EVENT

Week 11: Context-aware learning

VENUE & DATE

Course presentation
11/26/2025

PRESEDENTED BY

Jiying Zhang
First-year PHD
Patrick's Lab

École
polytechnique
fédérale
de Lausanne

Oveview

- Zhang et al., EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction, ICML 2024

- Li et al., Contextual AI models for single-cell protein biology, Nature Methods, 2024

Oveview

- Zhang et al., EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction, ICML 2024

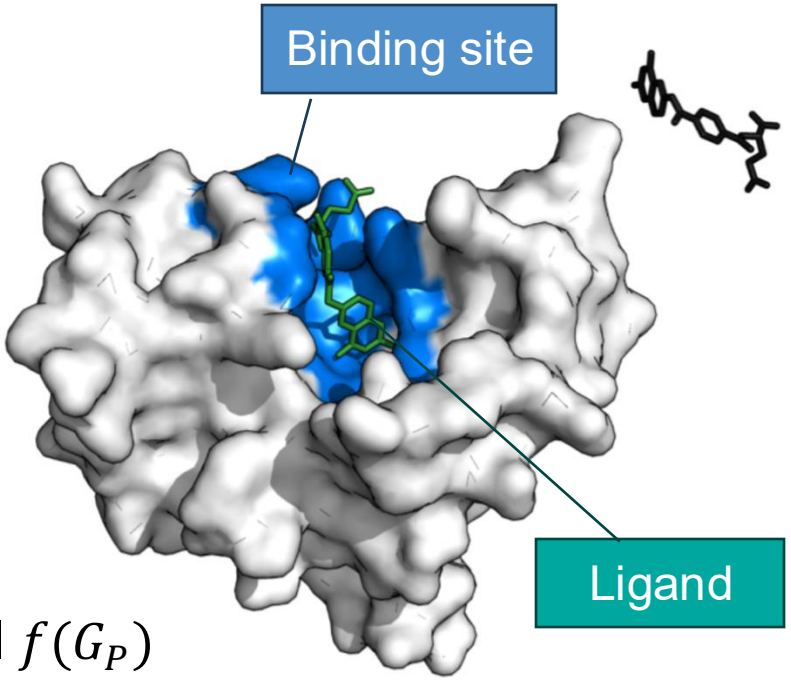
- Li et al., Contextual AI models for single-cell protein biology, Nature Methods, 2024

Background–Binding Site/Pocket

The Core Problem

Predicting the binding sites of target proteins.
It is a fundamental step in modeling drug discovery

- Protein–ligand docking tasks.
- Structure-based molecular generation
- Understanding biological interactions in living systems



Problem Statement. Given a protein $G_P = (\mathcal{V}_p, E_c, E_D)$, the model $f(G_P)$ is required to predict the atoms of **binding sites**: $\mathcal{V}_B \in \mathcal{V}_p$.

Previous works

Geometry-Based

Examples: Fpocket, LigSite

Exploit hand-crafted geometric algorithms to find hollow spaces. Limited expressivity

Machine Learning

Examples: P2Rank

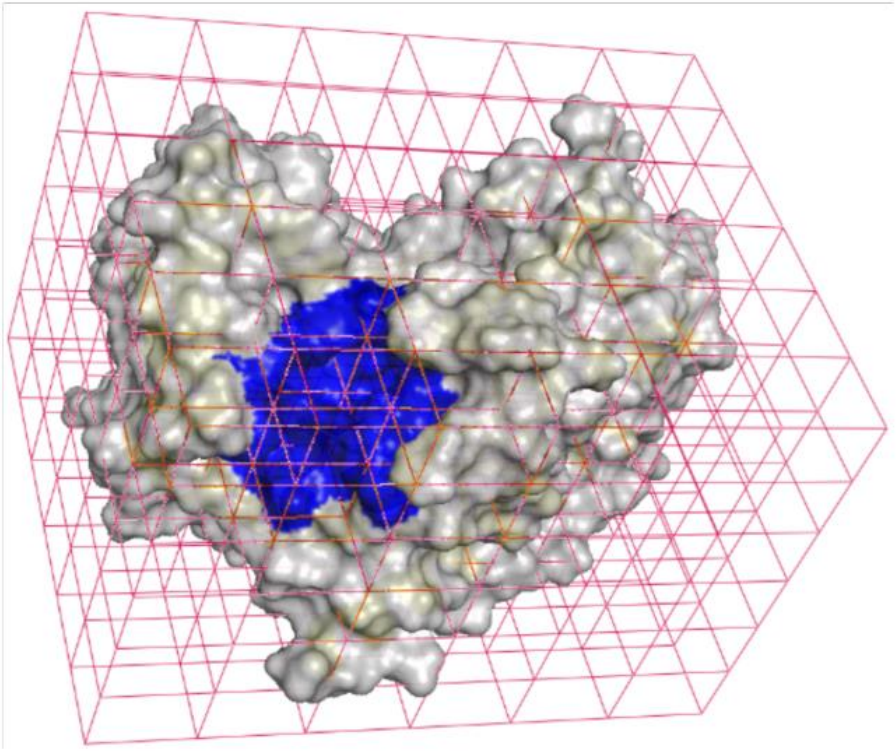
Use Random Forest with surface geometry features. Still relies on hand-crafted descriptors

Deep Learning (CNN)

Examples: DeepSite, DeepSurf

Treats protein as a 3D image (voxelization). Currently the dominant approach but has flaws

Limitations of CNN-based Methods



(a) CNN-based methods

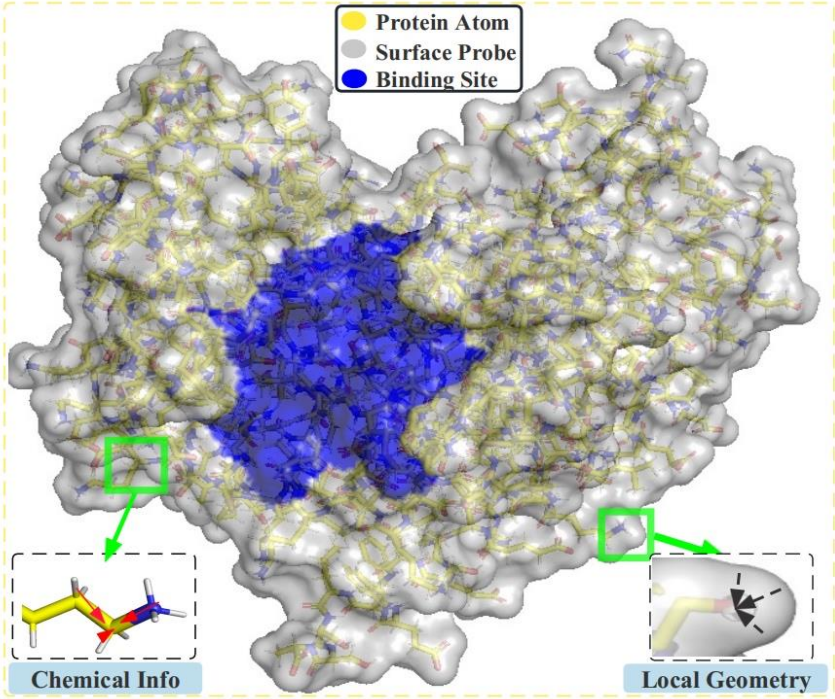
- **Irregular Structures**: Regular voxels struggle to model the naturally shape of proteins. (fixed box)
- **Rotation Sensitivity**: Rotation changes the input, confusing the model.
- **Surface Coarseness**: Voxel grids are too coarse to capture the fine-grained geometry of surface atoms.
- **Size Shift Unawareness**: Fixed grid sizes often cut off large proteins or waste space for small ones.

The Proposed Method: EquilPocket

Key Innovations

EquilPocket is an E(3)-equivariant GNN that addresses previous limitations

- **Graph-based:** naturally handles irregular protein
- **Equivariant:** Insensitive to rotation and translation
- **Surface-centric:** Specifically models surface geometry using probes
- **Adaptive:** Uses dense attention to handle variable protein sizes



Preliminaries & Definition

Protein Graphs

- Protein Graph $G_P = (V_P, E_C, E_D)$: Nodes V_P are atoms, edges E_C are chemical bonds and E_D are spatial neighbors

Surface Probe Set

- A set of 3D coordinates generated by rolling a solvent probe (e.g., water) across the protein surface (alg. **MSMS)**
  - $\boldsymbol{x_i} \in \mathrm{R}^3$: Its 3D coordinate.
  - $p_i \in V_p$: The index of its **Nearest Protein Atom**.

Surface Graph

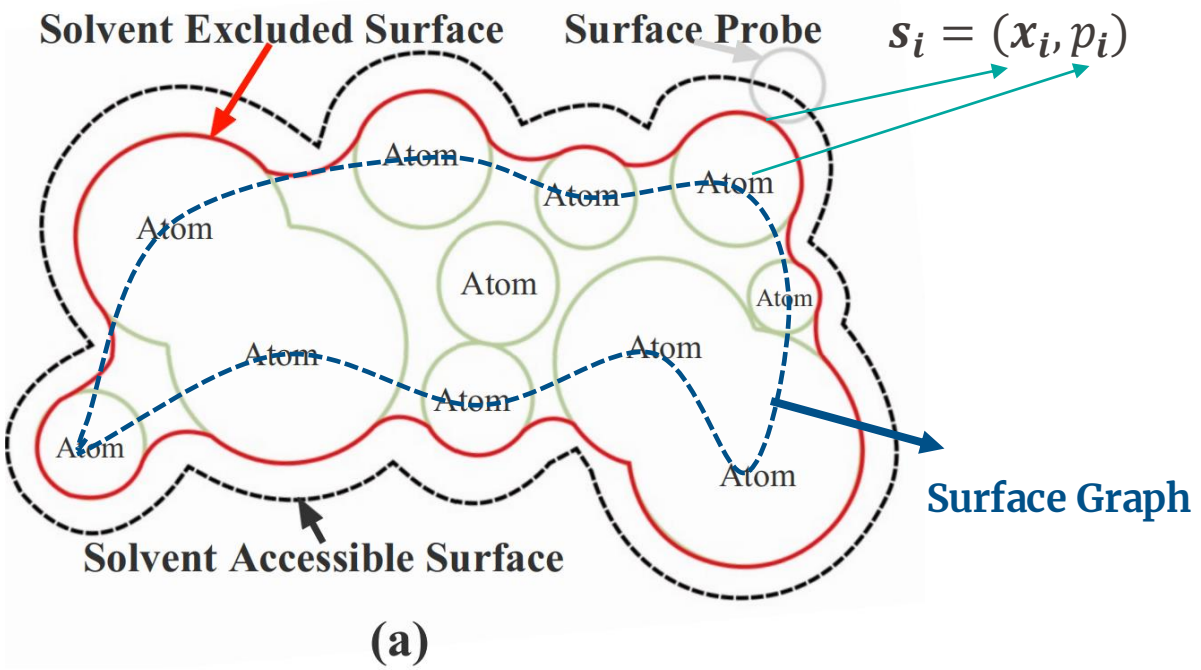
- $G_S = (V_s, E_s)$: $V_s = \{p_i\}$ (Nearest Protein Atoms).

E(3)-Equivariance

- A function f is equivariant if transforming the input transforms the output similarly:

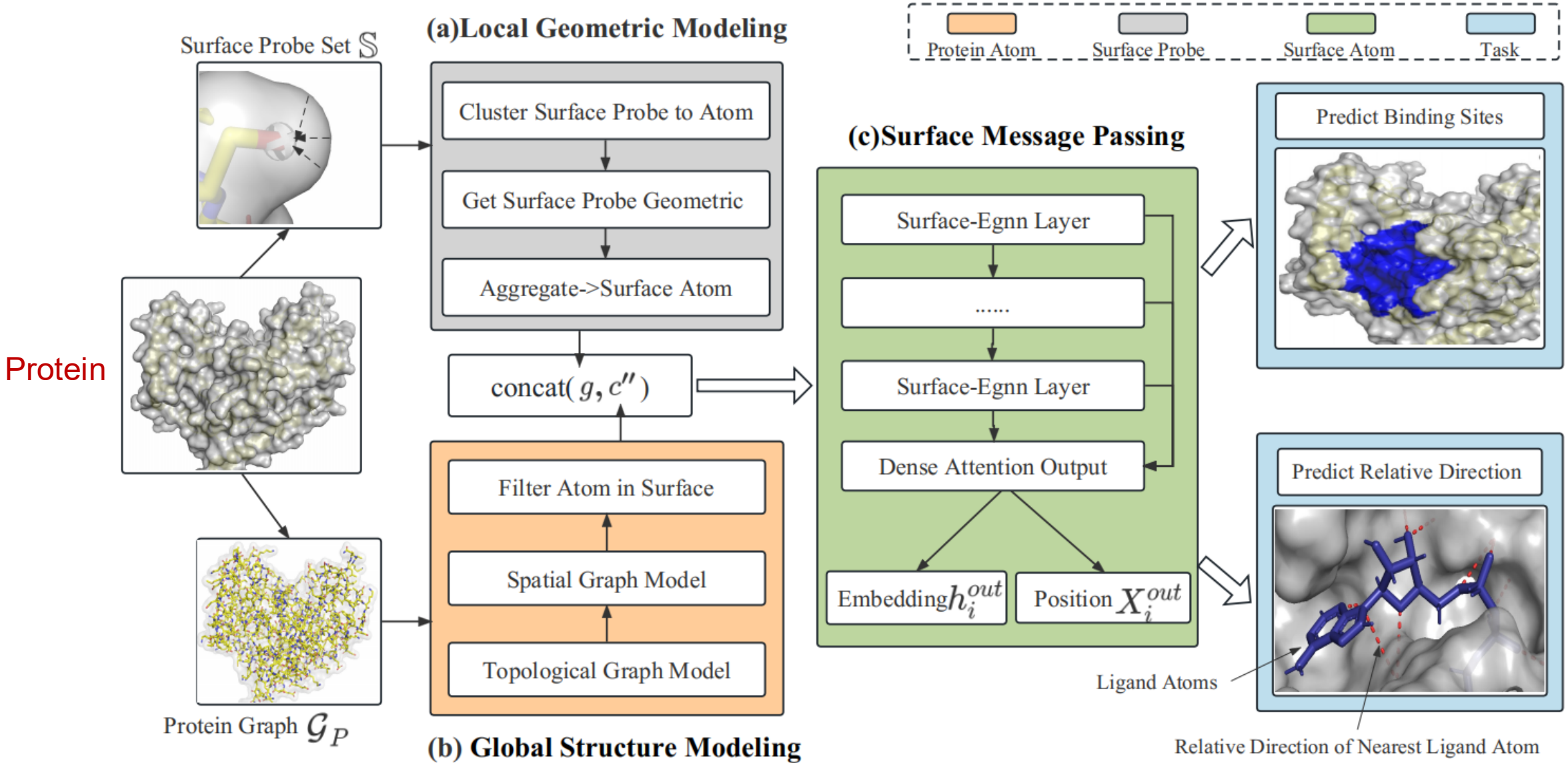
$$f(g \cdot X) = g \cdot f(X)$$

- Invariant: $f(g \cdot X) = f(X)$



EquilPocket Overview

Three modules：Local Geometric Modeling, Global Structure, Surface Message Passing Modeling



Extracting Surface Features (context)

- Collect surrounding probes $S_i = \{s_j = (\boldsymbol{x}_j, p_j) \in S \,|p_j = i\}$ for each surface atom $i \in V_S$.

Geometric descriptor g_i

- For each surrounding surface probe $s_j \in S_i$, we first search its two nearest surface probes from S as s_{j_1} and s_{j_2}, $\bar{\boldsymbol{x}}_i = \boldsymbol{mean}(S_i)$

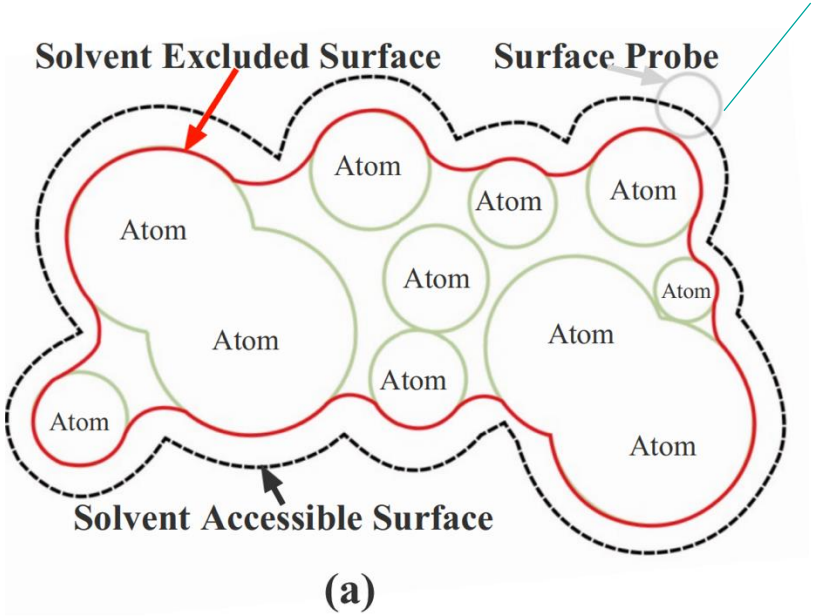
$$\boldsymbol{x}_{jj_1} = \boldsymbol{x}_j - \boldsymbol{x}_{j_1}, \boldsymbol{x}_{jj_2} = \boldsymbol{x}_j - \boldsymbol{x}_{j_2}, \boldsymbol{x}_{j,\text{center}} = \boldsymbol{x}_j - \bar{\boldsymbol{x}}_i,$$

$$\boldsymbol{x}_{j,\text{protein}} = \boldsymbol{x}_j - \boldsymbol{x}_i, \boldsymbol{x}_{\text{center,protein}} = \bar{\boldsymbol{x}}_i - \boldsymbol{x}_i.$$

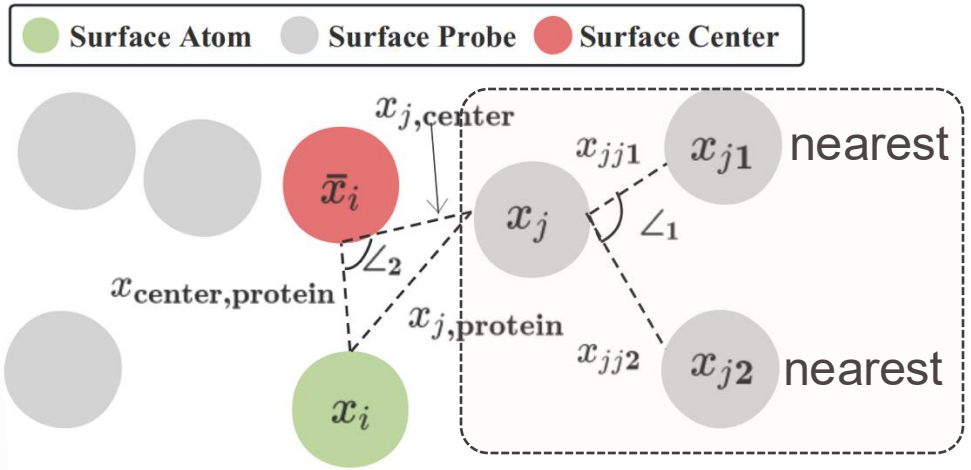
$$\boldsymbol{g}(s_j) := [\|\boldsymbol{x}_{jj_1}\|_2, \|\boldsymbol{x}_{jj_2}\|_2, \angle_1, \|\boldsymbol{x}_{j,\text{center}}\|_2, \|\boldsymbol{x}_{j,\text{protein}}\|_2, \|\boldsymbol{x}_{\text{protein,center}}\|_2, \angle_2],$$

$$\angle_1 = \frac{\boldsymbol{x}_{jj_1} \cdot \boldsymbol{x}_{jj_2}}{\|\boldsymbol{x}_{jj_1}\|_2 \|\boldsymbol{x}_{jj_2}\|_2} \qquad \angle_2 = \frac{\boldsymbol{x}_{j,\text{center}} \cdot \boldsymbol{x}_{\text{center,protein}}}{\|\boldsymbol{x}_{j,\text{center}}\|_2 \|\boldsymbol{x}_{\text{center,protein}}\|_2}$$

$$\boldsymbol{g}_i = [\text{Pooling}(\{\text{MLP}(\boldsymbol{g}(s_j))\}_{s_j \in \mathbb{S}_i}), \text{MLP}(\text{Pooling}\{(\boldsymbol{g}(s_j))\}_{s_j \in \mathbb{S}_i})].$$



(a)



(b)

Compute relative distances and angles between probes

Module 2 – Global Structure Modeling

Chemical-Graph Modeling

- Uses GNN to process chemical features (atom types) and chemical bonds

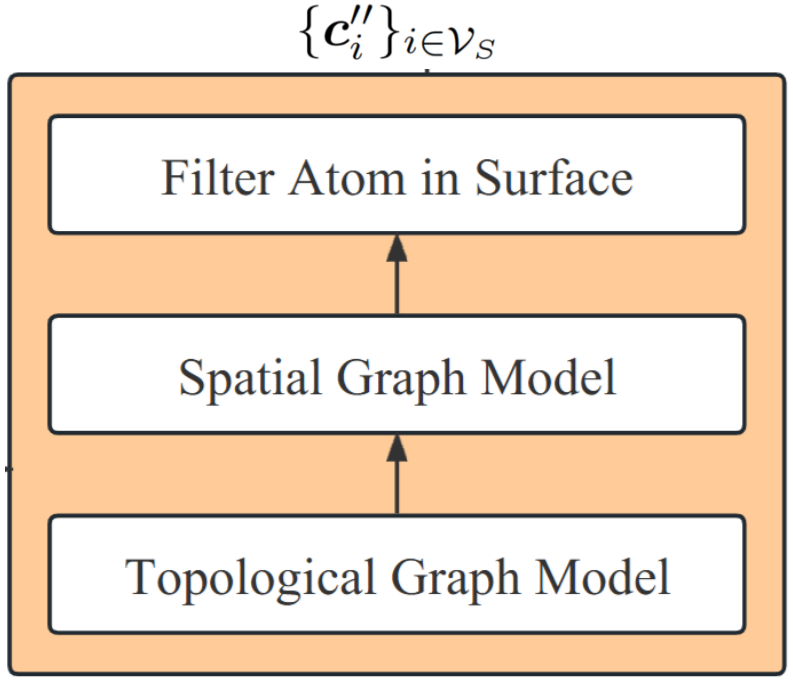
$$\{\boldsymbol{c}_i'\}_{i\in\mathcal{V}_P} = \mathbf{GNN}(\{\boldsymbol{c}_i\}_{i\in\mathcal{V}_P}, \mathcal{E}_C),$$

Spatial-Graph Modeling

- Uses EGNN to incorporate 3D coordinates x_i and spatial edges.

$$\{\boldsymbol{c}_i''\}_{i\in\mathcal{V}_P} = \mathbf{EGNN}(\{\boldsymbol{x}_i, \boldsymbol{c}_i'\}_{i\in\mathcal{V}_P}, \mathcal{E}_D).$$

- This captures the global structural context of the protein.



Module 3 – Surface Message Passing

We perform equivariant message passing exclusively on the **Surface Graph** to refine features.

Surface-EGNN Layer

- **Input:** Invariant features $h^{(0)} = [c'', g]$ and coordinates $X_i(0) = [x_i, \bar{x}_i]$.

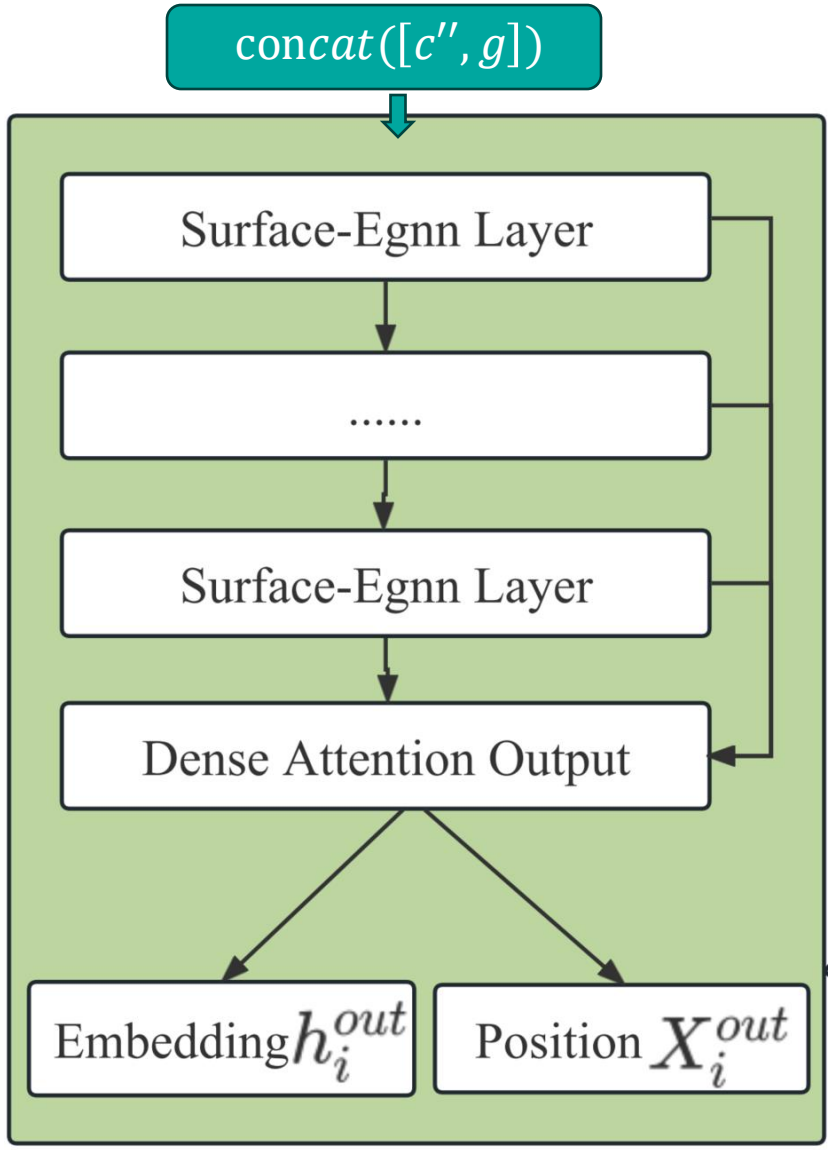
$$m_{ij} = \phi_m\left(h_i^{(l)}, h_j^{(l)}, f_x(X_i^{(l)}, X_j^{(l)}), e_{ij}\right),$$

$$h_i^{(l+1)} = \phi_h\left(h_i^{(l)}, \sum_{j\in\mathbb{N}(i)} m_{ij}\right),$$

$$X_i^{(l+1)} = X_i^{(l)} + \frac{1}{|\mathbb{N}(i)|}\sum_{j\in\mathcal{N}(i)}(x_{i,1}^{(l)} - x_{j,1}^{(l)})\phi_x(m_{ij}),$$

where ϕ_m, ϕ_h, ϕ_x are MLPs, and

$$f_x(X_i, X_j) = [\|x_{ij}\|_2, \|x_{ci}\|_2, \|x_{cj}\|_2, \angle_{ci,ij}, \angle_{cj,ij}, \angle_{ci,cj}],$$



Module 3 – Surface Message Passing

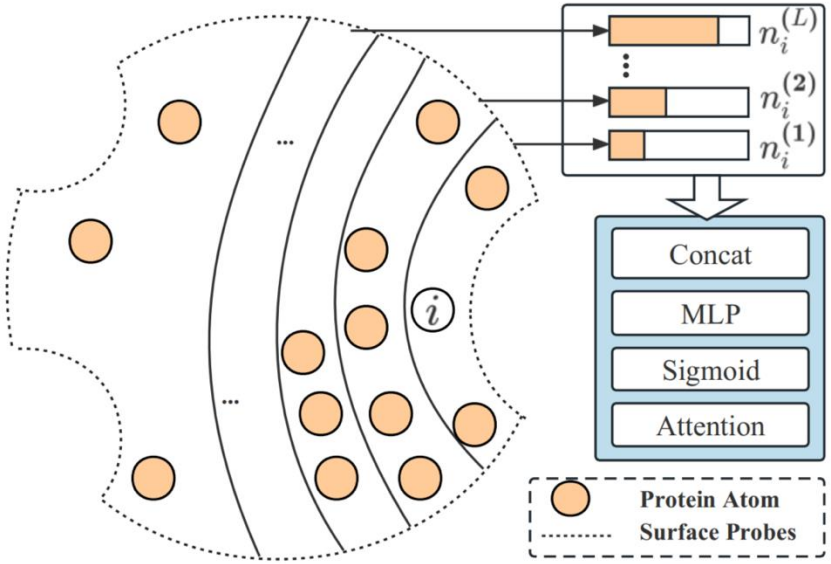
Challenge: A fixed receptive field causes over-smoothing in small proteins and insufficient coverage in large ones.

Dense Attention Output Layer.

1. Calculate the density of neighbors at different hop distances n_i

$$n_i^{(l)} = \frac{|\{j \in \mathcal{V}_P \mid 0 \leq \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 < l\theta\}|}{N_P},$$

$$\boldsymbol{n}_i = [n_i^{(0)}, n_i^{(1)}, \cdots, n_i^{(L)}, N_P] \in \mathbb{R}^{L+2}$$



2. An MLP generates attention weights a_i to adaptively balance information from different GNN layers.

$$\boldsymbol{a}_i = \text{Sigmoid}(\phi_a(\boldsymbol{n}_i))$$

$$\boldsymbol{h}_i^{\text{out}} = \text{Concat}(a_{i0}\boldsymbol{h}_i^{(0)}, ..., a_{iL}\boldsymbol{h}_i^{(L)}),$$

$$\boldsymbol{X}_i^{\text{out}} = \frac{1}{L+1}\sum_{l=0}^{L} \boldsymbol{X}_i^{(l)},$$

Training Objectives

Binding Site Prediction

- **Binary Classification**: label y =1 : surface atom i is within 4Å of a ligand.

$$\hat{y}_i = \textbf{Sigmoid}(\text{MLP}(\boldsymbol{h}_i^{\text{out}}))$$

- **Loss**: Dice Loss (L_b) handles class imbalance effectively.

Relative Direction

- **Auxiliary Task**: Predict the unit vector direction d_i from atom $\boldsymbol{x}_i \in G_p$ to the nearest ligand atom m_i. $\boldsymbol{d}_i = \frac{\boldsymbol{m}_i - \boldsymbol{x}_i}{\|\boldsymbol{m}_i - \boldsymbol{x}_i\|_2}$, $\hat{\boldsymbol{d}}_i = \frac{\boldsymbol{x}_i^{\text{out}} - \boldsymbol{x}_i}{\|\boldsymbol{x}_i^{\text{out}} - \boldsymbol{x}_i\|_2}$.

- **Loss**: Cosine Similarity Loss(L_d)

Total Loss

$$L = L_b + L_d$$

- Combines semantic segmentation with geometric regression

Experimental Setup

Datasets

Dataset	Role	Number of Binding Sites	Description
scPDB	Training	1	~16k entries, curated binding sites
COACH420	Testing	1,2,3,4,5…	Standard benchmark(mlig subset)
HOLO4K	Testing	1.2,3,4,5…	Larger, more complex proteins

Evaluation Metrics

- **DCC:** Distance between the predicted and true binding site center. (Success if < 4Å)
- **DCA:** Shortest distance from predicted center to any ligand atom

Failure Rate: % of proteins without any binding site center

$$\text{Success Rate(DCC/DCA)} = \frac{|\{\text{Predicted sites} \mid \text{DCC/DCA} < \overset{4\text{Å}}{\text{threshold}}\}|}{|\{\text{True sites}\}|}$$

$$\text{Failure Rate} = \frac{|\{\text{Protein}|\text{predicted binding center} = \emptyset\}|}{|\{\text{Protein}\}|}$$

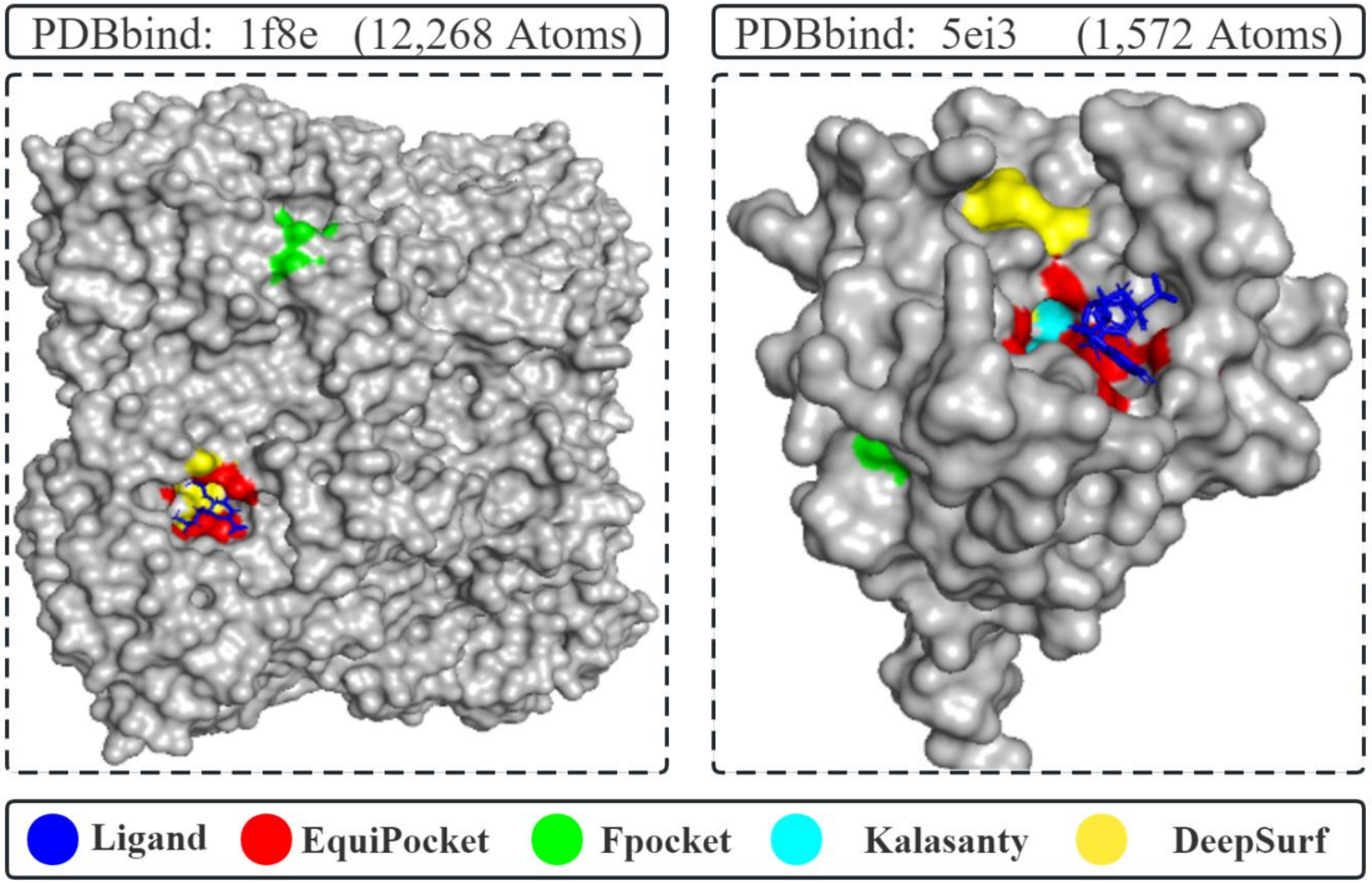
Comparison with SOTA

EquilPocket outperforms geometry-based, ML-based, and CNN-based methods

Methods	Type	Param (M)	Failure Rate ↓	COACH420		HOLO4K		PDBbind2020	
				DCC↑	DCA↑	DCC↑	DCA↑	DCC↑	DCA↑
Fpocket[b]	Geometric-based	\	**0.000**	0.228	0.444	0.192	0.457	0.253	0.371
P2rank[b]	Machine-learning	\	**0.000**	0.366	0.628	0.314	0.621	0.503	0.677
DeepSite[b]	CNN-based	1.00	\	\	0.564	\	0.456	\	\
Kalasanty[b]		70.6	0.120	0.335	0.636	0.244	0.515	0.416	0.625
DeepSurf[b]		33.1	0.054	0.386	**0.658**	0.289	0.635	0.510	0.708
RecurPocket[b]		21.2	0.075	0.354	0.593	0.277	0.616	0.492	0.663
GAT	Topological Graph	**0.03**	0.110	0.039(0.005)	0.130(0.009)	0.036(0.003)	0.110(0.010)	0.032(0.001)	0.088(0.011)
GCN		0.06	0.163	0.049(0.001)	0.139(0.010)	0.044(0.003)	0.174(0.003)	0.018(0.001)	0.070(0.002)
GCN2		0.11	0.466	0.042(0.098)	0.131(0.017)	0.051(0.004)	0.163(0.008)	0.023(0.007)	0.089(0.013)
SchNet	Spatial Graph	0.49	0.140	0.168(0.019)	0.444(0.020)	0.192(0.005)	0.501(0.004)	0.263(0.003)	0.457(0.004)
Egnn		0.41	0.270	0.156(0.017)	0.361(0.020)	0.127(0.005)	0.406(0.004)	0.143(0.007)	0.302(0.006)
EquiPocket-L	Ours	0.15	0.552	0.070(0.009)	0.171(0.008)	0.044(0.004)	0.138(0.006)	0.051(0.003)	0.132(0.009)
EquiPocket-G		0.42	0.292	0.159(0.016)	0.373(0.021)	0.129(0.005)	0.411(0.005)	0.145(0.007)	0.311(0.007)
EquiPocket-LG		0.50	0.220	0.212(0.016)	0.443(0.011)	0.183(0.004)	0.502(0.008)	0.274(0.004)	0.462(0.005)
EquiPocket		1.70	0.051	**0.423(0.014)**	0.656(0.007)	**0.337(0.006)**	**0.662(0.007)**	**0.545(0.010)**	**0.721(0.004)**

Case study

EquilPocket outperforms geometry-based, ML-based, and CNN-based methods

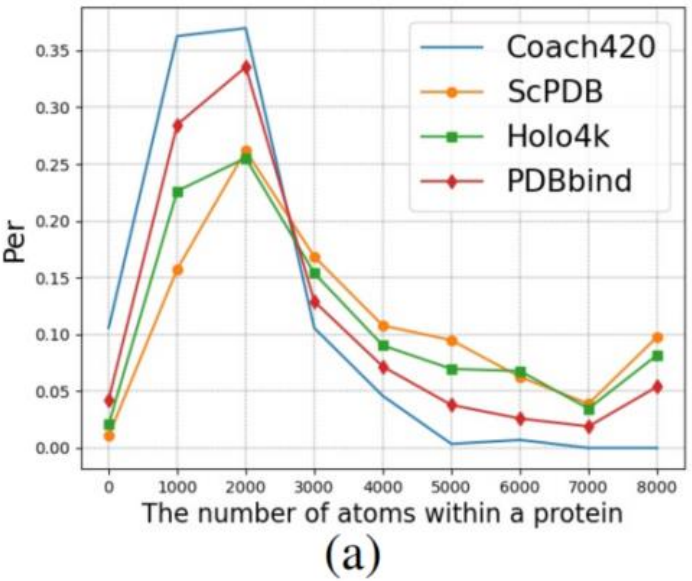


Ablation & Size Shift Analysis

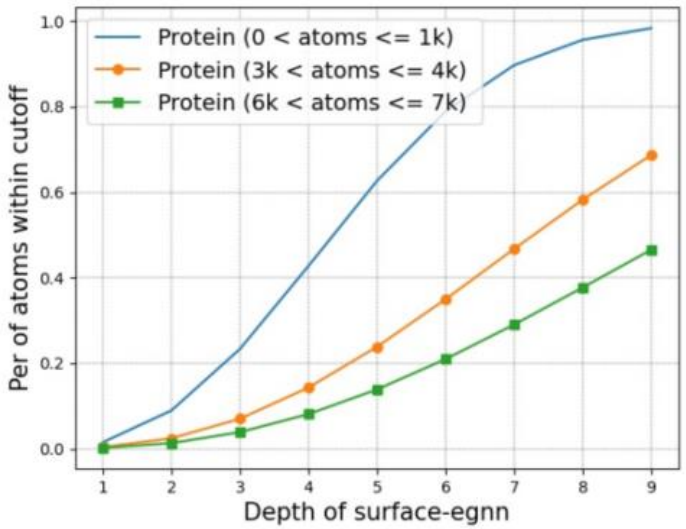
EquilPocket outperforms geometry-based, ML-based, and CNN-based methods

Methods	Type	Param (M)	Failure Rate ↓	COACH420		HOLO4K		PDBbind2020	
				DCC↑	DCA↑	DCC↑	DCA↑	DCC↑	DCA↑
EquiPocket-L	Ours	0.15	0.552	0.070(0.009)	0.171(0.008)	0.044(0.004)	0.138(0.006)	0.051(0.003)	0.132(0.009)
EquiPocket-G		0.42	0.292	0.159(0.016)	0.373(0.021)	0.129(0.005)	0.411(0.005)	0.145(0.007)	0.311(0.007)
EquiPocket-LG		0.50	0.220	0.212(0.016)	0.443(0.011)	0.183(0.004)	0.502(0.008)	0.274(0.004)	0.462(0.005)
EquiPocket		1.70	0.051	**0.423(0.014)**	0.656(0.007)	**0.337(0.006)**	**0.662(0.007)**	**0.545(0.010)**	**0.721(0.004)**

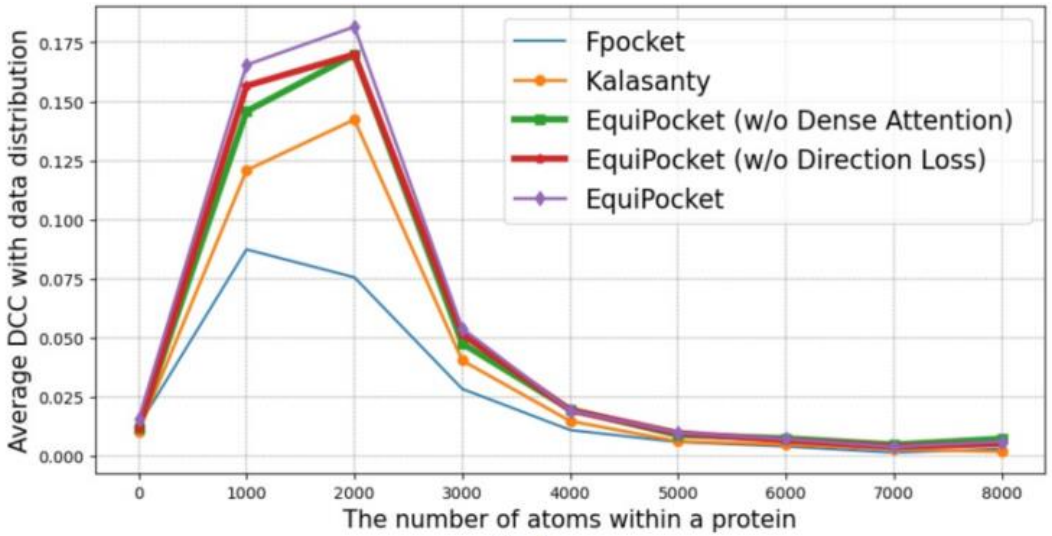
The protein size shift and model performances for proteins of various sizes



(a)



(b)



(c)

Summary: Positive & Negative

EquiPocket is an **E(3)-Equivariant** Graph Neural Network (GNN) that leverages a **dual-graph architecture** (atomic and surface) to integrate **geometric context** for rotation-invariant **ligand binding site prediction**.

👍 Positives:

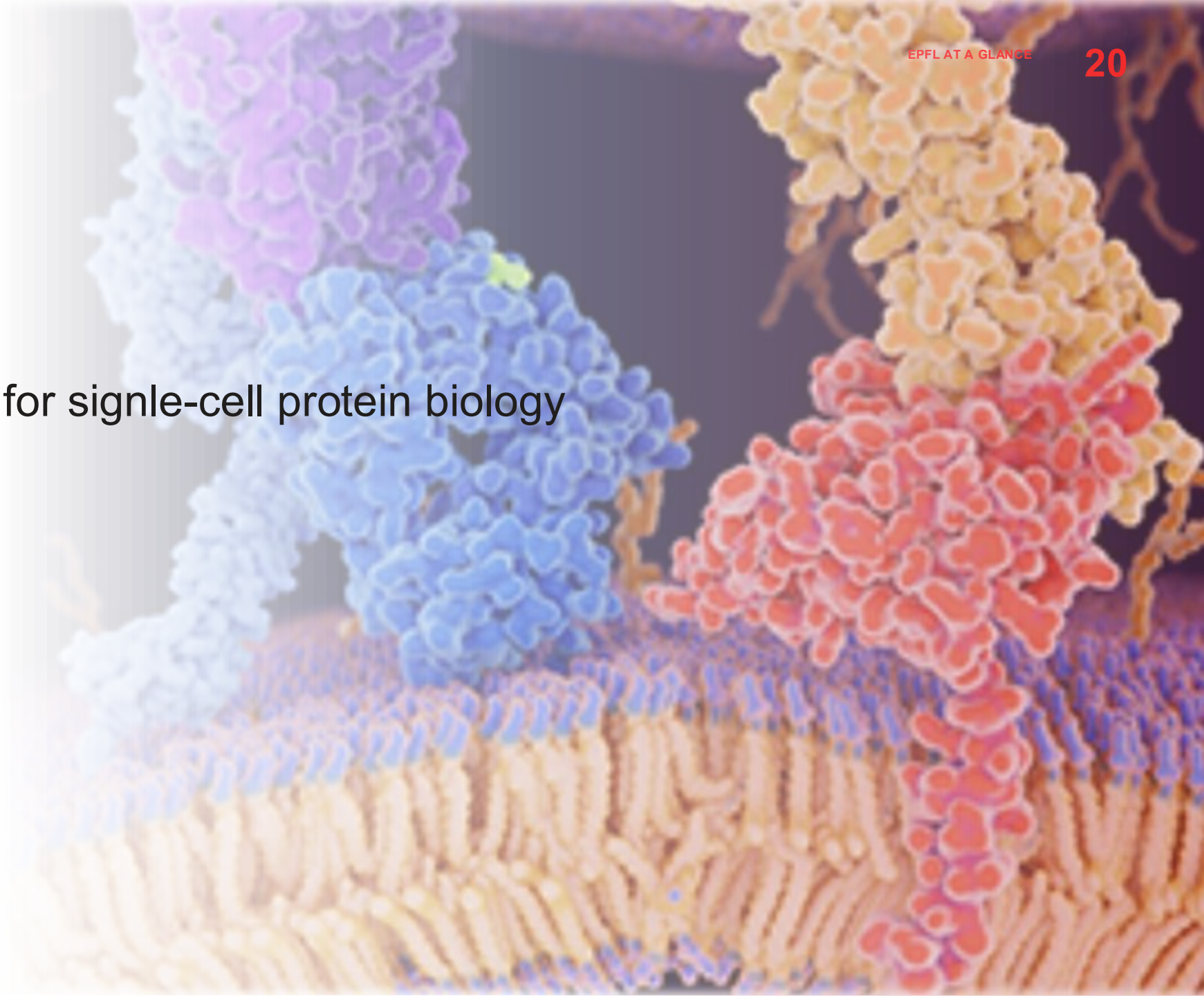
- **E(3) Equivariance**: Guarantees rotation/translation invariance for the final prediction and consistency of feature learning, improving generalization.
- **Geometric Context**: Effectively models the protein surface geometry to provide crucial context for binding site prediction.

⚠ Limitations:

- Scalability Issue: Performance can deteriorate or become inefficient for very large proteins due to the sheer size of the surface probe set and complex graph message passing.

Oveview

- EquiPocket:
- Contextual AI models for signle-cell protein biology

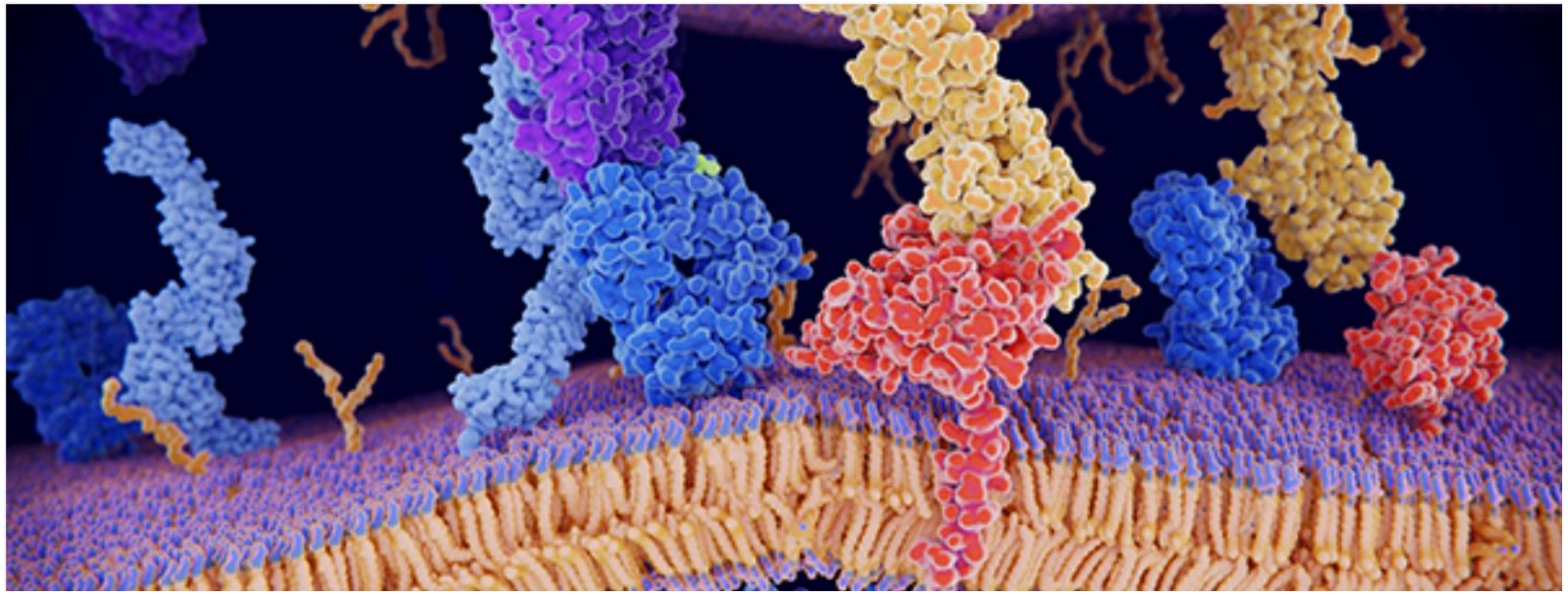


Background—protein in the biological contexts

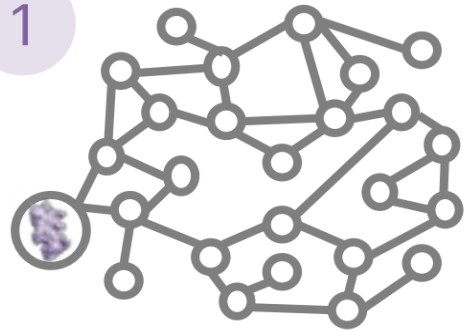
Importance of biological contexts (e.g. cellular and tissue)

Understanding protein function and developing molecular therapies require deciphering the **cell types** in which proteins act as well as the interactions between proteins.

- The expression of genes and the function of proteins encoded by these genes depend on **cellular and tissue contexts**
- Protein functions and interactions will be different in various biological contexts.
- Modeling protein interactions across biological contexts remains challenging for existing algorithms

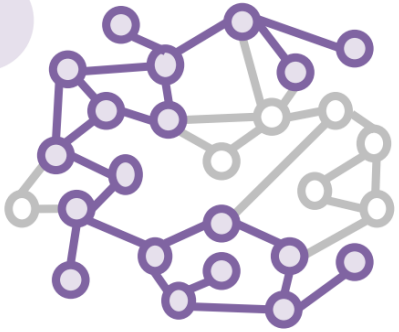


1



Protein interaction network

2

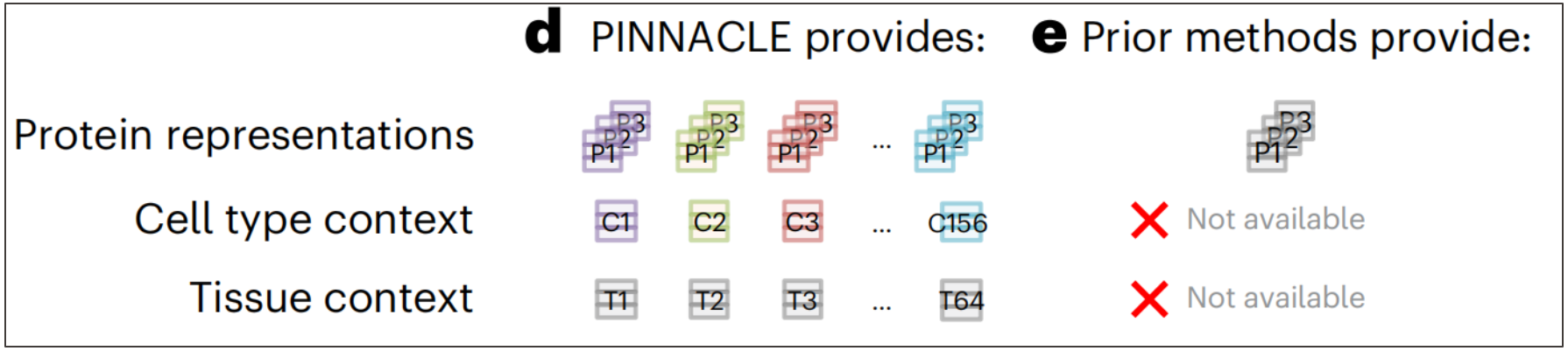


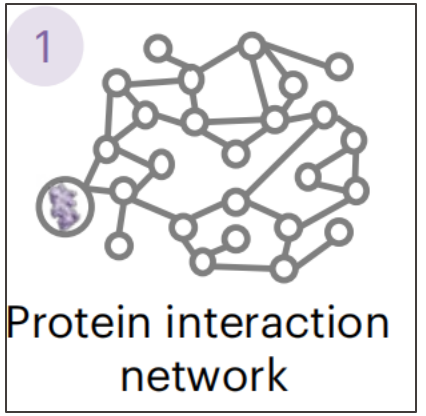
Activated genes in the cell type

Context-free protein representation learning

- **Existing works:** Generate one representation for each protein, not tailored to specific biological contexts (cell types and disease states) (e.g. GAT)

- **Weakness:** These representations cannot identify protein functions that vary across different cell types, which in turn hamper the prediction of protein roles in a cell type-specific manner.





The Data Revolution

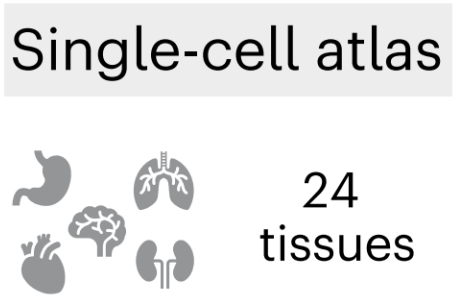
Single-Cell Atlases

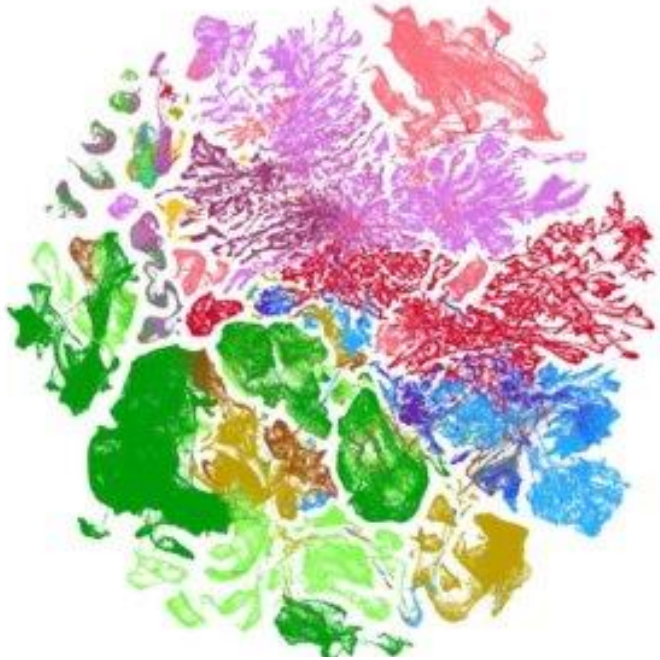
We are in the era of **single-cell transcriptomics**. Resources like *Tabula Sapiens* provide a high-resolution "parts list" of activated genes across hundreds of unique cell types and tissues.

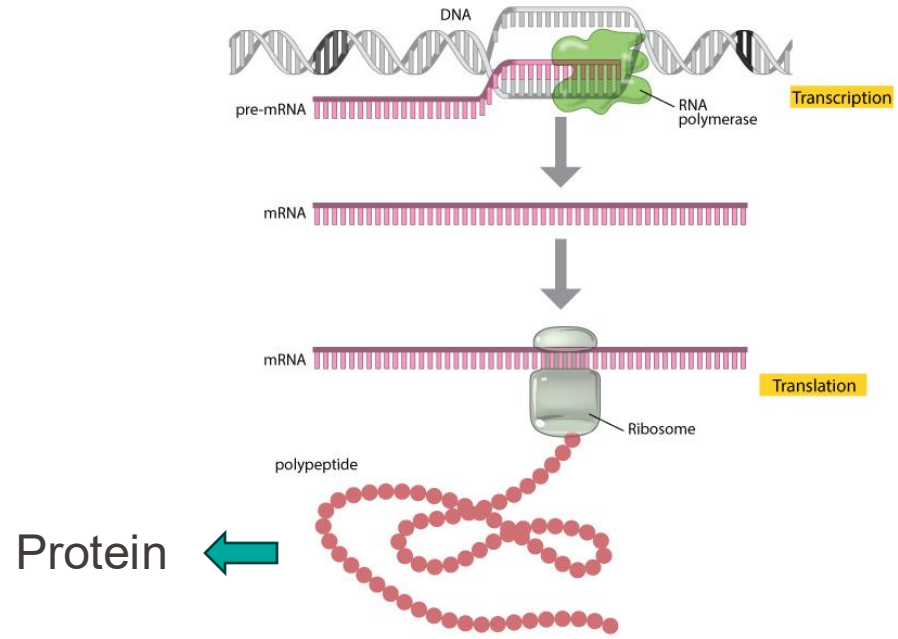
The Missing Link

While we have the **schematic** (protein networks) and the **inventory** (single-cell atlases), they have historically remained separate.

The challenge lies in effectively merging these massive, disparate datasets to inform protein models.







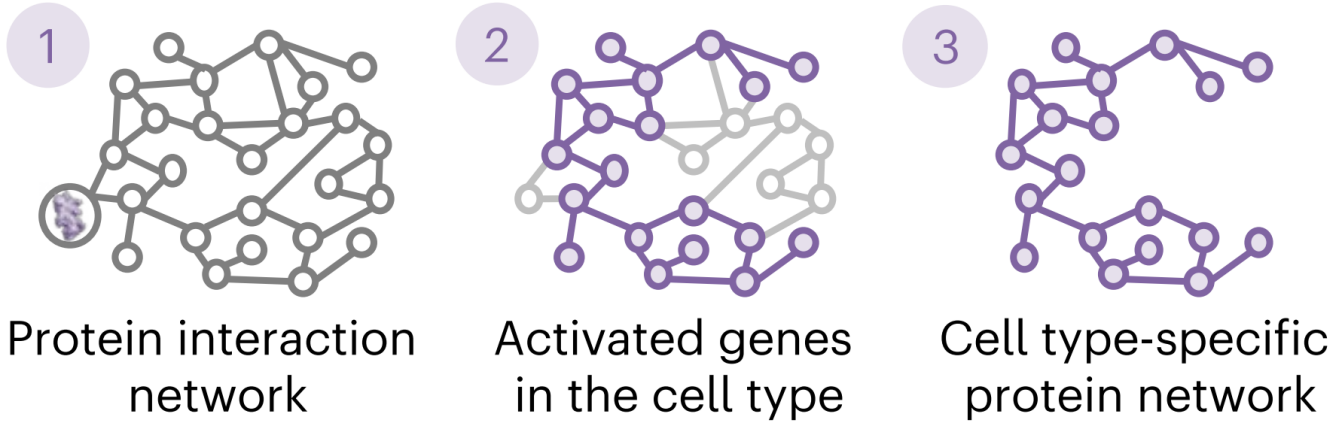
Contribution: Context-specific model PINNACLE (Protein Network-based Algorithm for Contextual Learning)

Geometric Deep Learning

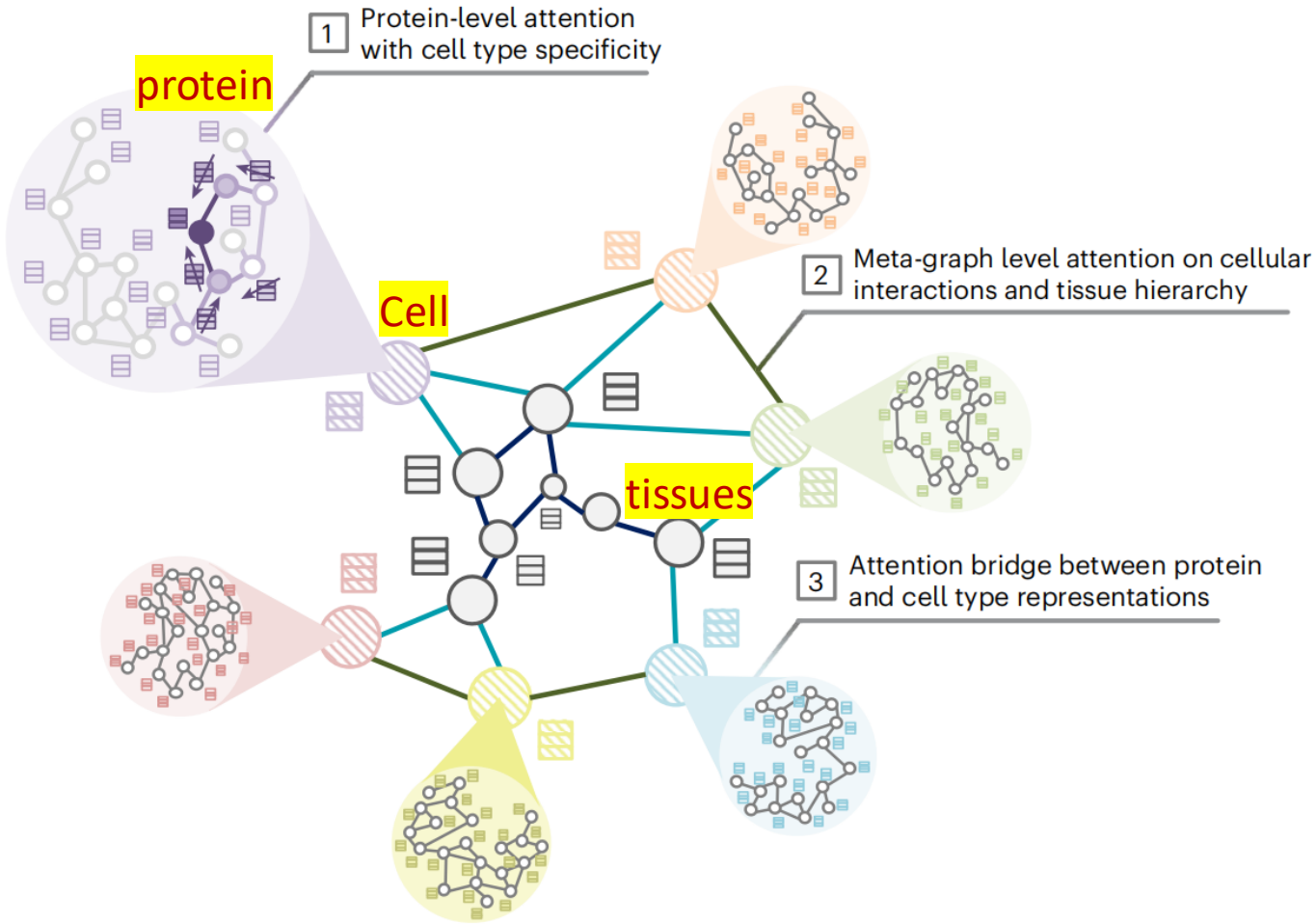
PINNACLE integrates single-cell transcriptomics, protein-protein interaction networks, and tissue hierarchies. It moves beyond static models to generate high-resolution protein representations.

Distinct Cell-Type Specificity

Instead of one single vector **per protein**, PINNACLE generates **394,760 distinct representations**—one for every cell type in which a protein is active. This allows for precise, cell-type-specific therapeutic targeting.



Overview of PINNACLE



➢ PPI

Cell type-specific PPI (Protein-protein interation) networks $\{G_i\}$

- Edges: Protein-Protein interation

➢ Metagraph:

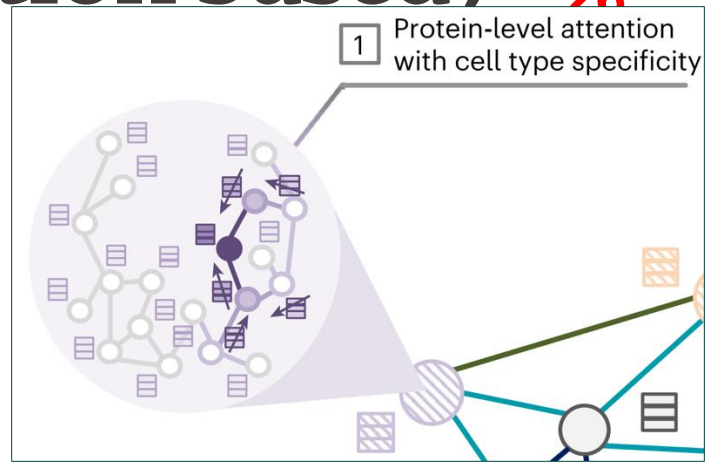
Cell types and tissues form a network.

- Edges
  - cell type-cell type interactions
  - cell type-tissue associations
  - Tissue–tissue relationships

PINNACLE: Multiscale GNN (Attention based)

1. Protein-level attention with cell type specificity.

$$\mathbf{h}_u^{\mathrm{PP}} \leftarrow \mathrm{AGG}\left(\sigma\left(\sum_{v\in\mathcal{N}_u} \alpha_{u,v} W^{\mathrm{PP}} \mathbf{h}_v^{\mathrm{PP}}\right)\right) \qquad \alpha_{u,v} = \frac{\exp(\sigma(\mathbf{a}^T\cdot[\mathbf{h}_u \| \mathbf{h}_v]))}{\sum_{v\in\mathcal{N}_u}\exp(\sigma(\mathbf{a}^T\cdot[\mathbf{h}_u \| \mathbf{h}_v]))}$$



2. Metagraph-level attention on cellular interactions and tissue hierarchy.

$$\mathbf{h}_{c_i}^{\mathrm{CC}} \leftarrow \mathrm{AGG}\left(\sigma\left(\sum_{c\in\mathcal{N}_{c_i}} \alpha_{c_i,c} W^{\mathrm{CC}} \mathbf{h}_c^{\mathrm{CC}}\right)\right) \quad (2) \qquad \mathbf{h}_{t_i}^{\mathrm{TT}} \leftarrow \mathrm{AGG}\left(\sigma\left(\sum_{t\in\mathcal{N}_{t_i}} \alpha_{t_i,t} W^{\mathrm{TT}} \mathbf{h}_t^{\mathrm{TT}}\right)\right) \quad (5)$$

$$\mathbf{h}_{c_i}^{\mathrm{CT}} \leftarrow \mathrm{AGG}\left(\sigma\left(\sum_{t\in\mathcal{N}_{c_i}} \alpha_{c_i,t} W^{\mathrm{CT}} \mathbf{h}_t^{\mathrm{CT}}\right)\right) \quad (3) \qquad \mathbf{h}_{t_i}^{\mathrm{TC}} \leftarrow \mathrm{AGG}\left(\sigma\left(\sum_{c\in\mathcal{N}_{t_i}} \alpha_{t_i,c} W^{\mathrm{TC}} \mathbf{h}_c^{\mathrm{TC}}\right)\right) \quad (6)$$

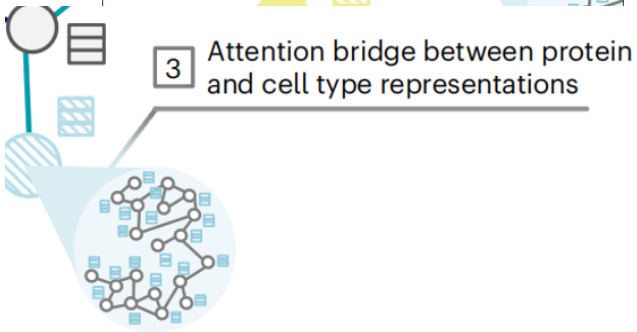
$$\mathbf{h}_{c_i} = \beta^{\mathrm{CC}}\mathbf{h}_{c_i}^{\mathrm{CC}} + \beta^{\mathrm{CT}}\mathbf{h}_{c_i}^{\mathrm{CT}} \quad (4) \qquad \mathbf{h}_{t_i} = \beta^{\mathrm{TT}}\mathbf{h}_{t_i}^{\mathrm{TT}} + \beta^{\mathrm{TC}}\mathbf{h}_{t_i}^{\mathrm{TC}} \quad (7)$$

learned attention weights $\beta^r, r = CC, CT, TC, TT,$ $\beta^r = \frac{\exp(m_r)}{\sum_{r\in R}\exp(m_r)}$;

$$m_r = \sum_{u\in V_q} \mathbf{s}^T \cdot \tanh(M\cdot \mathbf{h}_u^r + \mathbf{b})$$

3. Bridge between protein and cell type embeddings.

$$\mathbf{h}_{c_i} \leftarrow \mathbf{h}_{c_i} + \mathrm{AGG}\left(\sigma\left(\sum_{u\in V_{c_i}} \gamma_{c_i,u}\mathbf{h}_u\right)\right). \qquad \mathbf{h}_u \leftarrow \mathbf{h}_u + \gamma_{c_i,u}\mathbf{h}_{c_i}.$$



Training objective

Self-supervised learning

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

- Each cell type-specific PPI network (link prediction) and cell type identity

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

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

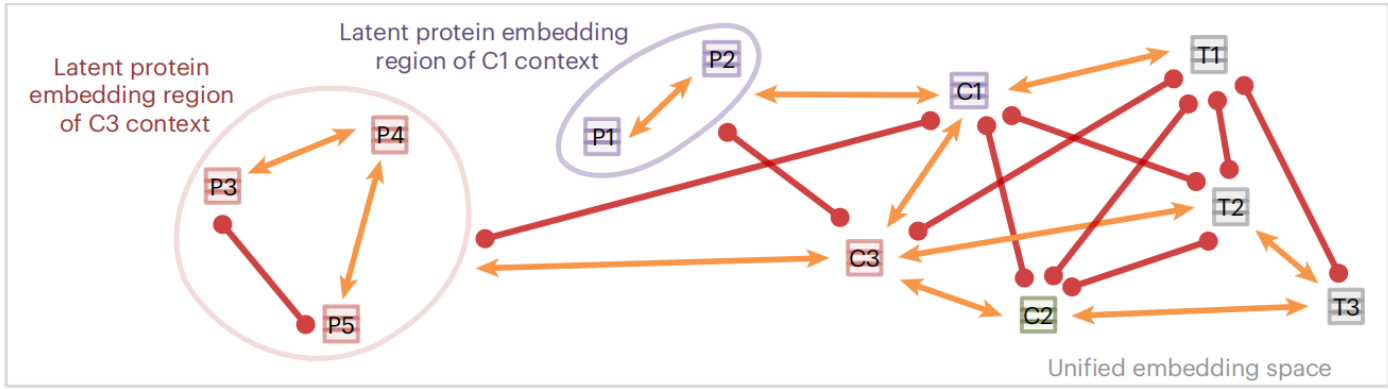
$$\mathcal{L}_{\text{celltypeid}} = \sum_{c_i \in \mathcal{C}} \sum_{u \in V_{c_i}} ||\mathbf{z}_u - \mathbf{z}_{c_i}||_2^2$$ (cell type identity)

- Cell type level (link prediction) BCE(C-C + C-T)

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

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

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



- Tissue level (link prediction) BCE(T-T + T-C)

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

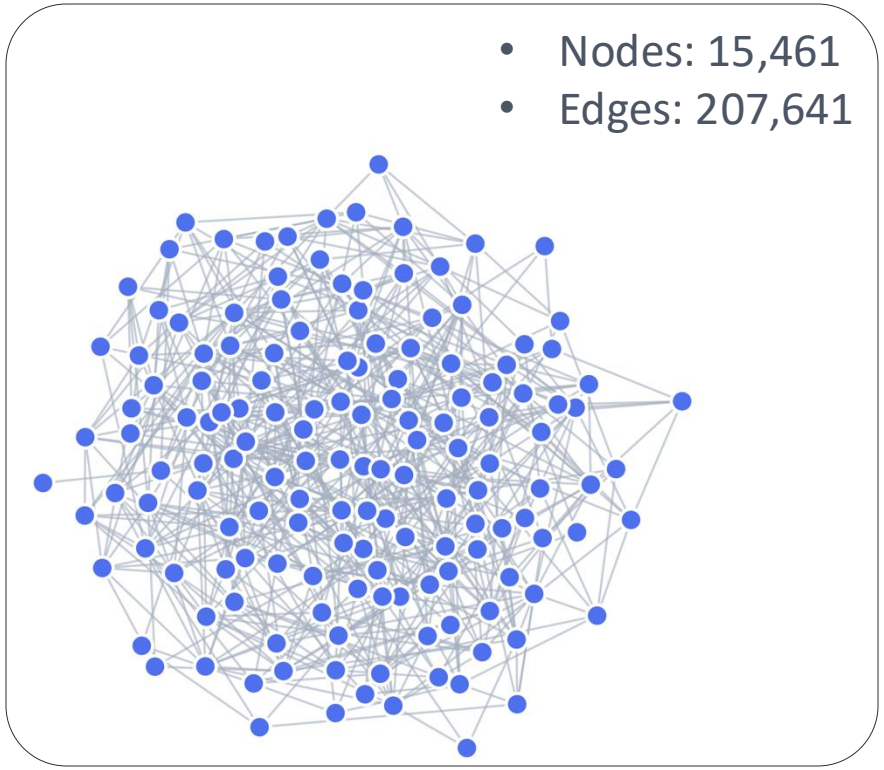
$$\mathcal{L}_{\text{tissue}}^{\text{TT}} = \sum_{t_k, t_q \in \mathcal{T}} y_{t_k,t_q} \log(\hat{y}_{t_k,t_q}) + (1-y_{t_k,t_q}) \log(1-\hat{y}_{t_k,t_q})$$

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

Pretraining data

Reference human physical PPI network

BioGRID, HuRI and Menche et al[1] with 15,461 nodes and 207,641 edges.

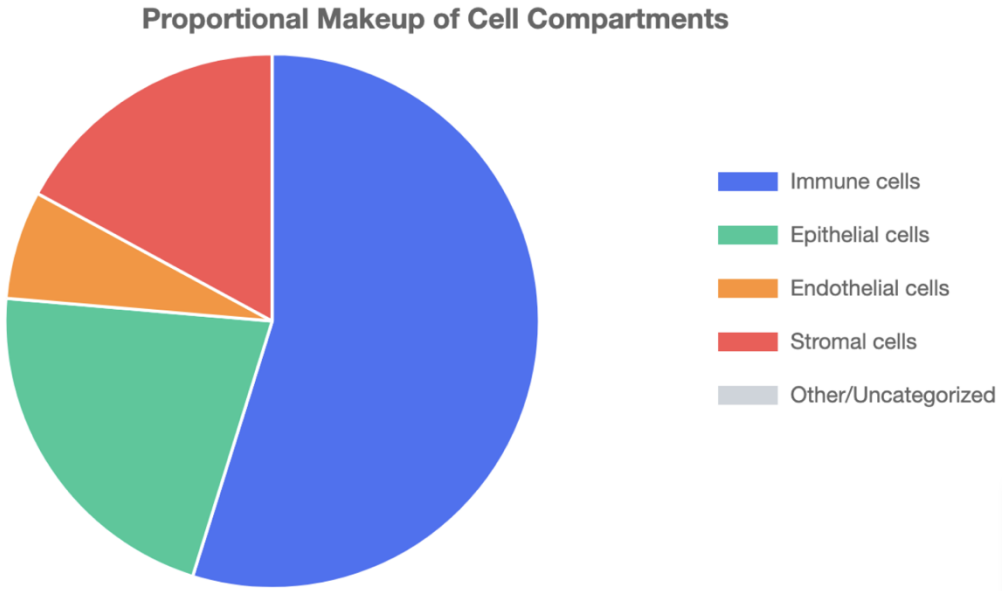


Multiorgan, single-cell transcriptomic atlas of humans.

Total Cells: 483,152 (from 15 Donors, 59 Specimens) *Tabula Sapiens*

Data Summary:

- Immune cells: **264,824** (54.81%)
- Epithelial cells: **104,148** (21.56%)
- Endothelial cells: **31,691** (6.56%)
- Stromal cells: **82,478** (17.07%)

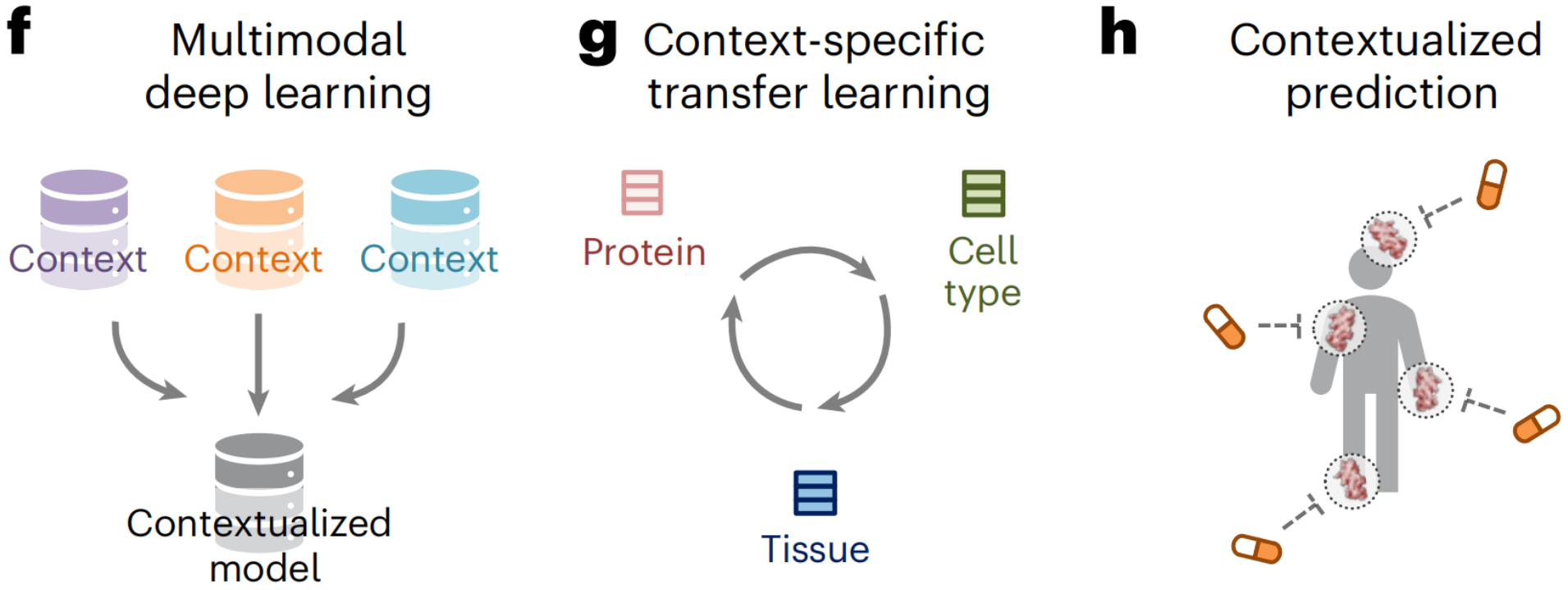


Menche, J. et al. Uncovering disease–disease relationships through the incomplete interactome. *Science* **347**, 1257601(2015).

Application

Tasks:

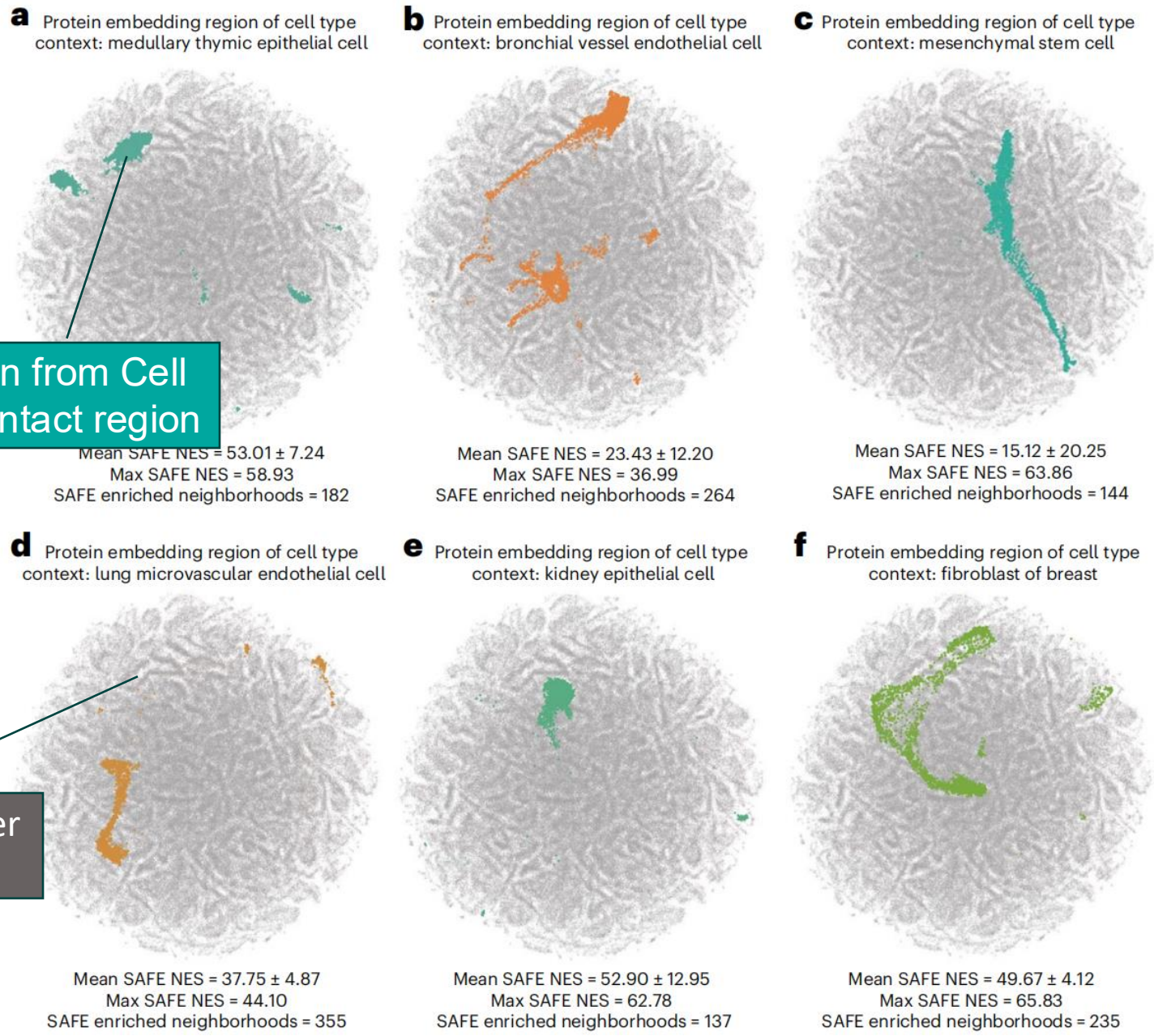
- enhance three-dimensional (3D) structural protein representations,
- analyze the effects of drugs across cell type contexts, nominate therapeutic targets in a cell type-specific manner
- retrieve tissue hierarchy in a zero-shot manner
- perform context-specific transfer learning



Results: captures cellular and tissue organization

Two dimensional UMAP(uniform manifold approximate projection) plots of contextualized **protein representations** generated by PINNACLE from six different cell type context.

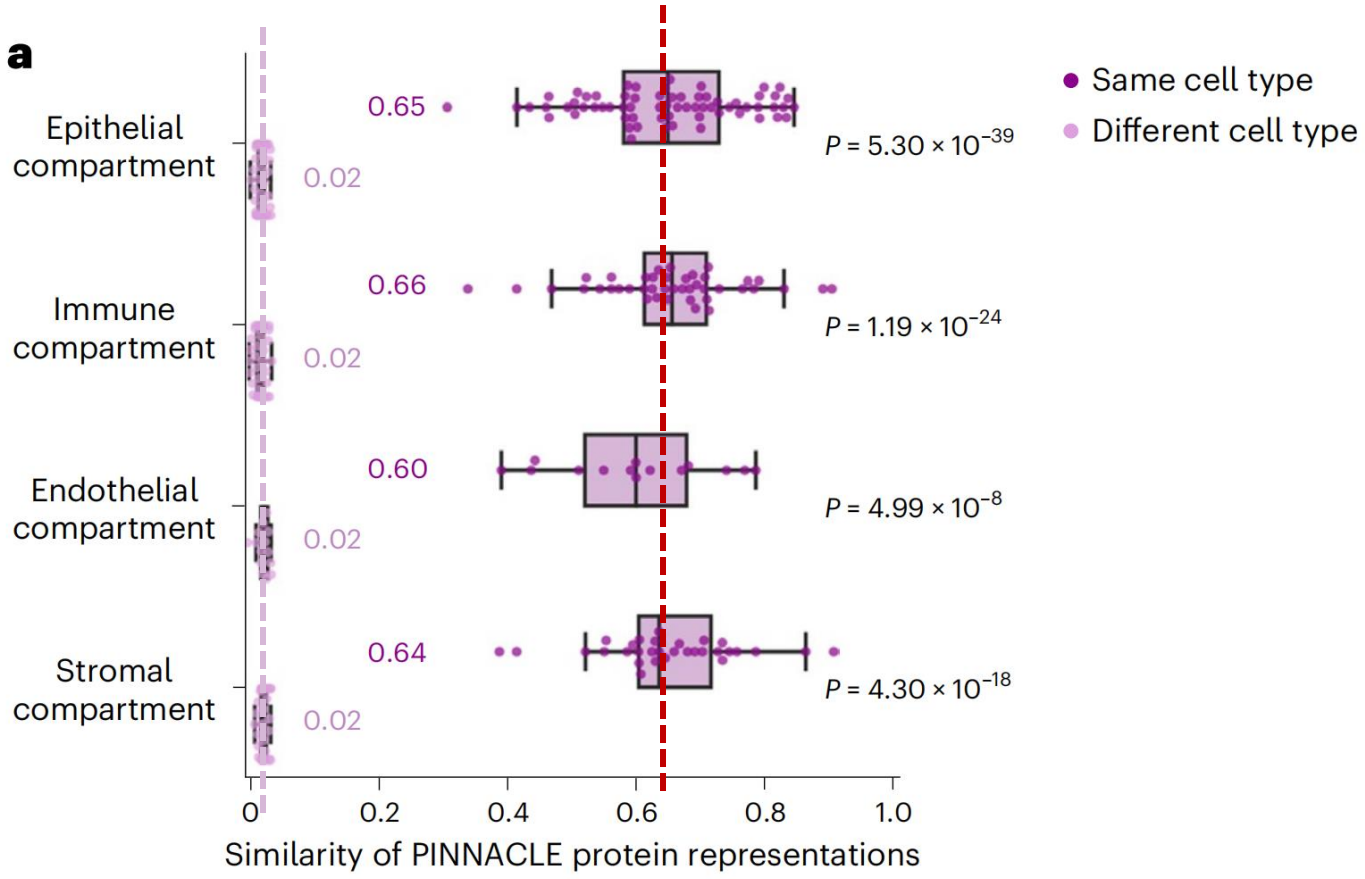
Quantified by neighborhood enrichment scores (NES), the higher is better.



Embedding similarities using PINNACLE's protein representations

Same cell type v.s. Different Cell type

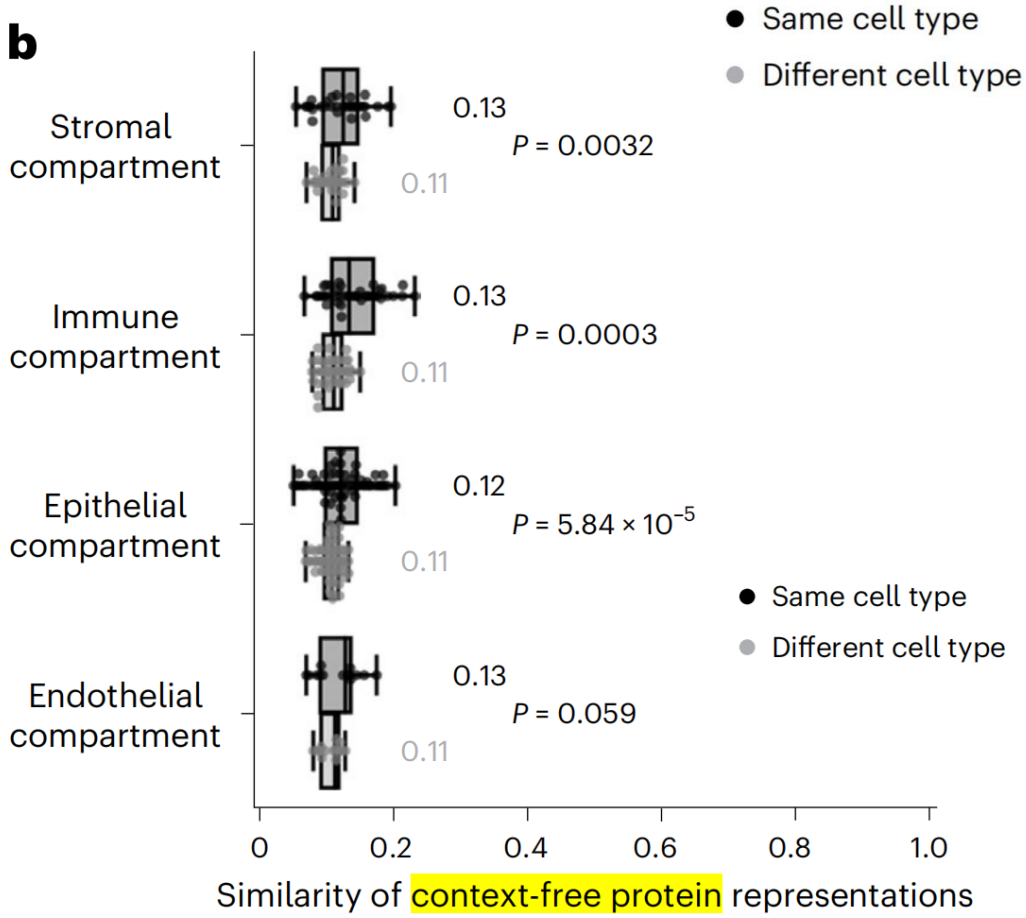
Protein presentations from the same cell type show high similarity



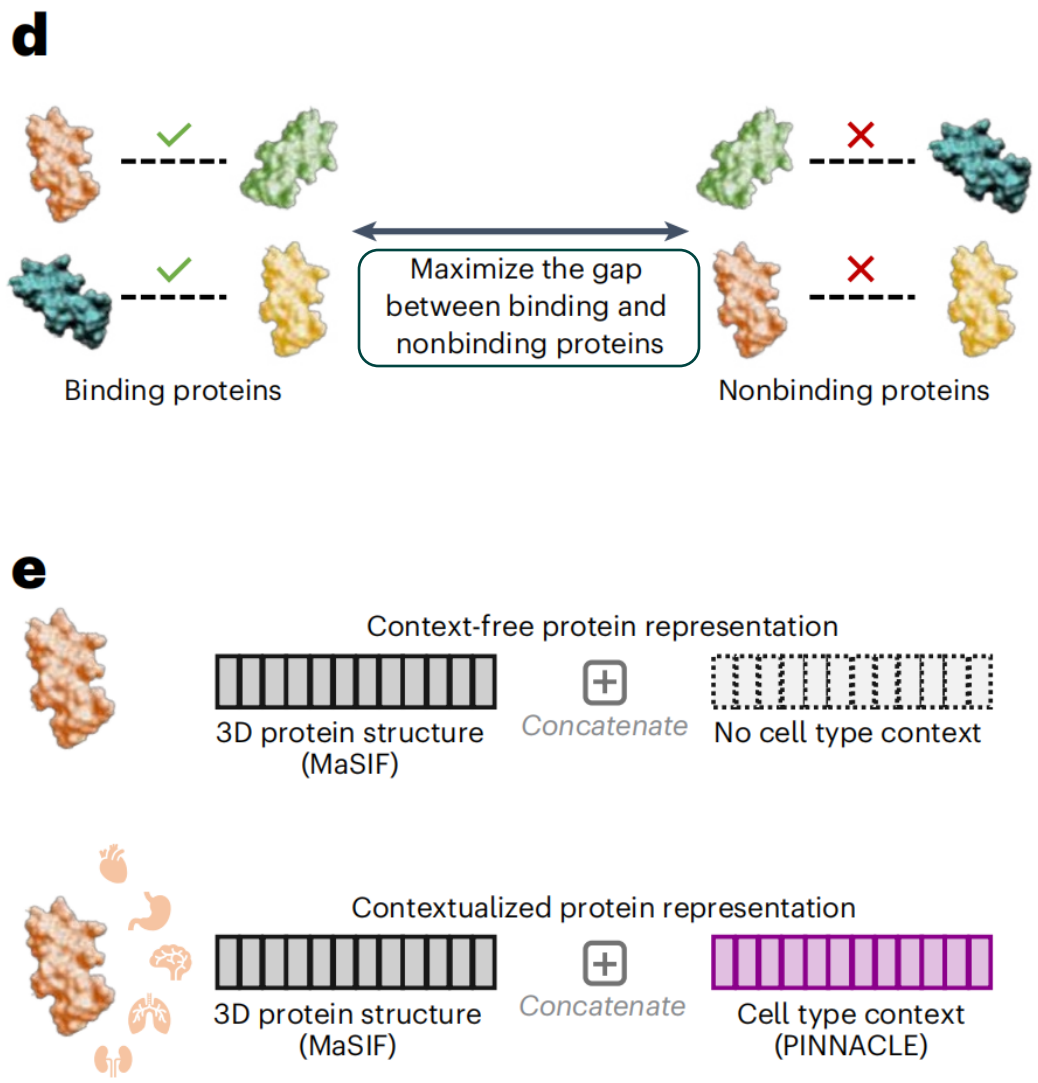
KS two-sample KS test P-value, lower is better (from two distribution)

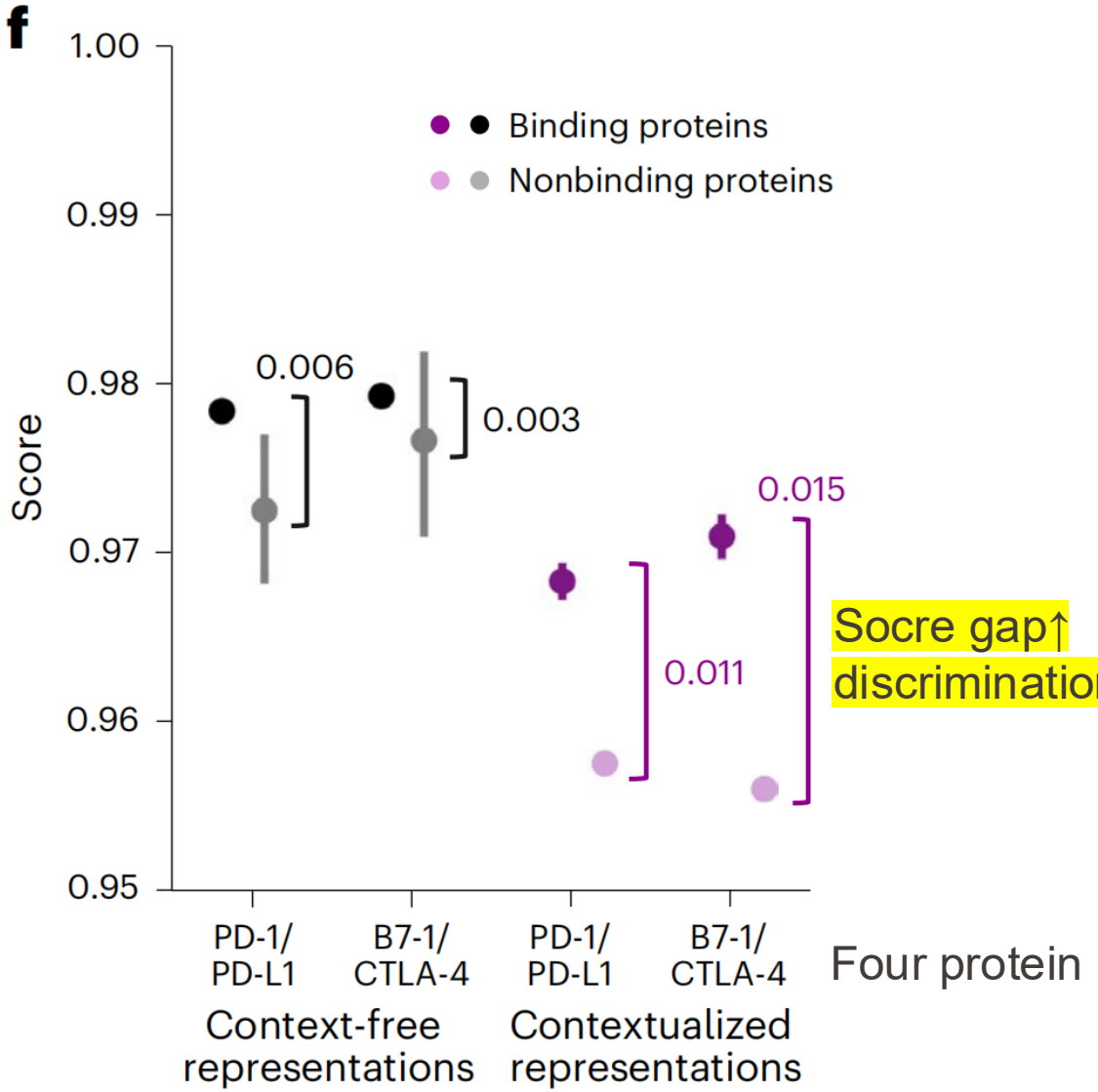
Context specific v.s. Context-free

Context-free PINNACLE: remove cell type and tissue network and all cell type- and tissue-related components of PINNACLE's architecture and objective function



PINNACLE enhances 3D structural representations of PPIs





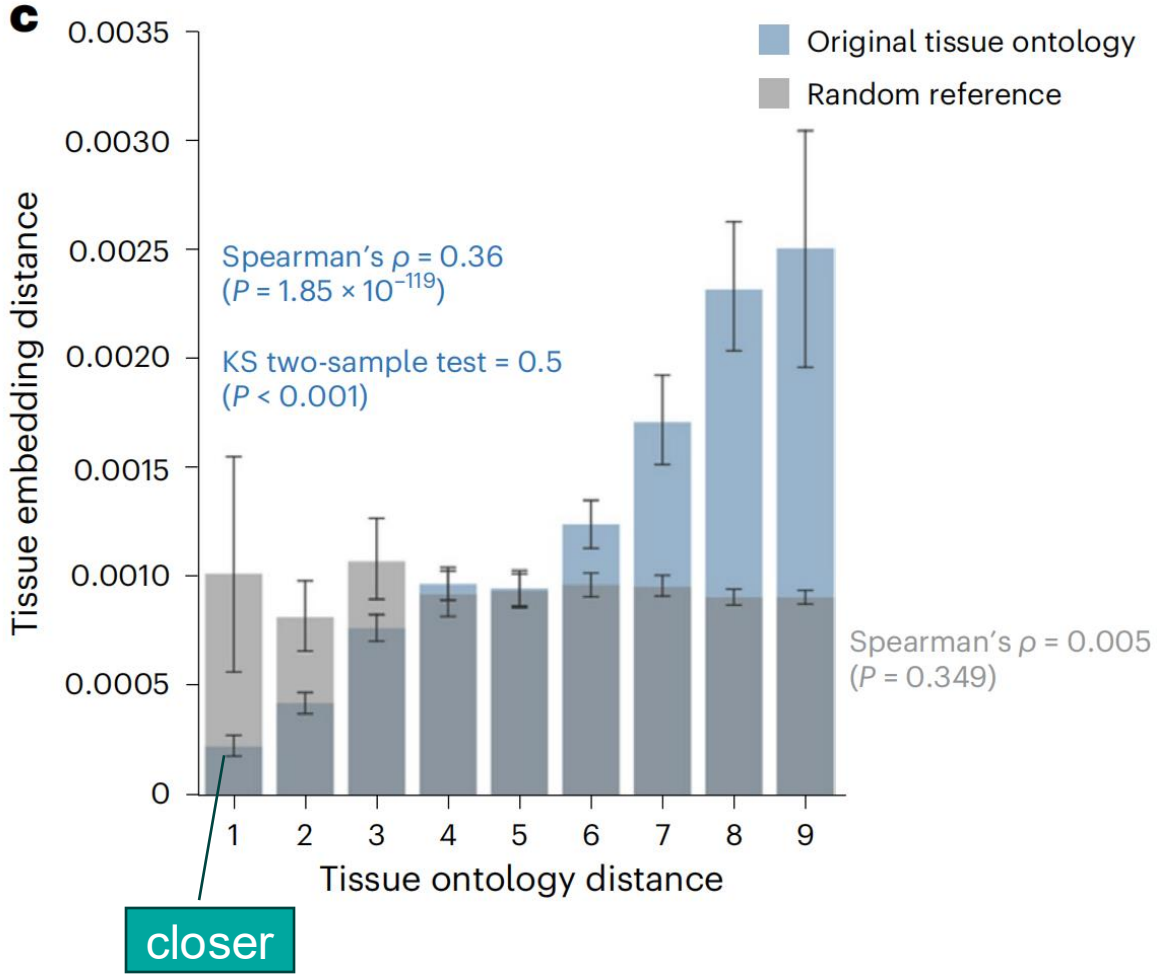
Socre gap↑
discrimination ↑

Context information increases the discrimination of the protein representations

Tissue embedding evaluation

Tissues discrimination ↑

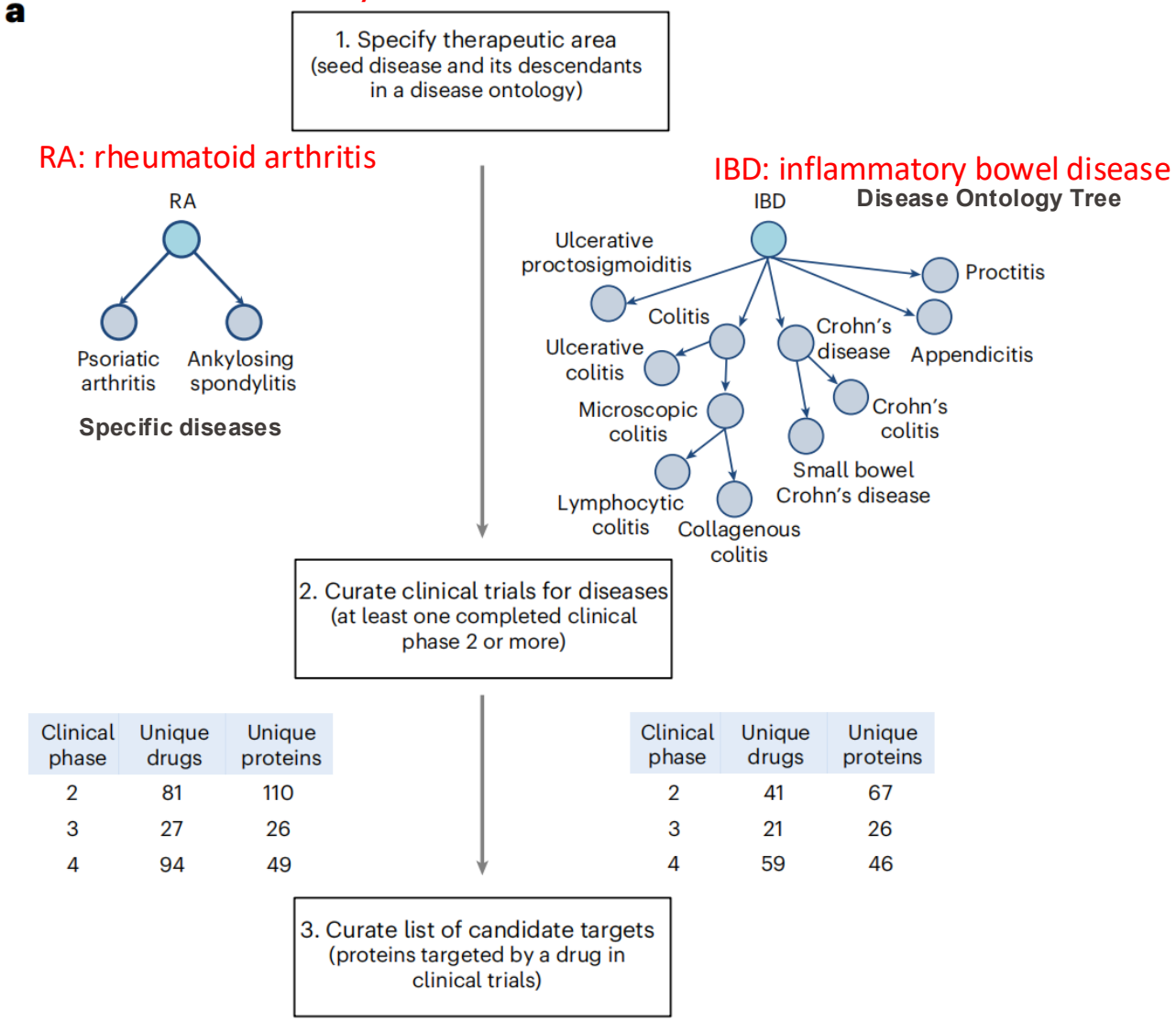
This validates PINNACLE's ability **to learn and encode the underlying biological structure of tissues**. The closer the tissues are in the real world (ontology), the closer they are in the model's embedding space.



As the **biological distance (X-axis) increases**, the **computational distance (Y-axis) also increases**.

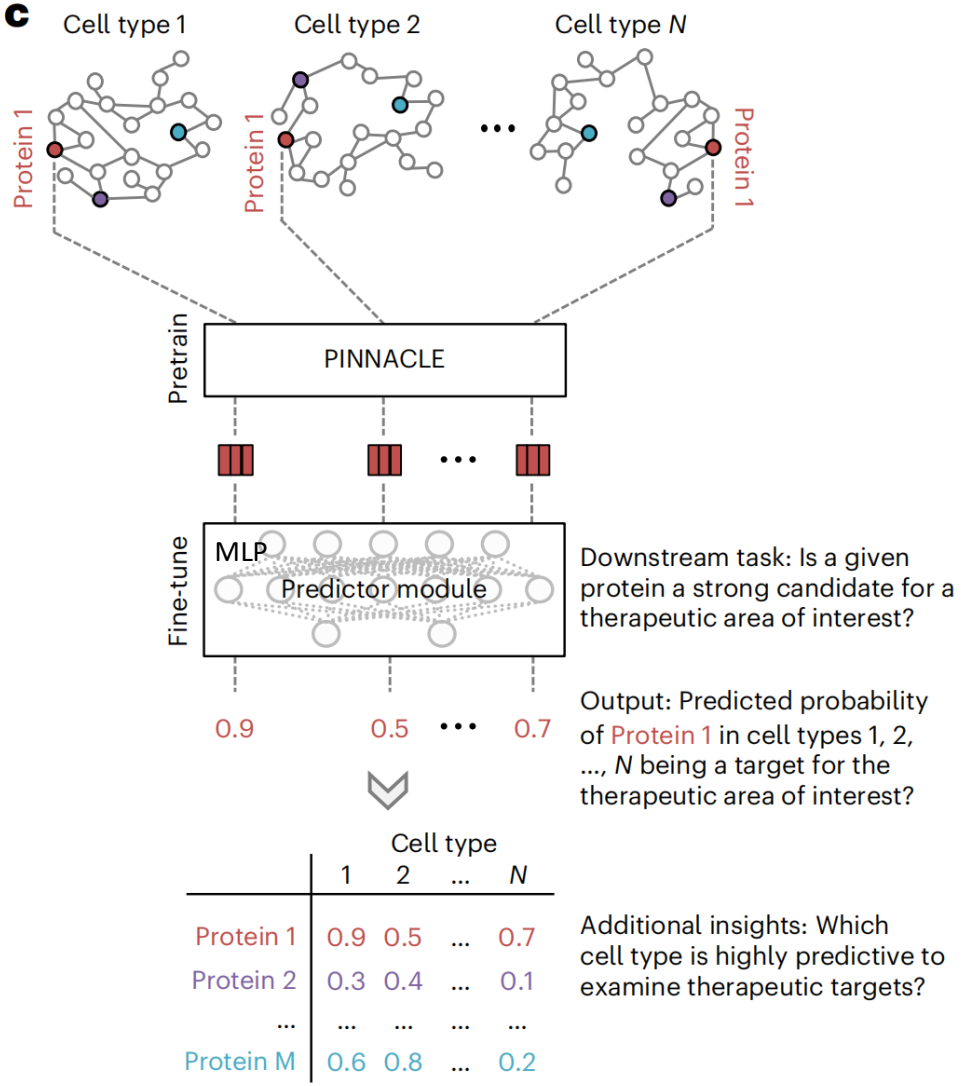
Therapeutic target prioritization (via fine-tuning contextualized protein representations)

The predictor module (that is, MLP) fine-tunes the (pretrained) contextualized protein representations for predicting whether a given protein is a strong candidate for the therapeutic area of interest. (RA: rheumatoid arthritis + IBD: inflammatory bowel disease)



Clinical phase	Unique drugs	Unique proteins
2	81	110
3	27	26
4	94	49

Clinical phase	Unique drugs	Unique proteins
2	41	67
3	21	26
4	59	46



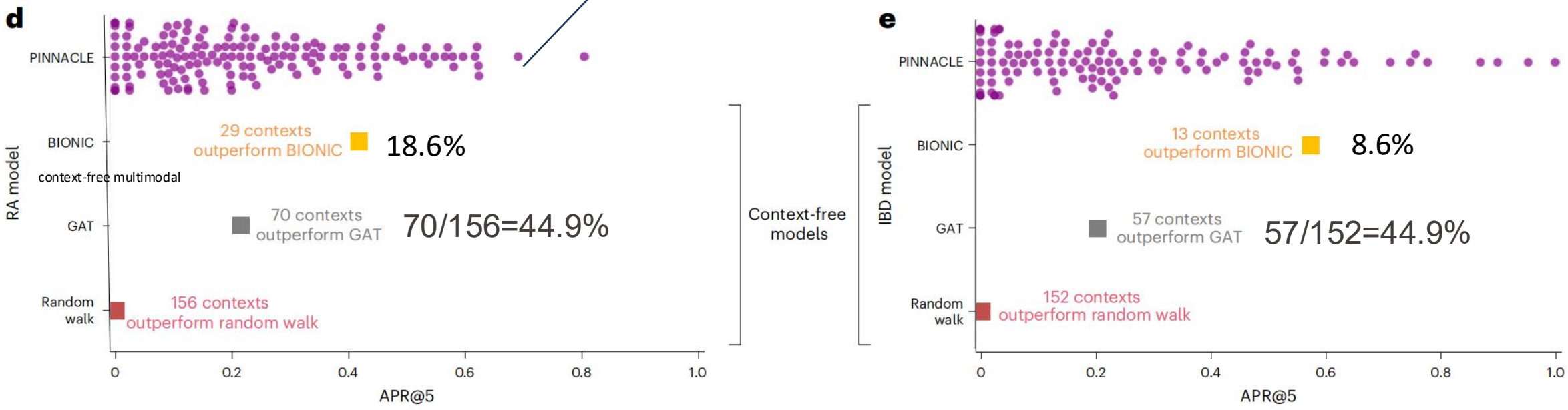
	Cell type 1	2	...	N
Protein 1	0.9	0.5	...	0.7
Protein 2	0.3	0.4	...	0.1
...	...	...	...	...
Protein M	0.6	0.8	...	0.2

Contextual models outperform context-free target prediction

Benchmarking of context-aware and context-free approaches for RA (**d**) and IBD (**e**) therapeutic areas.

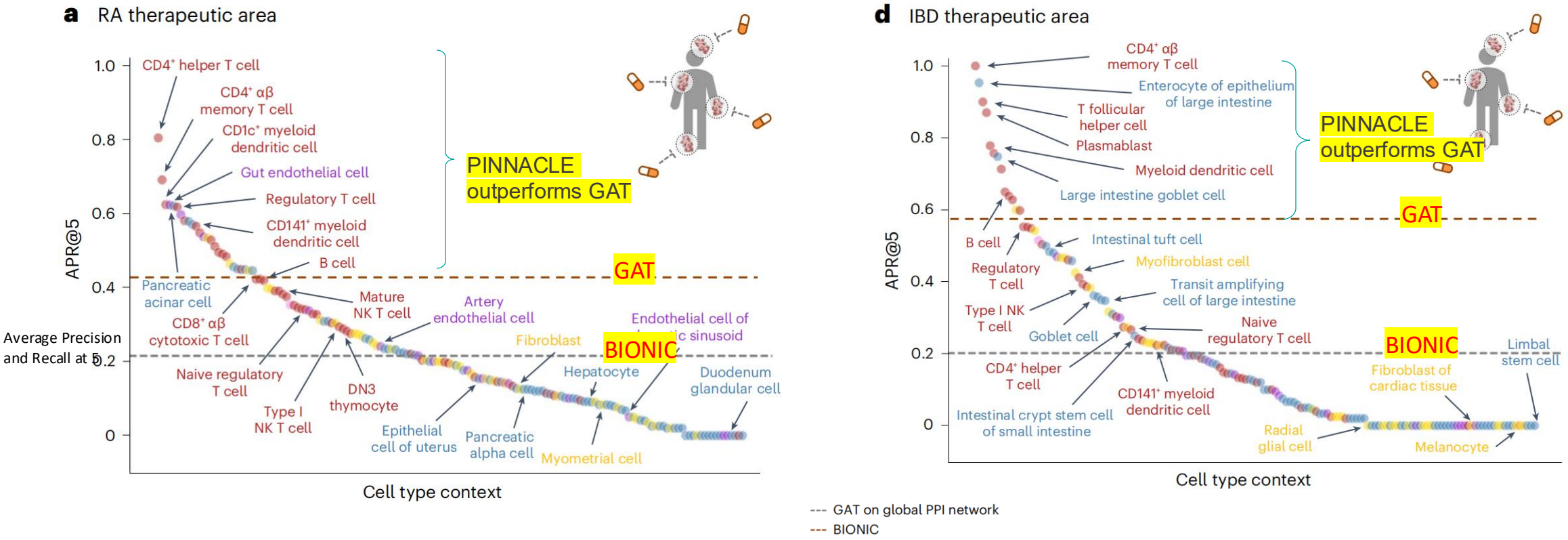
Baselines: Random work, GAT, BIONIC (context free model)

Dot: performance (averaged across ten random seeds) of protein representations from a given context



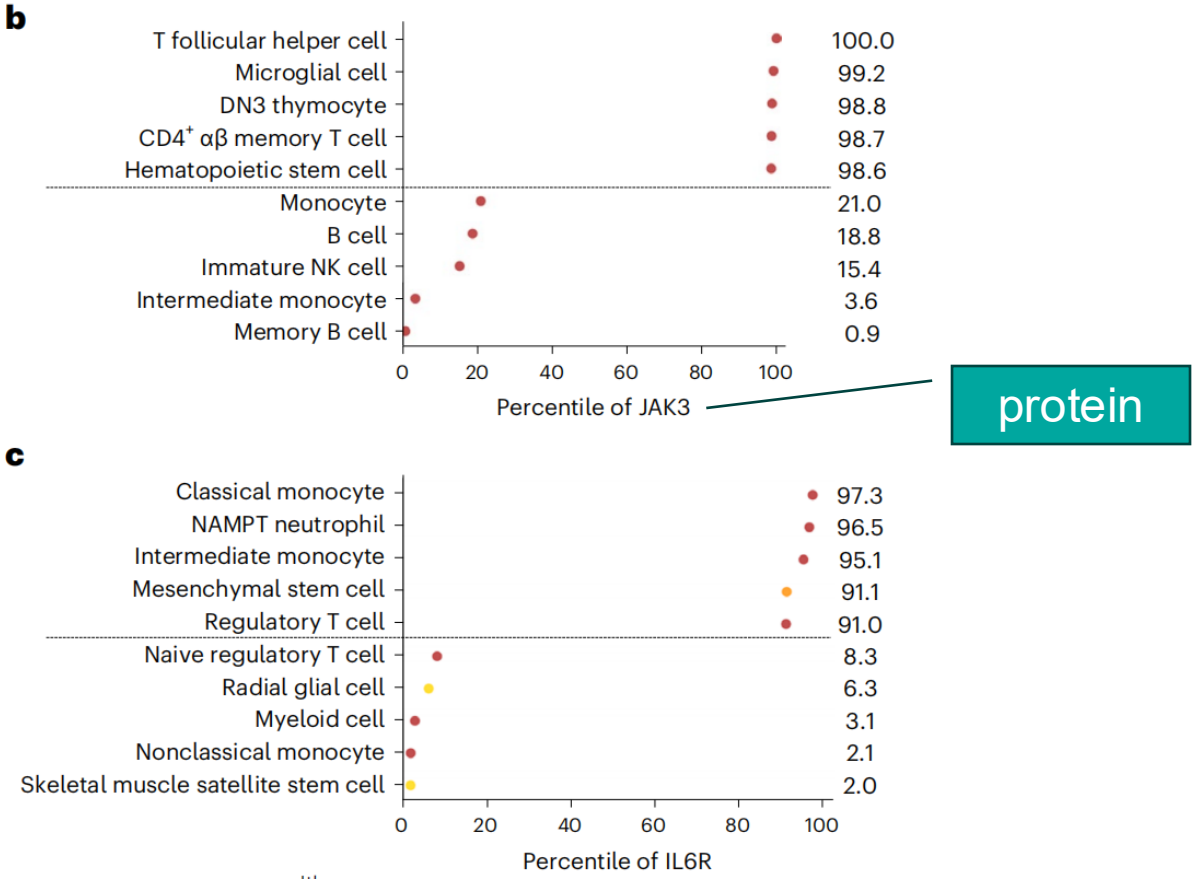
Performance of contextualized target prioritization for RA and IBD therapeutic areas

PINNACLE can nominate targets across cell type contexts



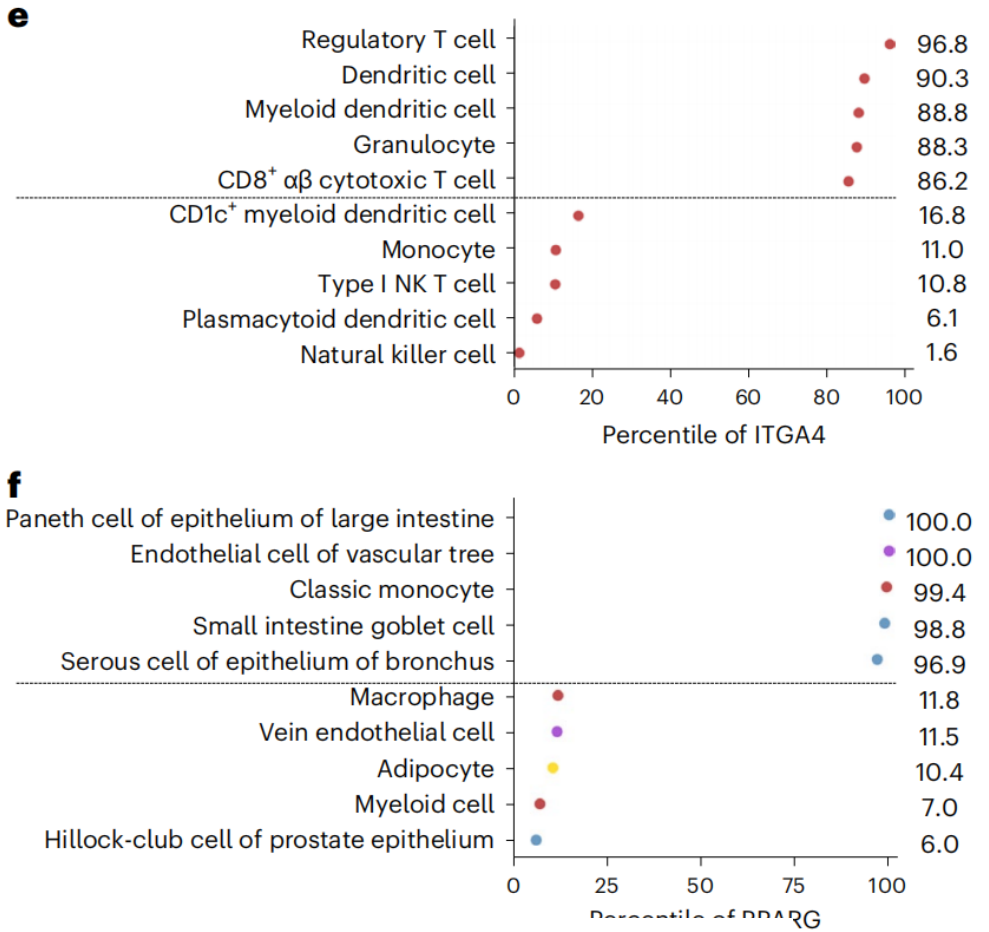
Cell type contexts reflect MoAs (mechanisms of actions)

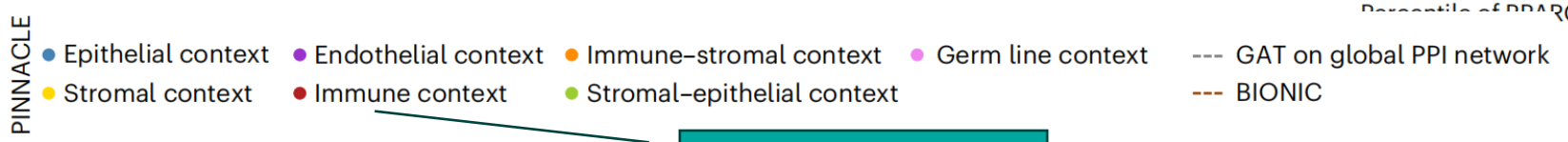
- **Predictive cell type contexts reflect MoAs in RA therapies**



protein

- **Predictive cell type contexts elucidate MoAs in IBD therapies**





compartment

Summary: Positive & Negative

PINNACLE is a protein representation learning method that use the context learning for enhancing the context-specific property of the presentations

👍 Positives:

- **Sophisticated Modeling**: Utilizes GNNs to capture complex topological interactions within the static PPI network backbone.
- **High-Resolution Targeting**: Achieves cell-type specificity in protein representations, moving beyond "average" network views using *single cell atlases* context.

⚠ Limitations:

- **Inter-Cell Signaling**: Long-range connections (protein interactions between different cell types) are currently only modeled implicitly through the Metagraph structure, simplifying direct molecular dialogue.

EVENT

Thank you for your attention !

PRESEDENTED BY

Jiying Zhang

École polytechnique fédérale de Lausanne