional spatial transcriptomics uncovers cell type localizations in the human rheumatoid arthritis synovium. *Communications Biology* **5,** 129 (2022).

40. Chang, J. T. Pathophysiology of Inflammatory Bowel Diseases. *New England Journal of Medicine* **383** (ed Longo, D. L.) 2652–2664 (2020).

41. Abbasi, M. *et al.* Strategies toward rheumatoid arthritis therapy; the old and the new. *Journal of Cellular Physiology* **234,** 10018–10031 (2018).

42. Orange, D. E. *et al.* RNA Identification of PRIME Cells Predicting Rheumatoid Arthritis Flares. *New England Journal of Medicine* **383,** 218–228 (2020).

43. Agrawal, M., Zitnik, M. & Leskovec, J. Large-scale analysis of disease pathways in the human interactome. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **23,** 111–122 (2018).

44. Brody, S., Alon, U. & Yahav, E. How attentive are graph attention networks? *ICLR* (2022).

45. Evans, C. H. *et al.* Gene transfer to human joints: Progress toward a gene therapy of arthritis. *Proceedings of the National Academy of Sciences* **102,** 8698–8703 (2005).

46. Marel, S. V. D. Gene and cell therapy based treatment strategies for inflammatory bowel diseases. *World Journal of Gastrointestinal Pathophysiology* **2,** 114 (2011).

47. Parker, K. R. *et al.* Single-Cell Analyses Identify Brain Mural Cells Expressing CD19 as Potential Off-Tumor Targets for CAR-T Immunotherapies. *Cell* **183,** 126–142.e17 (2020).

48. Verma, P., Srivastava, A., Srikanth, C. V. & Bajaj, A. Nanoparticle-mediated gene therapy strategies for mitigating inflammatory bowel disease. *Biomaterials Science* **9,** 1481–1502 (2021).

49. Zhang, Q. *et al.* Novel gene therapy for rheumatoid arthritis with single local injection: adeno-associated virus-mediated delivery of A20/TNFAIP3. *Military Medical Research* **9,** 34 (2022).

50. Craig, E. & Cappelli, L. C. Gastrointestinal and Hepatic Disease in Rheumatoid Arthritis. *Rheumatic Disease Clinics of North America* **44,** 89–111 (2018).

51. Dios, I. D. Inflammatory role of the acinar cells during acute pancreatitis. *World Journal of Gastrointestinal Pharmacology and Therapeutics* **1,** 15 (2010).

52. Yang, X., Chang, Y. & Wei, W. Endothelial Dysfunction and Inflammation: Immunity in Rheumatoid Arthritis. *Mediators of Inflammation* **2016,** 1–9 (2016).

53. Parker, B. & Chattopadhyay, C. A case of rheumatoid vasculitis involving the gastrointestinal tract in early disease. *Rheumatology* **46,** 1737–1738 (2007).

54. Roda, G. Intestinal epithelial cells in inflammatory bowel diseases. *World Journal of Gastroenterology* **16,** 4264 (2010).

55. Uzzan, M. *et al.* Ulcerative colitis is characterized by a plasmablast-skewed humoral response associated with disease activity. *Nature Medicine* **28,** 766–779 (2022).

56. Imam, T., Park, S., Kaplan, M. H. & Olson, M. R. Effector T Helper Cell Subsets in Inflammatory Bowel Diseases. *Frontiers in Immunology* **9,** 1212 (2018).

57. Casalegno Garduño, R. & Däbritz, J. New Insights on CD8+ T Cells in Inflammatory Bowel Disease and Therapeutic Approaches. *Frontiers in Immunology* **12,** 738762 (2021).

58. Tindemans, I., Joosse, M. E. & Samsom, J. N. Dissecting the Heterogeneity in T-Cell Mediated Inflammation in IBD. *Cells* **9,** 110 (2020).

59. Yokoi, T. *et al.* Identification of a unique subset of tissue-resident memory CD4 + T cells in Crohn's disease. *Proceedings of the National Academy of Sciences* **120,** e2204269120 (2023).

60. Harrington, R., Al Nokhatha, S. A. & Conway, R. JAK Inhibitors in Rheumatoid Arthritis: An Evidence-Based Review on the Emerging Clinical Data. *Journal of Inflammation Research* **13,** 519–531 (2020).

61. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Research* **49,** D1302–D1310 (2020).

62. Sonomoto, K. *et al.* Effects of tofacitinib on lymphocytes in rheumatoid arthritis: relation to efficacy and infectious adverse events. *Rheumatology* **53,** 914–918 (2014).

63. Gotthardt, D., Trifinopoulos, J., Sexl, V. & Putz, E. M. JAK/STAT Cytokine Signaling at the Crossroad of NK Cell Development and Maturation. *Frontiers in Immunology* **10,** 2590 (2019).

64. Betts, B. C. *et al.* Janus kinase-2 inhibition induces durable tolerance to alloantigen by human dendritic cell–stimulated T cells yet preserves immunity to recall antigen. *Blood* **118,** 5330–5339 (2011).

65. Kotschenreuther, K., Yan, S. & Kofler, D. M. Migration and homeostasis of regulatory T cells in rheumatoid arthritis. *Frontiers in Immunology* **13,** 947636 (2022).

66. Luo, P. *et al.* Immunomodulatory role of T helper cells in rheumatoid arthritis: a comprehensive research review. *Bone Joint Research* **11,** 426–438 (2022).

67. Sharfe, N., Dadi, H. K., O'Shea, J. J. & Roifman, C. M. Jak3 activation in human lymphocyte precursor cells. *Clinical and Experimental Immunology* **108,** 552–556 (1997).

68. Liu, M.-F. *et al.* Distribution of Double-Negative (CD4- CD8-, DN) T Subsets in Blood and Synovial Fluid from Patients with Rheumatoid Arthritis. *Clinical Rheumatology* **18,** 227–231 (1999).

69. Fuggle, N. R., Howe, F. A., Allen, R. L. & Sofat, N. New insights into the impact of neuroinflammation in rheumatoid arthritis. *Frontiers in neuroscience* **8,** 357 (2014).

70. Jain, M. *et al.* Role of JAK/STAT in the Neuroinflammation and its Association with Neurological Disorders. *Annals of Neurosciences* **28,** 191–200 (2021).

71. Dayer, J.-M. & Choy, E. Therapeutic targets in rheumatoid arthritis: the interleukin-6 receptor. *Rheumatology* **49,** 15–24 (2009).

72. Xu, Y.-D., Cheng, M., Shang, P.-P. & Yang, Y.-Q. Role of IL-6 in dendritic cell functions. *Journal of Leukocyte Biology* **111,** 695–709 (2021).

73. Choy, E. H. *et al.* Translating IL-6 biology into effective treatments. *Nature Reviews Rheumatology* **16,** 335–345 (2020).

74. Lopez-Santalla, M., Bueren, J. A. & Garin, M. I. Mesenchymal stem/stromal cell-based therapy for the treatment of rheumatoid arthritis: An update on preclinical studies. *eBioMedicine* **69,** 103427 (2021).

75. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nature Medicine* **28,** 2486–2496 (2022).

76. Gubatan, J. *et al.* Anti-Integrins for the Treatment of Inflammatory Bowel Disease: Current Evidence and Perspectives. *Clinical and Experimental Gastroenterology* **14,** 333–342 (2021).

77. Dotan, I. *et al.* The role of integrins in the pathogenesis of inflammatory bowel disease: Approved and investigational anti-integrin therapies. *Medicinal Research Reviews* **40,** 245–262 (2019).

78. Baumgart, D. C. Patients with active inflammatory bowel disease lack immature peripheral blood plasmacytoid and myeloid dendritic cells. *Gut* **54,** 228–236 (2005).

79. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46,** D1074–D1082 (2017).

80. Annese, V., Rogai, F., Settesoldi, A. & Bagnoli, S. PPARγ in Inflammatory Bowel Disease. *PPAR Research* **2012,** 1–9 (2012).

81. Duszka, K. *et al.* Intestinal PPARγ signalling is required for sympathetic nervous system activation in response to caloric restriction. *Scientific Reports* **6,** 36937 (2016).

82. Zhao, J., Zhao, R., Cheng, L., Yang, J. & Zhu, L. Peroxisome proliferator-activated receptor gamma activation promotes intestinal barrier function by improving mucus and tight junctions in a mouse colitis model. *Digestive and Liver Disease* **50,** 1195–1204 (2018).

83. Klepsch, V., Moschen, A. R., Tilg, H., Baier, G. & Hermann-Kleiter, N. Nuclear Receptors Regulate Intestinal Inflammation in the Context of IBD. *Frontiers in Immunology* **10,** 1070 (2019).

84. Duan, S. Z., Usher, M. G. & Mortensen, R. M. Peroxisome Proliferator-Activated Receptor-γ–Mediated Effects in the Vasculature. *Circulation Research* **102,** 283–294 (2008).

85. Kotlinowski, J. & Jozkowicz, A. PPAR Gamma and Angiogenesis: Endothelial Cells Perspective. *Journal of Diabetes Research* **2016,** 1–11 (2016).

86. Alkim, C., Alkim, H., Koksal, A. R., Boga, S. & Sen, I. Angiogenesis in Inflammatory Bowel Disease. *International Journal of Inflammation* **2015,** 1–10 (2015).

87. Yu, L., Gao, Y., Aaron, N. & Qiang, L. A glimpse of the connection between PPARγ and macrophage. *Frontiers in Pharmacology* **14,** 1254317 (2023).

88. Caër, C. & Wick, M. J. Human Intestinal Mononuclear Phagocytes in Health and Inflammatory Bowel Disease. *Frontiers in Immunology* **11,** 410 (2020).

89. Lakshmi, S. P., Reddy, A. T., Banno, A. & Reddy, R. C. Airway Epithelial Cell Peroxisome Proliferator–Activated Receptor γ Regulates Inflammation and Mucin Expression in Allergic Airway Disease. *The Journal of Immunology* **201,** 1775–1783 (2018).

90. Ghosh, S. *et al.* Pulmonary Manifestations of Inflammatory Bowel Disease and Treatment Strategies. *CHEST Pulmonary* **1,** 100018 (2023).

91. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580,** 402–408 (2020).

92. Zerrouk, N., Aghakhani, S., Singh, V., Augé, F. & Niarakis, A. A mechanistic cellular atlas of the rheumatic joint. *Frontiers in Systems Biology* **2,** 925791 (2022).

93. Korsunsky, I. *et al.* Cross-tissue, single-cell stromal atlas identifies shared pathological fibroblast phenotypes in four chronic inflammatory diseases. *Med* **3,** 481–518 (2022).

94. Ma, S., Chen, X., Zhu, X., Tsao, P. S. & Wong, W. H. Leveraging cell-type-specific regulatory networks to interpret genetic variants in abdominal aortic aneurysm. *Proceedings of the National Academy of Sciences* **119,** e2115601119 (2022).

95. Prieto-Vila, M. *et al.* Single-Cell Analysis Reveals a Preexisting Drug-Resistant Subpopulation in the Luminal Breast Cancer Subtype. *Cancer Research* **79,** 4412–4425 (2019).

96. Wang, Y.-Y. *et al.* CeDR Atlas: a knowledgebase of cellular drug response. *Nucleic Acids Research* **50,** D1164–D1171 (2021).

97. Hanley, C. J. *et al.* Single-cell analysis reveals prognostic fibroblast subpopulations linked to molecular and immunological subtypes of lung cancer. *Nature Communications* **14** (2023).

98. Huang, K., Jin, Y., Candes, E. & Leskovec, J. *Uncertainty Quantification over Graph with Conformalized Graph Neural Networks* in *Advances in Neural Information Processing Systems* (eds Oh, A. *et al.*) **36** (Curran Associates, Inc., 2023), 26699–26721.

99. Contextual learning is nearly all you need. *Nature Biomedical Engineering* **6,** 1319–1320 (2022).

100. Bode, D., Cull, A. H., Rubio-Lara, J. A. & Kent, D. G. Exploiting single-cell tools in gene and cell therapy. *Frontiers in immunology* **12,** 702636 (2021).

101. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47,** D529–D541 (2018).

102. Stark, C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34,** D535–D539 (2006).

103. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347,** 1257601 (2015).

104. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols* **15,** 1484–1506 (2020).

105. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Systems* **8,** 483–493.e7 (2019).

106. Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research* **39,** D507–D513 (2010).

107. Wen, Y., Zhang, K., Li, Z. & Qiao, Y. *A Discriminative Feature Learning Approach for Deep Face Recognition* in *Computer Vision – ECCV 2016* (eds Leibe, B., Matas, J., Sebe, N. & Welling, M.) (Springer International Publishing, 2016), 499–515.

108. Biewald, L. *Experiment Tracking with Weights and Biases* 2020.

109. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3,** 1–27 (1974).

110. Paszke, A. *et al. PyTorch: an imperative style, high-performance deep learning library* in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019).

111. Fey, M. & Lenssen, J. E. *Fast Graph Representation Learning with PyTorch Geometric* in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

112. Abid, A. *et al.* Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. *ICML Workshop on Human in the Loop Learning* (2019).

113. Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6,** 3021 (2021).

114. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine* **9,** eaag1166 (2017).

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

116. Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference* (2008).

117. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272 (2020).