# CAN PROTEIN-DERIVED KNOWLEDGE IMPROVE PATHOLOGY FOUNDATION MODELS?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Molecularly guided pathology foundation models (PFMs) exploit transcriptomic or proteomic information to enrich whole-slide image (WSI) representations, yet effectively leveraging large standalone molecular corpora remains challenging. First, existing molecular foundation models encode protein sequences or single-cell states, not the patient-level bulk expression profiles paired with WSIs. Second, because cross-modal supervision is restricted to paired WSI–omics samples, knowledge from standalone molecular corpora reaches the pathology encoder only indirectly, creating a paired-support bottleneck. To address these challenges, we propose a three-stage framework that decouples proteomic knowledge acquisition from cross-modal transfer, yielding ProSlide, a slide-level hierarchical pathology foundation model. First, to close the modality gap, we pretrain a Proteomic Foundation Encoder (PFE) on 12,695 sample-level bulk protein profiles using virtual profile generation and expression-space multi-view pretraining. Second, we pretrain ProSlide, a patch–region–slide encoder, to predict protein expression from paired WSI–protein samples. Third, to relax the paired-support bottleneck, we introduce Prot2Path, a cross-modal relational distillation objective. For each paired sample, it aligns the similarity distributions of the WSI and its protein profile over a shared, frozen bank of PFE-encoded paired and standalone profiles. We evaluate ProSlide on 12 downstream tasks across breast, lung, and renal cancers. Despite being pretrained with only 2,229 WSIs and 12,695 sample-level protein profiles, ProSlide achieves the highest mean accuracy and AUC within each cancer group.

## 1 INTRODUCTION

Foundation models have advanced computational pathology by learning transferable representations from large-scale histopathology data (Gao et al., 2026; Chen et al., 2024), supporting diverse diagnostic (Gao et al., 2025), prognostic (Yang et al., 2026), and molecular prediction tasks (Zhang et al., 2026b). More recently, pathology foundation models (PFMs) have extended morphology-driven pretraining by incorporating complementary molecular supervision (Xu et al., 2025). Molecular profiles provide an additional view of disease biology, offering supervision that connects tissue morphology with underlying molecular characteristics. Multimodal frameworks such as Threads (Vaidya et al., 2025) and CARE (Zhang et al., 2026a) exemplify this direction by integrating histopathological and molecular information to enrich whole-slide image (WSI) representations. These advances highlight the value of biological information for pathology representation learning.

However, finding the optimal strategy to exploit such molecular knowledge for pathology pretraining is challenging. Existing molecularly guided PFMs largely rely on paired WSI–omics cohorts for cross-modal learning (Vaidya et al., 2025; Zhang et al., 2026a). As shown in Fig. 1(a), the biological units behind typical molecular encoders are much smaller in physical scale than a WSI. For example, protein language models (Lin et al., 2022) primarily encode protein identity and sequence-derived properties, and single-cell models (Cui et al., 2023) characterize cellular expression states. In contrast, the proteomic measurements associated with WSIs normally represent sample-level bulk protein abundance patterns. This motivates learning a sample-level proteomic representation space aligned with the molecular states used for pathology supervision.

A second challenge arises when transferring such molecular knowledge to pathology. Even when a molecular encoder is pretrained on a substantially larger collection of standalone profiles, conven-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

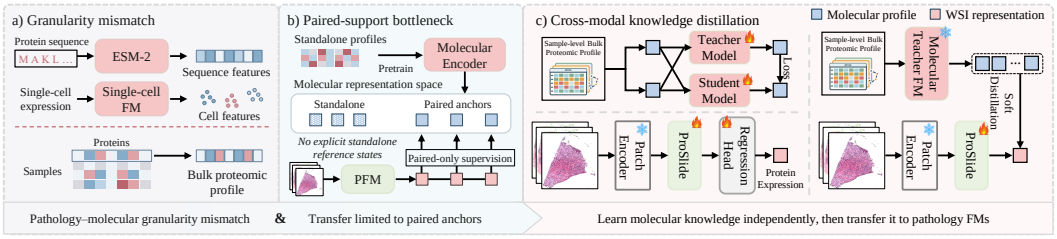



Figure 1: (a) Granularity mismatch between existing molecular encoders and sample-level proteomic states. (b) Cross-modal transfer is restricted to paired WSI–omics anchors. (c) Our framework leverages standalone molecular profiles through cross-modal knowledge distillation.

tional cross-modal learning typically queries this representation space only at molecular samples with corresponding WSIs (Zhang et al., 2026a). As illustrated in Fig. 1(b), standalone profiles can shape the molecular encoder indirectly during pretraining, but they do not explicitly participate in the pathology training objective. Consequently, the pathology encoder is supervised only at a limited set of paired molecular anchors and is not directly constrained to preserve its relationships to molecular states outside the paired cohort. This creates a paired-support bottleneck between large-scale molecular representation learning and cross-modal pathology training.

To address these limitations, we introduce ProSlide, a hierarchical pathology encoder trained through a three-stage framework that separates proteomic knowledge acquisition from cross-modal transfer (Fig. 1(c)). First, we pretrain a Proteomic Foundation Encoder (PFE) on sample-level bulk protein profiles using expression-space multi-view pretraining and virtual profile generation. This stage establishes a molecular representation space by learning from proteomic cohorts independently of WSI pairing. Second, we pretrain ProSlide to predict protein expression from paired WSI–protein samples, using region-level attention to aggregate patch features into slide representations. Third, we introduce Prot2Path to transfer PFE-defined inter-sample relations to ProSlide, complementing the direct supervision provided by protein-expression prediction. Prot2Path constructs a frozen bank of PFE embeddings from both paired and standalone protein profiles as a shared molecular reference. Standalone profiles thus contribute through two routes: shaping the teacher's representation space during PFE pretraining and providing molecular references during cross-modal distillation. We evaluate PFE against a BERT-style pretrained counterpart through profile retrieval under partial protein masking and downstream pathology distillation. Our default PFE configuration achieves higher retrieval accuracy and better downstream performance, supporting its utility as a proteomic teacher for pathology representation learning. We further evaluate ProSlide on 12 tasks, with four each for breast, lung, and renal cancers. Despite being pretrained with only 2,229 WSIs and 12,695 sample-level protein profiles, ProSlide achieves the highest mean accuracy and AUC within each cancer group. Our main contributions are summarized as follows:

**(1) Independent proteomic representation learning.** We introduce **PFE**, which learns sample-level proteomic representations through expression-space multi-view pretraining and virtual profile generation without requiring paired WSIs.

**(2) Cross-modal relational distillation.** We develop **ProSlide**, a protein-supervised hierarchical pathology encoder, and train it with **Prot2Path** to match PFE-defined similarity distributions over a shared bank of paired and standalone protein profiles.

**(3) Effective learning under limited pairing.** Across 12 downstream tasks, ProSlide achieves the highest mean performance within each of three cancer groups using limited paired training data.

## 2 Related Work

**Molecularly Guided Pathology Representation Learning.** Pathology foundation models learn transferable representations from large histology collections (Chen et al., 2024), while hierarchical encoders such as HIPT aggregate patch features into slide representations (Chen et al., 2022). Molecularly guided approaches extend this paradigm using paired WSI–omics data. Threads in-

corporates genomic and transcriptomic supervision (Vaidya et al., 2025), and CARE uses RNA and protein profiles to refine adaptive region representations (Zhang et al., 2026a). Our framework complements these approaches by independently pretraining a proteomic encoder on standalone data before transferring its knowledge to ProSlide through paired samples.

**Protein and Proteomic Representation Learning.** Protein language models such as ESM-2 learn sequence representations that support structure prediction (Lin et al., 2022). At the expression level, scPROTEIN uses graph contrastive learning to generate single-cell proteomic embeddings (Li et al., 2024). PFE targets bulk protein expression profiles, where each sample contains continuous measurements of multiple identified proteins. Building on multi-view self-supervised learning (Oquab et al., 2023), PFE combines virtual profile generation with identity-preserving expression perturbations to learn proteomic representations for subsequent pathology distillation.

**Knowledge Distillation.** Knowledge distillation transfers teacher predictions or representations to a student, while relational variants transfer relationships among samples (Park et al., 2019). In pathology, Pathryoshka consolidates multiple pretrained pathology teachers into a compact model (Grashei et al., 2026). Cross-modal approaches such as MoMKD transfer genomic supervision through a momentum-updated memory for histology-only inference (Guo et al., 2026). Prot2Path uses an independently pretrained proteomic teacher and a frozen bank containing paired and standalone profiles. It transfers proteomic relations by aligning protein- and WSI-derived similarity distributions over the same sampled references.

# 3 METHOD

Our framework uses standalone proteomic data for both proteomic representation learning and cross-modal knowledge transfer through three training stages (Fig. 2). First, we pretrain a Proteomic Foundation Encoder (PFE) on sample-level protein profiles independently of WSI pairing. Second, we pretrain ProSlide to predict protein expression from paired WSI–protein samples. Third, Prot2Path transfers inter-sample relations from the frozen PFE to ProSlide through a shared representation bank containing both paired and standalone protein profiles.

## 3.1 STAGE I: PROTEOMIC FOUNDATION ENCODER PRETRAINING

To support subsequent cross-modal transfer, we establish a proteomic representation space that captures variation across sample-level bulk protein profiles. PFE combines protein identity with continuous expression values to represent each profile. We train the encoder using expression-space multi-view pretraining, with complementary objectives operating at the profile and protein levels.

### 3.1.1 PFE ARCHITECTURE

We curate RPPA-based bulk protein expression profiles from four resources, TCGA (Akbani et al., 2014), TCPA (Li et al., 2013), CCLE (Ghandi et al., 2019), and CPPA (Zhao et al., 2020), covering sample tumor specimens, baseline cancer cell lines, and pharmacologically perturbed cell-line samples (Appendix A). Each profile is represented as

$$\mathcal{X}_i = \{(p_j, x_{i,j}, m_{i,j})\}_{j=1}^{M}, \tag{1}$$

where $i$ indexes samples and $j$ indexes proteins. Here, $p_j$ denotes protein identity, $x_{i,j} \in \mathbb{R}$ is the normalized expression value, and $m_{i,j} \in \{0, 1\}$ indicates whether the measurement is observed. PFE constructs each protein token using a learnable identity embedding $e_j^{\mathrm{id}} \in \mathbb{R}^d$ and a shared value encoder $g_{\mathrm{val}} : \mathbb{R} \to \mathbb{R}^d$:

$$e_{i,j} = e_j^{\mathrm{id}} + m_{i,j} g_{\mathrm{val}}(x_{i,j}) + (1 - m_{i,j}) e^{\mathrm{mask}}, \tag{2}$$

where $e^{\mathrm{mask}} \in \mathbb{R}^d$ is a shared learnable embedding for missing measurements.

An $L$-layer Transformer processes these tokens with a prepended learnable `[CLS]` token, without positional embeddings. Its outputs $h_{i,0}$ and $\{h_{i,j}\}_{j=1}^{M}$ provide the profile-level and contextualized protein-level representations, respectively.

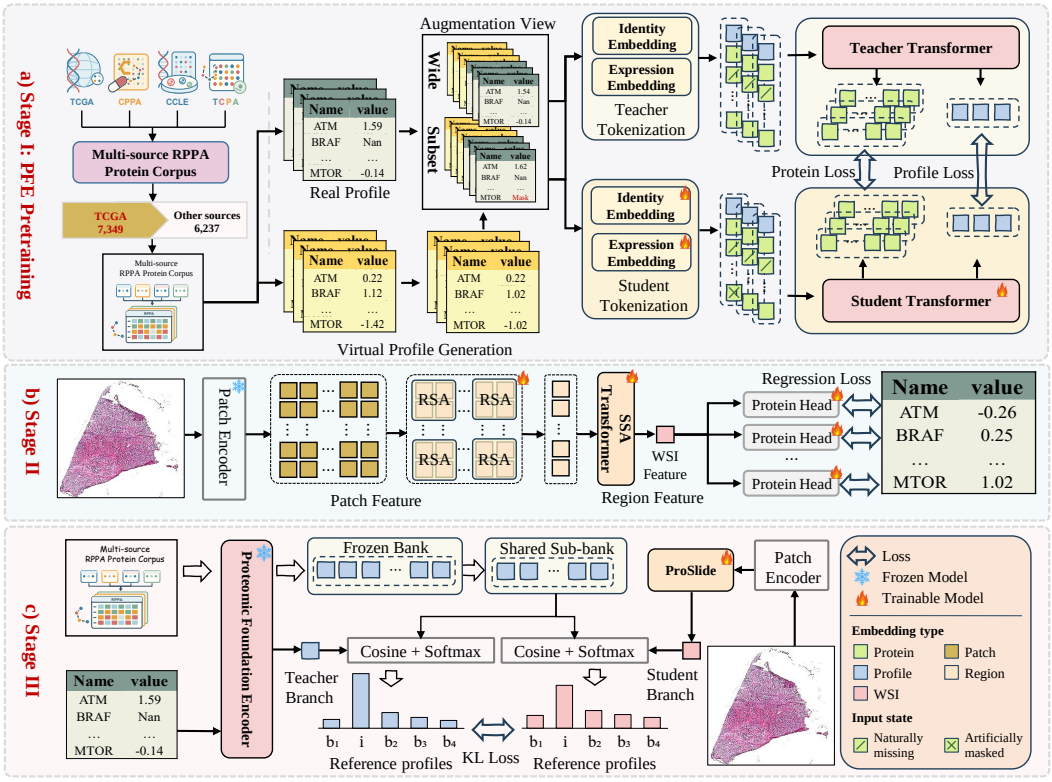



Figure 2: **Overview of the three-stage framework.** (a) PFE learns proteomic representations through virtual profile generation and expression-space multi-view pretraining. (b) ProSlide learns hierarchical WSI representations through protein-expression prediction. (c) Prot2Path distills proteomic relational knowledge into ProSlide through similarity distribution alignment over a shared sampled protein bank.

### 3.1.2 Expression-Space Multi-View Pretraining

Inspired by multi-view self-supervised learning (Oquab et al., 2023), we construct proteomic views by augmenting expression values while preserving protein identities. Our strategy combines virtual profile generation to diversify training inputs with asymmetric teacher–student learning over profile-wide and subset-perturbed views.

**Virtual profile generation.** Each mini-batch contains $B_{\mathrm{r}}$ real profiles from $\mathcal{D} = \{\mathcal{X}_i\}_{i=1}^{N}$ and $B_{\mathrm{v}}$ virtual profiles generated online. For each virtual profile $a$, we independently sample a source profile $\mathcal{X}_{k_a}$ from the full corpus. We select a small subset $\mathcal{S}_a \subseteq \{j : m_{k_a,j} = 1\}$ of observed proteins as candidates for expression perturbation:

$$\widetilde{\mathcal{X}}_a = \{(p_j, \tilde{x}_{a,j}, m_{k_a,j})\}_{j=1}^{M}, \qquad \tilde{x}_{a,j} = \begin{cases} \mathcal{T}_{a,j}(x_{k_a,j}), & j \in \mathcal{S}_a, \\ x_{k_a,j}, & \text{otherwise.} \end{cases} \tag{3}$$

Here, $\mathcal{T}_{a,j}$ denotes the action on protein $j$ of the augmentation sampled for virtual profile $a$. We sample one of three augmentation types: *partial posterization* reduces expression precision, *partial Gaussian perturbation* adds independent noise, and *partial mixup* interpolates values with another profile. Mixup modifies only proteins observed in both profiles, using one mixing coefficient shared across the interpolated coordinates. Each virtual-profile augmentation modifies only a small subset of observed proteins, preserving the source profile's protein identities, missingness pattern, and unselected expression values while introducing additional local expression configurations for pretraining. Operator definitions are provided in Appendix B.1.

**Expression-space multi-view construction.** For each real or virtual profile, we construct profile-wide and subset-perturbed views. Both retain all $M$ protein identities and preserve naturally missing measurements.

*Profile-wide views.* We construct three views for each profile: one clean view and two stochastically augmented views. The augmented views apply mild posterization to all observed expression values, introducing variations in numerical precision while retaining the complete protein context.

*Subset-perturbed views.* Additional views apply Gaussian noise and random masking to independently sampled subsets of observed proteins. Random masking hides expression values while retaining protein identity tokens, and unselected expression values remain unchanged.

The resulting views are used to train PFE with complementary profile- and protein-level self-supervised objectives following DINOv2. View assignment, objective definitions, and training configurations are provided in Appendix B.

### 3.2 Stage II: Protein-Supervised ProSlide Pretraining

ProSlide adopts a patch-to-region-to-slide hierarchy, operating on pre-extracted features from the frozen CONCH v1.5 patch encoder (Lu et al., 2024). Its region and slide Transformers are randomly initialized and jointly trained to predict protein expression from paired WSI–protein data. This stage provides a molecularly informed initialization for subsequent Prot2Path distillation.

**Hierarchical slide encoder.** For slide $i$, the input is represented as $\mathcal{W}_i = \{(h_{i,n}, u_{i,n})\}_{n=1}^{N_i}$, where $N_i$ is the number of patches, $h_{i,n} \in \mathbb{R}^{768}$ denotes the feature of patch $n$, and $u_{i,n} \in \mathbb{R}^2$ denotes the patch-grid coordinate of patch $n$ in the WSI. We group patches into $R_i$ non-overlapping spatial regions, with $\mathcal{I}_{i,k}$ denoting the patch indices in region $k$. Within each region, a region self-attention (RSA) module aggregates patch features into a region representation:

$$r_{i,k} = \mathrm{RSA}\left(c^r, \{h_{i,n}\}_{n \in \mathcal{I}_{i,k}}\right), \tag{4}$$

where $c^r$ is a learnable region `[CLS]` token. RSA projects the patch features into the region embedding space, prepends $c^r$, and processes the resulting sequence with a region Transformer. The output corresponding to $c^r$ serves as the region representation $r_{i,k}$. The region representations are further aggregated by a slide self-attention (SSA) module to obtain the WSI representation:

$$z_i^{\mathrm{slide}} = \mathrm{SSA}\left(c^s, \{r_{i,k}\}_{k=1}^{R_i}\right), \tag{5}$$

where $c^s$ is a learnable slide `[CLS]` token. SSA projects the region representations into the slide embedding space, prepends $c^s$, and models interactions across regions using a slide Transformer. The output corresponding to $c^s$ serves as the slide representation $z_i^{\mathrm{slide}} \in \mathbb{R}^{768}$. Both RSA and SSA are implemented as Transformer encoders with multi-head self-attention. RSA models interactions among patches within each region, whereas SSA models interactions among regions across the slide. Region partitioning and architectural configurations are detailed in the Appendix C.1.

**Protein-expression supervision.** A protein-specific regression head $g_j$ predicts the expression of protein $j$ as $\widehat{y}_{i,j} = g_j(z_i^{\mathrm{slide}})$. We minimize masked mean-squared error:

$$\mathcal{L}_{\mathrm{Pro}} = \frac{\sum_{i \in \mathcal{B}} \sum_j m_{i,j} \left(\widehat{y}_{i,j} - y_{i,j}\right)^2}{\sum_{i \in \mathcal{B}} \sum_j m_{i,j}}, \tag{6}$$

where $\mathcal{B}$ is a mini-batch of paired samples, $j$ indexes the supervised protein panel, $y_{i,j}$ is the measured expression value, and $m_{i,j}$ indicates whether that measurement is available. This objective trains ProSlide to capture morphological features associated with individual protein expression levels. After pretraining, the regression heads are removed, and the ProSlide encoder provides the initialization for Prot2Path distillation in Stage III.

### 3.3 Stage III: Prot2Path Cross-Modal Knowledge Distillation

To further incorporate the proteomic structure learned by PFE, we introduce Prot2Path, which aligns the similarity distributions of paired protein and WSI representations over a shared proteomic representation bank. This relational objective transfers PFE-defined neighborhoods to ProSlide and allows standalone protein profiles to serve as molecular references.

**Shared proteomic representation bank.** We freeze PFE and use it to encode a reference set $\mathcal{D}_{\text{bank}} = \{\mathcal{X}_b\}_{b=1}^{K}$ containing both paired and standalone protein profiles. For each reference profile, we obtain its normalized PFE representation as

$$t_b = \text{PFE}(\mathcal{X}_b), \qquad \widehat{t}_b = \frac{t_b}{\|t_b\|_2}. \tag{7}$$

The resulting representations form a fixed proteomic representation bank

$$T = \left[\widehat{t}_1; \widehat{t}_2; \ldots; \widehat{t}_K\right] \in \mathbb{R}^{K \times d}, \tag{8}$$

where $d$ is the PFE embedding dimension. The bank is computed once and remains frozen throughout distillation.

For each paired sample $(\mathcal{W}_i, \mathcal{X}_i)$, we construct a reference index set $\mathcal{J}_i$ containing all paired training profiles and randomly sampled standalone profiles, with $|\mathcal{J}_i| = K_{\text{sub}}$. The sampled sub-bank is defined as $T_{\mathcal{J}_i} = \left[\widehat{t}_b\right]_{b \in \mathcal{J}_i} \in \mathbb{R}^{K_{\text{sub}} \times d}$. The index set $\mathcal{J}_i$ is resampled across epochs, allowing the same paired sample to be compared with different proteomic references during training. PFE defines the proteomic relational target over the sampled sub-bank:

$$\pi_i^{\text{prot}} = \text{Softmax}\left(\frac{\widehat{t}_i T_{\mathcal{J}_i}^{\top}}{\tau_{\text{t}}}\right). \tag{9}$$

where $\tau_{\text{t}}$ is the teacher temperature. Each element of $\pi_i^{\text{prot}}$ measures the relative proteomic similarity between the paired profile $\mathcal{X}_i$ and one reference profile in the bank. Consequently, profiles without corresponding WSIs can still participate in defining the relational target.

**Cross-modal relational distillation.** Given the paired WSI $\mathcal{W}_i$, ProSlide produces the slide representation $z_i^{\text{slide}}$. An alignment head $g_\phi$ maps this representation into the PFE embedding space:

$$s_i = g_\phi\left(z_i^{\text{slide}}\right), \qquad \widehat{s}_i = \frac{s_i}{\|s_i\|_2}. \tag{10}$$

Using the same sampled sub-bank, we construct the pathology-derived relational distribution as

$$\pi_i^{\text{path}} = \text{Softmax}\left(\frac{\widehat{s}_i T_{\mathcal{J}_i}^{\top}}{\tau_{\text{s}}}\right). \tag{11}$$

where $\tau_{\text{s}}$ is the student temperature. Because $\pi_i^{\text{prot}}$ and $\pi_i^{\text{path}}$ are defined over the same reference profiles, their corresponding elements describe relations to identical bank entries.

**Knowledge distillation loss.** Prot2Path minimizes the divergence between the proteomic and pathology-derived distributions:

$$\mathcal{L}_{\text{Prot2Path}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} D_{\text{KL}}\left(\pi_i^{\text{prot}} \,\|\, \pi_i^{\text{path}}\right), \tag{12}$$

where $\mathcal{B}$ denotes a mini-batch of paired samples. This objective trains ProSlide to reproduce the PFE-defined proteomic neighborhood of each paired sample. The graded target probabilities preserve similarities to both paired and standalone reference profiles.

During distillation, PFE and the representation bank remain frozen, while the alignment head and ProSlide are optimized. After training, PFE, the bank, and the alignment head are discarded. The distilled ProSlide encoder is retained for WSI-only downstream inference. Additional details of the Prot2Path reference bank and distillation objective are provided in the Appendix C.2.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets and downstream tasks.** We evaluate our models on 12 WSI-level classification tasks drawn from eight cancer cohorts. These include four CPTAC cohorts (Edwards et al., 2015),

Table 1: Logistic-regression results averaged within the selected renal cell carcinoma (RCC), breast, and lung task groups. Each entry reports the mean of the four task-level means ± the arithmetic mean of their reported standard deviations.

| Model | Renal | | Breast | | Lung | |
|---|---|---|---|---|---|---|
| | ACC | AUC | ACC | AUC | ACC | AUC |
| mean-pooling | 57.82±0.25 | 72.16±0.80 | 67.74±0.87 | 75.70±1.01 | 59.61±0.20 | 75.31±0.20 |
| FEATHER | 54.95±0.44 | 70.96±1.01 | 66.79±0.92 | 74.80±0.82 | 60.22±0.22 | 74.78±0.26 |
| GigaPath | 56.04±0.25 | 75.33±0.72 | 67.19±0.84 | 77.50±1.08 | 58.93±0.35 | 73.43±0.27 |
| PRISM | 56.28±0.36 | 72.72±0.73 | 66.39±0.88 | 74.88±0.90 | 61.83±0.16 | 74.86±0.17 |
| CARE | 58.43±0.31 | 75.24±0.69 | 67.63±0.67 | 75.07±0.83 | 61.94±0.20 | 76.99±0.16 |
| TITAN | 58.67±0.45 | 75.74±0.54 | 68.84±0.66 | 78.19±0.72 | 62.18±0.26 | 76.15±0.19 |
| ProSlide | **60.85±0.46** | **76.67±0.61** | **70.02±0.78** | **78.53±0.87** | **63.17±0.21** | **77.90±0.24** |

Table 2: Comparison of protein models on internal and external masked-protein retrieval tasks. The internal results are evaluated on 891 held-out TCGA samples, while the external results are evaluated using the combined 275-patient CPTAC gallery. All retrieval results are reported as the mean R@1 (%) over five runs.

| Model | Mask Results | | | External Mask Results | | | Distillation Results | |
|---|---|---|---|---|---|---|---|---|
| | M1 R@1 | M2 R@1 | M3 R@1 | M1 R@1 | M2 R@1 | M3 R@1 | ACC | AUC |
| Protein-BERT | 97.62 | 94.25 | 91.40 | 96.22 | 92.15 | 84.87 | 62.64 | 75.99 |
| PFE & $B_v = 0$ | 97.42 | 93.89 | 89.90 | 95.35 | 91.71 | 83.35 | 63.21 | 76.94 |
| PFE & $B_v = 32$ | 98.47 | 96.36 | 93.76 | **97.09** | 94.55 | **88.51** | 64.57 | 77.13 |
| **PFE & $B_v = 64$** | **98.92** | **96.92** | **95.17** | 96.58 | **95.05** | 88.44 | **64.68** | 77.70 |
| PFE & $B_v = 128$ | 96.14 | 87.77 | 75.29 | 94.25 | 88.00 | 75.35 | 64.57 | **78.13** |

CPTAC-BRCA, CPTAC-LUAD, CPTAC-LSCC, and CPTAC-CCRCC, together with the BCNB (Xu et al., 2021), Local-LUNG, DHMC-LUNG (Wei et al., 2019), and MUT-HETRCC (Acosta et al., 2022) cohorts. The benchmark contains four tasks each for breast, lung, and renal cancers, covering molecular prediction, histological subtyping, and cross-cohort transfer. WSI pretraining in Stages II and III uses only TCGA data, whereas all downstream evaluations are conducted on non-TCGA cohorts. Detailed task definitions, cohort statistics, and split protocols are provided in Appendix D.1.

**Implementation details.** We evaluate frozen representation quality through linear probing, training a logistic regression classifier on fixed slide representations. For the within-cohort CPTAC tasks, we follow the train/test-only evaluation protocol of Vaidya et al. (2025) and use 50 repeated patient-level 80%/20% train/test splits without a separate validation set. For these tasks, logistic regression uses $L_2$ regularization ($C = 0.5$), an intercept term, and the L-BFGS solver, with a maximum of 10,000 optimization iterations. The remaining tasks use five patient-level repetitions with separate training, validation, and test partitions. Task-specific cohort compositions and splitting protocols are provided in Appendix D.1. Test partitions are reserved for final evaluation. All downstream experiments are conducted on a system equipped with a single NVIDIA RTX 4090 GPU. Code is available at the anonymous repository: `https://anonymous.4open.science/r/ICLR-ProSlide`.

**Evaluation metrics.** We evaluate all downstream tasks using balanced accuracy and the area under the receiver-operating-characteristic curve (AUROC). Balanced accuracy measures the mean recall across classes and is therefore robust to class imbalance. For multi-class tasks, we report macro-averaged one-versus-rest AUROC. All results are reported as mean ± standard deviation.

### 4.2 Main Results

To evaluate the quality and transferability of the learned slide representations, we compare ProSlide with mean pooling and five representative pathology foundation models: FEATHER (Shao et al., 2025), GigaPath (Xu et al., 2024), PRISM (Shaikovski et al., 2024), CARE (Zhang et al., 2026a),

Table 3: Ablation study of supervision and knowledge distillation. S2-Sub and S2-Prot denote Stage II subtype supervision and protein-expression supervision, respectively. S3-KD denotes Stage III knowledge distillation. Eff. rank denotes the effective rank of representations on the CPTAC datasets.

| S2-Sub | S2-Prot | S3-KD | Eff. rank | Renal | | Breast | | Lung | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | ACC | AUC | ACC | AUC | ACC | AUC |
| ✓ | ✗ | ✗ | 11.95 | 52.61 | 69.79 | 57.94 | 69.56 | 60.12 | 73.18 |
| ✗ | ✓ | ✗ | 232.92 | 58.07 | 75.86 | 69.22 | 76.69 | 61.34 | 75.39 |
| ✗ | ✗ | ✓ | 295.43 | 59.44 | 74.88 | 68.20 | 77.51 | 61.92 | 76.94 |
| ✓ | ✗ | ✓ | 12.12 | 52.32 | 70.26 | 58.34 | 72.48 | 61.11 | 74.41 |
| ✗ | ✓ | ✓ | 201.29 | **60.85** | **76.67** | **70.02** | **78.53** | **63.17** | **77.90** |

Table 4: **Ablation of Prot2Path components.** We evaluate relational targets, standalone references, and sub-bank construction.

| Method | Target | Reference | Renal | | Breast | | Lung | |
|---|---|---|---|---|---|---|---|---|
| | | | ACC | AUC | ACC | AUC | ACC | AUC |
| Direct alignment | Cosine | Paired only | 60.07 | **77.86** | 69.22 | **78.69** | 62.34 | 77.39 |
| Hard matching | CE | All profiles | 60.83 | 76.49 | 68.47 | 76.57 | 62.89 | 77.69 |
| Paired-bank KD | Soft KL | Paired only | 60.62 | 76.61 | 69.35 | 77.65 | 62.69 | 77.47 |
| Full-bank KD | Soft KL | All profiles | 60.33 | 76.90 | **70.14** | **78.69** | 62.60 | 77.77 |
| Prot2Path | Soft KL | Sub-bank | **60.85** | 76.67 | 70.02 | 78.53 | **63.17** | **77.90** |

and TITAN (Ding et al., 2025). All models are evaluated using frozen slide representations under the same logistic-regression protocol. Table 1 summarizes the results averaged over four tasks each for renal, breast, and lung cancers, with complete task-level results provided in Appendix Table S10. ProSlide achieves the highest mean ACC and AUC across all three cancer groups. It performs consistently well across renal mutation-prediction tasks, while the relative strengths of different models vary across breast and lung molecular endpoints. ProSlide also performs well on pooled and cross-cohort fine-grained lung subtyping, supporting the utility of protein-informed representations for both molecular and morphological prediction. However, its competitive AUC and weaker ACC on cross-cohort binary lung subtyping indicate that transferability remains dependent on the endpoint and evaluation metric.

### 4.3 Ablation Studies

**PFE pretraining.** We evaluate proteomic encoders through masked-protein retrieval on held-out TCGA and external CPTAC profiles, together with their downstream distillation performance (Table 2). For $\mathrm{M}i$ R@1, we mask $i$ protein measurements in each query profile and compare its embedding against all original profile embeddings in the corresponding gallery. R@1 reports the percentage of queries whose matching original profile ranks first. Distillation results report the mean ACC and AUC across the 12 downstream tasks after using each proteomic encoder as the teacher for ProSlide. PFE provides stronger distillation supervision than Protein-BERT even without virtual profiles, while moderate virtual profile generation improves both retrieval and downstream performance. The largest virtual batch achieves the highest downstream AUC despite weaker retrieval, indicating that retrieval accuracy alone does not fully characterize the value of a proteomic teacher.

**Supervision and training stages.** To assess the roles of Stage II supervision and Stage III distillation, we compare supervision strategies and combinations of the two stages (Table 3). Subtype-supervised pretraining produces low-effective-rank representations that remain low-rank after distillation and yield weaker downstream performance. Protein-expression supervision instead produces higher-rank representations and a more effective initialization for knowledge transfer. Combining protein-supervised pretraining with Prot2Path achieves the best ACC and AUC across all three cancer groups, outperforming either stage alone and supporting their complementary contributions.

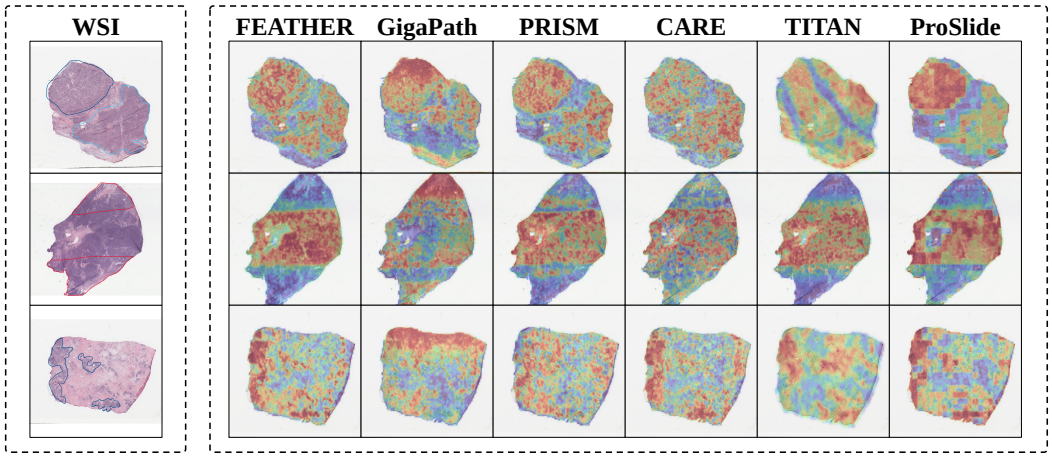



Figure 3: Visualization comparison of original WSIs and heatmaps. Heatmaps overlay patch aggregation weights for multiple-instance models or final-layer [CLS]-to-image-token attention weights for transformer-based models onto the corresponding WSIs.

**Prot2Path components.** To examine the contributions of relational supervision and reference-bank construction, we compare alternative alignment objectives and reference configurations (Table 4). Prot2Path outperforms hard matching and paired-bank KD in both ACC and AUC across all three cancer groups, supporting soft relational targets and the inclusion of standalone molecular references. Compared with direct alignment, its gains are consistent in ACC, while AUC favors different methods across cancer groups. Sampling standalone references maintains competitive performance with fewer reference entries, although neither sampled nor full-bank distillation consistently dominates across metrics. These findings support the effectiveness of the combined design while identifying reference sampling as a trade-off between bank size and predictive performance.

### 4.4 Interpretability Analysis

To examine the pathological relevance of ProSlide's attention, we compare its heatmaps with those of other foundation models on three representative CPTAC-BRCA cases (Fig. 3). The annotations provided by an invited pathologist support the assessment of tissue composition, preparation artifacts, and tumor localization. In the first case, tumor occupies most of the tissue, but the light-blue region contains more inflammatory and stromal components than the dark-blue region. ProSlide assigns stronger attention to the dark-blue region, indicating preferential attention to relatively tumor-rich tissue. In the second case, the tissue is also predominantly tumor-containing, while the red contours identify blurred areas introduced during slide preparation. ProSlide assigns lower attention to these areas and emphasizes the better-preserved central tissue. In the third case, the blue contours delineate localized tumor regions. ProSlide's higher-attention areas broadly overlap with the annotated regions along the left and lower portions of the tissue. Together, these examples suggest that ProSlide's attention is sensitive to tumor distribution, tissue composition, and image quality, providing complementary qualitative evidence for the pathological relevance of its representations.

## 5 Conclusion

This work investigated whether protein-derived knowledge could improve pathology foundation models. We introduced a three-stage framework that combined independent PFE pretraining, protein-supervised ProSlide pretraining, and Prot2Path relational distillation. Standalone protein profiles contributed to both proteomic representation learning and the distillation reference bank, extending molecular supervision beyond paired cohorts. Evaluation across 12 tasks in breast, renal, and lung cancers supported the effectiveness of this approach and highlighted the value of standalone proteomics for data-efficient pathology representation learning.

### AI USE STATEMENT

We used generative AI tools to assist with language editing. The authors reviewed and revised all AI-assisted text to verify its factual accuracy, preserve the intended scientific meaning, and ensure consistency with the methods and experimental results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This study analyzes existing patient-derived whole-slide images and RPPA profiles, as described in Appendix A. The reported experiments do not establish clinical safety or utility, and ProSlide should not be used to guide diagnosis or treatment without independent clinical validation and appropriate oversight.

### REPRODUCIBILITY STATEMENT

Appendix A describes the proteomic data sources, protein harmonization, and preprocessing. Section 3.1 and Appendix B describe the PFE architecture and pretraining. Sections 3.2–3.3 and Appendix C describe ProSlide and Prot2Path. Section 4.1 and Appendix D describe the downstream cohorts, patient-level splits, evaluation protocol, and baselines. Section 4.1 provides a link to an anonymous code repository.

## REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Paul H Acosta, Vandana Panwar, Vipul Jarmale, Alana Christie, Jay Jasti, Vitaly Margulis, Dinesh Rakheja, John Cheville, Bradley C Leibovich, Alexander Parker, et al. Intratumoral resolution of driver gene mutation heterogeneity in renal cancer using deep learning. *Cancer Research*, 82(15): 2792–2806, 2022.

Johanna K Ahlskog, Brian D Larsen, Kavya Achanta, and Claus S Sørensen. Atm/atr-mediated phosphorylation of palb2 promotes rad51 function. *The EMBO Reports*, 17(5):671–681, 2016.

Rehan Akbani, Patrick Kwok Shing Ng, Henrica MJ Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, Ji-Yeon Yang, Kosuke Yoshihara, Jun Li, et al. A pan-cancer proteomic perspective on the cancer genome atlas. *Nature communications*, 5(1):3887, 2014.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155, 2022.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, 2023.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, 2025.

Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713, 2015.

Zeyu Gao, Anyu Mao, Yuxing Dong, Hannah Clayton, Jialun Wu, Jiashuai Liu, ChunBao Wang, Kai He, Tieliang Gong, Chen Li, et al. Smmile enables accurate spatial quantification in digital pathology using multiple-instance learning. *Nature cancer*, 6(12):2025–2041, 2025.

Zeyu Gao, Kai He, Weiheng Su, Xiaobo Pang, Ines P Machado, Mercedes Jimenez-Linan, Brian Rous, Chunbao Wang, Chengzu Li, William McGough, et al. Alpaca: Adapting llama for pathology context analysis to enable slide-level question answering. *Nature Communications*, 2026.

Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald III, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, 2019.

Christian Grashei, Christian Brechenmacher, Rao Muhammad Umer, Jingsong Liu, Carsten Marr, Peter Schüffler, and Ewa Szczurek. Pathryoshka: Compressing pathology foundation models via multi-teacher knowledge distillation with nested embeddings. In *European Conference on Computer Vision*, pp. 76–94, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37029-7.

Yongxin Guo, Hao Lu, Onur C Koyun, Zhengjie Zhu, Muhammet F Demir, and Metin N Gurcan. Momentum memory for knowledge distillation in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6889–6899, 2026.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Jun Li, Yiling Lu, Rehan Akbani, Zhenlin Ju, Paul L Roebuck, Wenbin Liu, Ji-Yeon Yang, Bradley M Broom, Roeland GW Verhaak, David W Kane, et al. Tcpa: a resource for cancer functional proteomics data. *Nature methods*, 10(11):1046–1047, 2013.

Wei Li, Fan Yang, Fang Wang, Yu Rong, Linjing Liu, Bingzhe Wu, Han Zhang, and Jianhua Yao. scprotein: a versatile deep graph contrastive learning framework for single-cell proteomics embedding. *Nature Methods*, 21(4):623–634, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Drew M Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature reviews cancer*, 12(4):252–264, 2012.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54219–54238. PMLR, 2025.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports*, 9(1):3358, 2019.

Johann Wenckstern, Eeshaan Jain, Benedikt von Querfurth, Yexiang Cheng, Kiril Vasilev, Matteo Pariset, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Andreas Wicki, et al. The virtual tissues foundation model resolves spatial proteomics across scales. *Nature*, pp. 1–11, 2026.

Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, 11:759007, 2021.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Yingxue Xu, Yihui Wang, Fengtao Zhou, Jiabo Ma, Cheng Jin, Shu Yang, Jinbang Li, Zhengyu Zhang, Chenglong Zhao, Huajun Zhou, et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature communications*, 16(1):11406, 2025.

Chunze Yang, Qidong Liu, Wenjie Zhao, Yue Tang, Jiusong Ge, Di Zhang, Jiashuai Liu, Lei Wu, Junbo Lu, Ni Zhang, Xian Wu, Zeyu Gao, and Chen Li. Pathnavigate: A training-free pathology agent with surprise-guided scan and shared slide memory for whole-slide image vqa. *arXiv preprint arXiv:2605.23559*, 2026.

Di Zhang, Zhangpeng Gong, Xiaobo Pang, Jiashuai Liu, Junbo Lu, Hao Cui, Jiusong Ge, Zhi Zeng, Kai Yi, Yinghua Li, et al. Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21078–21088, 2026a.

Di Zhang, Jiashuai Liu, Youyuan Ma, Jiusong Ge, Zhi Zeng, Wenfang Sun, Qidong Liu, Kai He, Yefeng Zheng, Weimiao Yu, et al. Learnable frozen feature augmentation for few-shot biomarker prediction from pathology whole-slide images. *Bioinformatics*, pp. btag597, 2026b.

Wei Zhao, Jun Li, Mei-Ju M Chen, Yikai Luo, Zhenlin Ju, Nicole K Nesser, Katie Johnson-Camacho, Christopher T Boniface, Yancey Lawrence, Nupur T Pande, et al. Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. *Cancer cell*, 38(6): 829–843, 2020.

Qifeng Zhou, Lei Yu, Yuzhi Guo, Yuwei Miao, Hehuan Ma, Wenliang Zhong, Lin Xu, and Junzhou Huang. scpformer: A foundation model for unified representation and integration of the single-cell proteomics. *arXiv preprint arXiv:2604.20003*, 2026.

# Can Protein-Derived Knowledge Improve Pathology Foundation Models?

***Supplementary Material***

## Table of Contents


**A Proteomic Data and Preprocessing** **14**

A.1 Proteomic Data Resources . . . 14

A.2 Cross-Resource Protein Harmonization . . . 14

A.3 Missing Measurements and Expression Normalization . . . 15

A.4 Selection of the 38-Protein Panel . . . 15

**B PFE Pretraining** **15**

B.1 Details of Virtual Profile Generation . . . 16

B.2 Details of Expression-Space Multi-View Construction . . . 17

B.3 Profile- and Protein-Level Learning Objectives . . . 18

B.4 PFE Pretraining Configuration . . . 20

B.5 Comparison of Protein Modeling Paradigms . . . 20

**C ProSlide Pretraining and Prot2Path Distillation** **21**

C.1 ProSlide Architecture . . . 21

C.2 Prot2Path Implementation Details . . . 22

**D Downstream Evaluation and Additional Analyses** **22**

D.1 Datasets and Task Definitions . . . 22

D.2 Evaluation Protocols and Baselines . . . 24

D.3 Detailed Downstream Results . . . 25

D.4 Data Efficiency and Inference Memory Usage . . . 25

D.5 Analysis of PFE Protein Identity Embeddings . . . 26

# A PROTEOMIC DATA AND PREPROCESSING

To facilitate reproducibility, we provide additional details on the proteomic datasets and the corresponding preprocessing procedures. All four resources contain reverse-phase protein array (RPPA) measurements, which provide antibody-based quantification of both total proteins and phosphorylation-specific protein targets. After harmonization, all samples are represented using a common panel of 38 cancer-related protein measurements.

## A.1 PROTEOMIC DATA RESOURCES

We integrate RPPA profiles from four complementary resources: TCGA, TCPA, CCLE, and CPPA. These resources encompass patient tumor specimens, baseline cancer cell lines, and pharmacologically perturbed cancer cell lines, providing diverse protein-expression patterns for PFE pretraining. As summarized in Table S1, the resulting corpus comprises 13,586 protein profiles: 7,349 from TCGA, 878 from TCPA, 899 from CCLE, and 4,460 from CPPA. We curated a total of 13,586 proteomic profiles. Of these, 891 TCGA profiles were held out for evaluation, and the remaining 12,695 profiles were used for PFE pretraining.

**The Cancer Genome Atlas (TCGA).** The Cancer Genome Atlas (TCGA) is a large-scale pan-cancer initiative that provides genomic, molecular, and histopathological characterization across a broad range of human malignancies (Akbani et al., 2014). We use TCGA RPPA measurements as the primary source of patient-derived tumor proteomic profiles. The current cohort contains 7,349 RPPA records spanning 32 cancer cohorts. Unlike the external cell-line resources described below, TCGA provides naturally occurring tumor protein-expression profiles and constitutes the major patient tissue domain in PFE pretraining.

**The Cancer Proteome Atlas (TCPA).** The Cancer Proteome Atlas (TCPA) provides uniformly processed RPPA-based functional proteomic measurements for cancer samples and cell lines (Li et al., 2013). In our pretraining cohort, we use the CCLE RPPA500 profiles distributed through TCPA rather than treating TCPA as an additional independent TCGA patient cohort. After protein-name harmonization and coverage filtering, 878 cancer cell-line profiles are retained. These samples complement patient-derived TCGA tumors by introducing protein-expression patterns from experimentally controlled cancer cell-line systems.

**Cancer Cell Line Encyclopedia (CCLE).** The Cancer Cell Line Encyclopedia (CCLE) provides systematic molecular characterization of cancer cell lines originating from diverse tissue types (Ghandi et al., 2019). We use the CCLE RPPA measurements together with the corresponding antibody annotations to resolve protein identities and phosphorylation-specific targets. Following harmonization with the common 38-protein panel, 899 baseline cancer cell-line profiles are included in PFE pretraining. In contrast to TCGA tumor specimens, these samples represent protein expression in unperturbed cancer cell lines and therefore provide an additional molecular context for learning reusable proteomic representations.

**Cancer Perturbed Proteomics Atlas (CPPA).** The Cancer Perturbed Proteomics Atlas (CPPA) characterizes protein responses of cancer cell lines under pharmacological perturbations (Zhao et al., 2020). Unlike TCGA, TCPA, and CCLE profiles that primarily describe protein abundance states, CPPA records treatment-induced changes in protein abundance. Specifically, a perturbation response is represented as $\Delta p = p_1 - p_0$, where $p_0$ denotes the baseline protein level before treatment and $p_1$ denotes the corresponding level after drug exposure. We retain 4,460 CPPA perturbation records after protein mapping and coverage filtering. These profiles expose PFE to drug-, dose-, and treatment-dependent protein responses, thereby broadening the molecular variation available during pretraining. Importantly, each CPPA record corresponds to a perturbation condition rather than an independent cell line.

## A.2 CROSS-RESOURCE PROTEIN HARMONIZATION

The original RPPA resources contain different antibody panels and use heterogeneous naming conventions for proteins and phosphorylation sites. We therefore map all sources to a fixed panel of

Table S1: Composition of the proteomic cohort.

| Resource | Records | Sample Domain | Protein Profile Type |
|---|---|---|---|
| TCGA | 7,349 | Patient tumors | Baseline tumor RPPA |
| TCPA | 878 | Cancer cell lines | Baseline cell-line RPPA |
| CCLE | 899 | Cancer cell lines | Baseline cell-line RPPA |
| CPPA | 4,460 | Perturbed cell lines | Treatment-induced $\Delta p$ |
| **Total** | **13,586** | – | – |

$M = 38$ cancer-related protein targets before PFE pretraining (see Appendix A.4). Protein names are first standardized by normalizing capitalization and removing formatting differences and technical suffixes. Manually curated aliases are further used to reconcile established synonyms across resources. For phosphorylation-specific measurements, mapping is performed conservatively: only measurements corresponding to the same protein and the same modification site are treated as equivalent, while distinct phosphorylation sites are retained as separate molecular targets.

### A.3 Missing Measurements and Expression Normalization

Because individual RPPA resources do not uniformly measure all 38 target proteins, some protein measurements may be unavailable for a given sample. Missing entries arise when a target is absent from the original antibody panel, cannot be reliably mapped across resources, or contains an invalid measurement. These entries are explicitly marked as missing and are handled by the learnable mask representation described in Eq. 2. To maintain the original characteristics of each RPPA resource, we do not impose an additional cross-resource global normalization when constructing the unified dataset. Instead, source measurements are retained on their original processed scales after protein harmonization. During PFE training, valid expression measurements are standardized using protein-specific statistics before being mapped into protein tokens. This separates the preservation of the original RPPA measurements from the numerical normalization required for stable representation learning. External cell-line and perturbation samples are additionally filtered according to protein coverage. Only profiles containing measurements for a sufficient number of targets in the shared protein panel are retained, preventing extremely sparse profiles from dominating the self-supervised pretraining process.

### A.4 Selection of the 38-Protein Panel

We use a fixed panel of 38 cancer-related RPPA targets to establish a shared input space across the integrated proteomic resources. The panel spans DNA damage response and repair, growth signaling, cell-cycle and chromatin regulation, hormone receptor signaling, lineage-associated transcription, immune regulation, and hypoxia-related processes. This composition provides complementary molecular measurements relevant to tumor biology and the cellular context of bulk specimens. Table S2 lists the panel entries, standardized targets, and functional annotations. The panel comprises 32 distinct proteins and six additional site-specific phosphorylation measurements involving ESR1, ERBB2, EGFR, MET, and MTOR. Each phosphorylation measurement is retained as a separate target from its corresponding total-protein measurement, preserving the distinction between protein abundance and site-specific phosphorylation.

## B Proteomic Pretraining

Conventional multi-view self-supervised learning is largely developed for images (Oquab et al., 2023), where semantically related views can be constructed through spatial cropping and photometric transformations. Directly transferring this strategy to bulk proteomic profiles is non-trivial, because a protein-expression profile has no intrinsic spatial structure, each dimension is tied to an explicitly defined protein identity, and the corresponding measurements are continuous expression values. We therefore reformulate multi-view learning directly in the protein-expression space, preserving protein identities while perturbing only their expression states.

Table S2: **Composition of the 38-protein panel.** Total-protein and site-specific phosphorylation measurements are retained as distinct targets.

| Panel Name | Standardized Target | Functional Annotation |
|---|---|---|
| ATM | ATM | DNA damage response |
| ATR | ATR | DNA damage response |
| BRCA2 | BRCA2 | Homologous recombination repair |
| P53 | TP53 | Tumor suppression / DNA damage response |
| RB | RB1 | Cell-cycle regulation |
| BAP1C4 | BAP1 | Chromatin regulation / tumor suppression |
| SETD2 | SETD2 | Chromatin regulation / tumor suppression |
| AXL | AXL | Receptor tyrosine kinase signaling |
| BRAF | BRAF | RAS–MAPK signaling |
| NRAS | NRAS | RAS–MAPK signaling |
| EGFR | EGFR | Receptor tyrosine kinase signaling |
| EGFR_pY1068 | EGFR pTyr1068 | Receptor tyrosine kinase signaling |
| EGFR_pY1173 | EGFR pTyr1173 | Receptor tyrosine kinase signaling |
| HER2 | ERBB2 | Receptor tyrosine kinase signaling |
| HER2_pY1248 | ERBB2 pTyr1248 | Receptor tyrosine kinase signaling |
| CMET | MET | Receptor tyrosine kinase signaling |
| CMET_pY1235 | MET pTyr1235 | Receptor tyrosine kinase signaling |
| PDGFRB | PDGFRB | Receptor tyrosine kinase signaling |
| MTOR | MTOR | PI3K–AKT–mTOR signaling |
| MTOR_pS2448 | MTOR pSer2448 | PI3K–AKT–mTOR signaling |
| PTEN | PTEN | Negative regulation of PI3K–AKT signaling |
| SMAD4 | SMAD4 | TGF-$\beta$ signaling |
| ERALPHA | ESR1 (ER-$\alpha$) | Hormone receptor signaling |
| ERALPHA_pS118 | ESR1 pSer118 | Hormone receptor signaling |
| PR | PGR | Hormone receptor signaling |
| GATA3 | GATA3 | Lineage-associated transcriptional regulation |
| MITF | MITF | Lineage-associated transcriptional regulation |
| PAX8 | PAX8 | Lineage-associated transcriptional regulation |
| CD4 | CD4 | Immune-cell marker / T-cell co-receptor |
| CD86 | CD86 | Immune co-stimulation |
| CIITA | CIITA | Regulation of antigen presentation |
| CTLA4 | CTLA4 | Immune checkpoint |
| PDCD1 | PDCD1 (PD-1) | Immune checkpoint |
| PDL1 | CD274 (PD-L1) | Immune checkpoint ligand |
| CD44 | CD44 | Cell adhesion / stemness-associated processes |
| MYH11 | MYH11 | Smooth-muscle contractile machinery |
| VHL | VHL | Regulation of hypoxia signaling |
| CA9 | CA9 (CAIX) | Hypoxia-associated pH regulation |

Our expression-space pretraining strategy addresses two complementary challenges of bulk proteomic foundation modeling. First, the limited number of available patient-level protein profiles restricts the diversity of molecular states observed during pretraining. We therefore introduce *Virtual Profile Generation* to expand the effective proteomic cohort from real parent profiles. Second, because bulk protein profiles lack the spatial structure that enables conventional image-based multi-view augmentation, we construct semantically related views directly in the protein-expression space using *profile-wide views* and *subset-perturbed views*. These designs are subsequently coupled with profile- and protein-level self-supervised objectives to learn both global molecular representations and contextualized protein representations.

### B.1 Details of Virtual Profile Generation

To clarify how virtual protein profiles are generated, we describe the source-profile sampling procedure and the three expression perturbation operators used during PFE pretraining.

Each mini-batch contains $B_{\mathrm{r}}$ real profiles sampled from $\mathcal{D}$ and $B_{\mathrm{v}}$ virtual profiles generated online. For each virtual profile $a$, we independently sample a source profile $\mathcal{X}_{k_a}$ from the full pretraining corpus. Source selection is not restricted to the real profiles present in the current mini-batch.

Let $\mathcal{O}_{k_a}$ denote the observed proteins in the source profile:

$$\mathcal{O}_{k_a} = \{j \in \{1, \dots, M\} : m_{k_a,j} = 1\}. \tag{13}$$

We sample a small subset $\mathcal{S}_a \subseteq \mathcal{O}_{k_a}$ and modify only its expression values according to Eq. 3. One augmentation type is sampled per virtual profile, with probabilities $\pi_{\mathrm{post}}$, $\pi_{\mathrm{gau}}$, and $\pi_{\mathrm{mix}}$:

$$\pi_{\mathrm{post}} + \pi_{\mathrm{gau}} + \pi_{\mathrm{mix}} = 1. \tag{14}$$

Next, we introduce three Virtual Profile Generation Operators.

**Partial posterization.** Posterization reduces the numerical precision of selected expression values. Let $\ell_j < u_j$ denote the scaling bounds for protein $j$, and let $K \geq 2$ denote the number of quantization levels. For an expression value within these bounds, we first map it to the unit interval, quantize it, and restore the original scale:

$$z_{a,j} = \frac{x_{k_a,j} - \ell_j}{u_j - \ell_j}, \tag{15}$$

$$\bar{z}_{a,j} = Q_K(z_{a,j}) = \frac{\mathrm{round}((K-1)z_{a,j})}{K-1}, \tag{16}$$

$$\tilde{x}_{a,j} = \ell_j + (u_j - \ell_j)\bar{z}_{a,j}, \qquad j \in \mathcal{S}_a. \tag{17}$$

The quantizer maps values in $[0, 1]$ to $K$ uniformly spaced levels. Inverse scaling restores the expression scale while retaining the loss of within-bin precision introduced by quantization.

**Partial Gaussian perturbation.** For each selected protein, we independently add Gaussian noise on the normalized expression scale:

$$\tilde{x}_{a,j} = x_{k_a,j} + \epsilon_{a,j}, \qquad \epsilon_{a,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\mathrm{virt}}), \qquad j \in \mathcal{S}_a. \tag{18}$$

The perturbation changes only selected expression values and does not alter protein identities or missingness indicators.

**Partial mixup.** For each virtual profile assigned mixup, we sample a donor uniformly from the corpus, excluding its source:

$$b_a \sim \mathrm{Unif}(\{1, \dots, N\} \setminus \{k_a\}). \tag{19}$$

Interpolation is restricted to selected proteins observed in both profiles:

$$\mathcal{S}^{\mathrm{mix}}_a = \{j \in \mathcal{S}_a : m_{b_a,j} = 1\}. \tag{20}$$

We draw one coefficient per virtual profile,

$$\lambda_a \sim \mathrm{Beta}(\alpha, \alpha), \qquad \alpha = 0.4, \tag{21}$$

and construct the expression values as

$$\tilde{x}_{a,j} = \begin{cases} \lambda_a x_{k_a,j} + (1 - \lambda_a) x_{b_a,j}, & j \in \mathcal{S}^{\mathrm{mix}}_a, \\ x_{k_a,j}, & \text{otherwise.} \end{cases} \tag{22}$$

The same $\lambda_a$ is shared across all interpolated proteins. Selected proteins missing from the donor remain unchanged, and the virtual profile retains the source protein identities and missingness pattern.

### B.2 Details of Expression-Space Multi-View Construction

To encourage consistent proteomic representations under expression perturbations, we construct complementary profile-wide and subset-perturbed views for each input profile. View construction is applied after real and virtual profiles have been assembled into a mini-batch. Let $\mathcal{X} = \{(p_j, x_j, m_j)\}_{j=1}^{M}$ denote either a real or a virtual profile, with observed set $\mathcal{O} = \{j : m_j = 1\}$. All views retain the complete set of $M$ protein identities, with transformations restricted to observed expression values.

Table S3: **Architecture configuration of the Proteomic Foundation Encoder (PFE).** PFE represents each protein as the sum of a learnable identity embedding and a continuous abundance-value embedding, followed by Transformer-based contextual modeling.

| Component | Configuration | Description |
|---|---|---|
| Protein panel size | 38 | Shared protein vocabulary |
| Input sequence length | 39 | One `[CLS]` and 38 protein tokens |
| Value projection $g_{\text{val}}$ | $1 \to 32 \to 768$ | Projects continuous abundance values |
| Missing-value embedding | $1 \times 768$ | Shared learnable value-mask embedding |
| Hidden dimension | 768 | Transformer token dimension |
| Transformer depth | 7 | Number of encoder blocks |
| Attention heads | 8 | Multi-head self-attention |
| MLP ratio | 4 | Transformer feed-forward expansion ratio |
| Dropout | 0.1 | Applied in Transformer blocks |
| Attention dropout | 0.1 | Applied to self-attention |
| Profile representation | 768-d `[CLS]` | Patient-level molecular representation |
| Protein representation | 768-d per token | Contextualized protein-specific representation |
| Projection head | $768 \to 2048 \to 2048 \to 256$ | DINO-style projection MLP |
| Prototype dimension | 2048 | Profile- and protein-level soft targets |

**Profile-wide views.** We construct three profile-wide views: one clean view and two stochastically augmented views. The clean view preserves the input profile. Each augmented view applies mild posterization to all observed expression values using independently sampled augmentation settings:

$$x_j^{(v)} = \begin{cases} \mathcal{P}_{\theta_v,j}(x_j), & j \in \mathcal{O}, \\ x_j, & \text{otherwise}, \end{cases} \qquad v \in \{1, 2\}. \tag{23}$$

Here, $\mathcal{P}_{\theta_v,j}$ denotes the posterization-based transformation for protein $j$ under the sampled configuration $\theta_v$. Equation 17 defines its base quantization operation. Both augmented views preserve protein identities and naturally missing measurements.

**Subset-perturbed views.** We additionally construct $V_{\text{sub}}$ views, each using an independently sampled subset $\mathcal{S}^{(v)} \subseteq \mathcal{O}$. Let $\mathcal{G}^{(v)} \subseteq \mathcal{S}^{(v)}$ and $\mathcal{R}^{(v)} \subseteq \mathcal{S}^{(v)}$ denote the positions selected for Gaussian perturbation and artificial masking, respectively. Before masking, the expression values are

$$x_j^{(v)} = \begin{cases} x_j + \epsilon_j^{(v)}, & j \in \mathcal{G}^{(v)}, \\ x_j, & \text{otherwise}, \end{cases} \qquad \epsilon_j^{(v)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{sub}}^2). \tag{24}$$

The expression state at each position in $\mathcal{R}^{(v)}$ is then hidden from the encoder while its protein identity token is retained. Artificially masked positions originate from observed measurements and are distinguished from naturally missing proteins. Positions outside the selected subset remain unchanged.

**Asymmetric view assignment.** Let $\mathcal{V}_{\text{wide}}(\mathcal{X})$ and $\mathcal{V}_{\text{sub}}(\mathcal{X})$ denote the profile-wide and subset-perturbed view collections. The teacher and student receive

$$\mathcal{V}_{\text{teacher}}(\mathcal{X}) = \mathcal{V}_{\text{wide}}(\mathcal{X}), \qquad \mathcal{V}_{\text{student}}(\mathcal{X}) = \mathcal{V}_{\text{wide}}(\mathcal{X}) \cup \mathcal{V}_{\text{sub}}(\mathcal{X}). \tag{25}$$

Thus, the teacher processes three profile-wide views, while the student additionally processes $V_{\text{sub}}$ locally perturbed views. This assignment is applied to both real and virtual profiles. The profile- and protein-level learning objectives are defined in Appendix B.3.

### B.3 Profile- and Protein-Level Learning Objectives

To learn proteomic representations at both the profile and protein levels, we pretrain PFE using a student–teacher self-distillation framework. The student parameters $\theta_s$ are optimized by gradient descent, while the teacher parameters $\theta_t$ are updated as an exponential moving average of $\theta_s$.

Table S4: **Self-supervised pretraining configuration of PFE.** The model is trained using Distributed Data Parallel (DDP) on four NVIDIA RTX 4090 GPUs.

| Category | Hyperparameter | Value |
|---|---|---|
| Expression views | Profile-wide views | 3 |
| | Profile-wide view composition | 1 clean + 2 posterized |
| | Subset-perturbed views | 6 |
| | Subset-perturbed Gaussian noise | $\sigma = 0.05$ |
| | Gaussian perturbation ratio | 30% of observed proteins |
| Protein masking | Profile-wide student mask ratios | 0.05 / 0.10 / 0.15 |
| | Teacher random masking | None |
| | Subset-perturbed mask ratios | 0.07–0.20 |
| Virtual profiles | Virtual profiles / GPU / step | 64 |
| | Real profiles / GPU / step | 32 |
| | Augmentation types | Posterization / Gaussian / mixup |
| | Augmentation probability | $1/3$ each |
| | Gaussian noise | $\sigma = 0.12$ |
| | Mixup coefficient | $\lambda \sim \mathrm{Beta}(0.4, 0.4)$ |
| Self-distillation | Profile loss weight | 1.0 |
| | Protein loss weight | 1.0 |
| | KoLeo weight | 0.1 |
| | Student temperature | 0.10 |
| | Teacher temperature | $0.04 \rightarrow 0.07$ |
| | Temperature warm-up | 10 epochs |
| | Center momentum | 0.90 |
| Optimization | Optimizer | Adam |
| | Peak learning rate | $7 \times 10^{-5}$ |
| | Minimum learning rate | $1 \times 10^{-6}$ |
| | Learning-rate schedule | Warm-up + cosine decay |
| | Warm-up epochs | 10 |
| | Weight decay | $1 \times 10^{-5}$ |
| | Gradient clipping | 3.0 |
| | Epochs | 100 |
| EMA teacher | Initial momentum | 0.996 |
| | Momentum schedule | Cosine to 1.0 |

**Profile-level distillation.** We project the `[CLS]` representation of each view into a probability distribution. Let $p_t^{(u)}$ and $p_s^{(v)}$ denote the teacher and student outputs for two different views of the same profile. The profile-level objective is

$$\mathcal{L}_{\mathrm{profile}} = \frac{1}{|\mathcal{P}_{\mathrm{profile}}|} \sum_{(u,v)\in\mathcal{P}_{\mathrm{profile}}} H\left(\mathrm{sg}[p_t^{(u)}], p_s^{(v)}\right), \tag{26}$$

where $\mathcal{P}_{\mathrm{profile}}$ contains valid cross-view pairs, $\mathrm{sg}[\cdot]$ stops gradients through the teacher, and $H$ denotes cross-entropy. This objective is averaged over both real and virtual profiles.

**Protein-level distillation.** For subset-perturbed student views, we additionally align each artificially masked protein with its unmasked counterpart in a profile-wide teacher view. Let $\mathcal{M}^{(v)}$ denote the masked protein positions in student view $v$. The protein-level objective is

$$\mathcal{L}_{\mathrm{protein}} = \frac{1}{|\mathcal{P}_{\mathrm{protein}}|} \sum_{(u,v)\in\mathcal{P}_{\mathrm{protein}}} \frac{1}{|\mathcal{M}^{(v)}|} \sum_{j\in\mathcal{M}^{(v)}} H\left(\mathrm{sg}[p_{t,j}^{(u)}], p_{s,j}^{(v)}\right), \tag{27}$$

where $\mathcal{P}_{\mathrm{protein}}$ pairs profile-wide teacher views with subset-perturbed student views. This objective is computed only for real profiles.

**Overall objective.** The complete pretraining objective is

$$\mathcal{L}_{\mathrm{PFE}} = \mathcal{L}_{\mathrm{profile}} + \mathcal{L}_{\mathrm{protein}} + \lambda_{\mathrm{K}}\mathcal{L}_{\mathrm{KoLeo}}, \tag{28}$$

Table S5: **Comparison of representative protein modeling approaches.** We summarize primary inputs, principal learning objectives, and representations or outputs.

| Model | Primary Input | Learning Objective | Representation / Output |
|---|---|---|---|
| ESM-2 (Lin et al., 2022) | Amino-acid sequences | Masked amino-acid prediction | Residue and protein embeddings |
| AlphaFold 2 (Jumper et al., 2021) | Sequences, alignments, and structural templates | Structure prediction | Three-dimensional protein structures |
| AlphaFold 3 (Abramson et al., 2024) | Biomolecular sequences and ligand identities | Diffusion-based structure prediction | Three-dimensional biomolecular complexes |
| scPROTEIN (Li et al., 2024) | Single-cell peptide and protein measurements | Graph contrastive learning and uncertainty estimation | Cell embeddings |
| scpFormer (Zhou et al., 2026) | Single-cell protein abundance and sequence-derived embeddings | Multi-objective masked expression reconstruction | Cell and contextualized protein embeddings |
| VirTues (Wenckstern et al., 2026) | Multiplex protein images and sequence-derived marker embeddings | Masked image reconstruction | Marker, cell, niche, and tissue representations |
| **PFE (Ours)** | Bulk RPPA profiles | Expression-space multi-view self-supervised learning | Profile and contextualized protein embeddings |

where $\mathcal{L}_{\mathrm{KoLeo}}$ discourages collapse among profile-level representations. Virtual profiles contribute to $\mathcal{L}_{\mathrm{profile}}$ but are excluded from $\mathcal{L}_{\mathrm{protein}}$, preventing artificially edited protein values from serving as fine-grained token targets.

### B.4 PFE Pretraining Configuration

PFE uses a seven-layer Transformer with a hidden dimension of 768 and eight attention heads. Each input profile is represented by 38 protein tokens and one learnable `[CLS]` token, producing both profile-level and contextualized protein-level representations. Table S3 summarizes the architectural configuration, including the continuous value encoder and self-distillation projection head.

We pretrain PFE for 100 epochs using distributed data parallelism on four NVIDIA RTX 4090 GPUs. Each GPU generates 64 virtual profiles per training step, and each real or virtual profile is used to construct three profile-wide views and six subset-perturbed views. Optimization uses Adam with a 10-epoch learning-rate warm-up followed by cosine decay. Detailed augmentation settings, self-distillation hyperparameters, and optimization schedules are provided in Table S4. The compact panel of 38 cancer-related targets yields only 39 tokens per profile, keeping the memory overhead of self-attention modest and making pretraining feasible on four NVIDIA RTX 4090 GPUs under this configuration.

### B.5 Comparison of Protein Modeling Paradigms

Protein representation learning encompasses amino-acid sequence modeling, molecular structure prediction, and the analysis of single-cell and spatial proteomic measurements. These approaches characterize individual proteins, cellular expression states, or spatial tissue organization, as summarized in Table S5. Learning reusable representations directly from sample-level bulk protein expression remains an emerging direction. Bulk proteomic pretraining faces practical constraints arising from limited cohort sizes and heterogeneous assay panels. To make effective use of available data, we integrate complementary RPPA resources into a shared protein panel and augment the resulting corpus through virtual profile generation. By perturbing selected expression values while preserving protein identities and natural missingness patterns, this strategy exposes PFE to additional local expression configurations anchored to observed samples.

Table S6: **Architectural configuration of ProSlide.**

| Parameter | Configuration |
|---|---|
| Patch encoder | CONCH v1.5 (frozen) |
| Patch feature dimension | 768 |
| Region grid | $4 \times 4$ patches |
| RSA depth | 3 |
| RSA attention heads | 8 |
| RSA MLP ratio | 2 |
| RSA aggregation | Learnable region `[CLS]` token |
| SSA depth | 7 |
| SSA attention heads | 8 |
| SSA MLP ratio | 4 |
| SSA aggregation | Learnable slide `[CLS]` token |
| Dropout | 0.1 |
| Attention dropout | 0.1 |
| Activation | GELU |
| Normalization | Pre-LN + final LayerNorm |
| Positional encoding | None |
| WSI representation | 768-dimensional slide `[CLS]` output |
| Protein regression heads | Independent linear layers: $768 \rightarrow 1$ |

Pretraining objectives also differ across modeling paradigms. ESM-2 and scpFormer employ masked prediction, whereas scPROTEIN uses graph contrastive learning. PFE adopts expression-space multi-view self-supervised learning to learn profile- and protein-level representations through self-distillation across perturbed views. While masked abundance reconstruction focuses on recovering withheld measurements, our objectives explicitly promote consistency under expression quantization, localized noise, and partial masking. This formulation encourages representations that remain stable under the specified perturbations, providing a proteomic reference space for subsequent Prot2Path distillation.

# C ProSlide Pretraining and Prot2Path Distillation

To facilitate reproducibility of Stages II and III, we detail the architecture of ProSlide, and the implementation of Prot2Path relational distillation.

## C.1 ProSlide Architecture

ProSlide follows a patch-to-region-to-slide hierarchy, with architectural settings summarized in Table S6. Using the patch-grid coordinates, we partition each WSI into non-overlapping regions spanning $4 \times 4$ patch positions. Within each region, the frozen CONCH v1.5 features are processed by a three-layer RSA encoder together with a learnable region `[CLS]` token. The corresponding output token serves as the region representation.

The region representations are processed by a seven-layer SSA encoder with a learnable slide `[CLS]` token, whose output forms the WSI representation. Both encoders use eight attention heads, GELU activations, pre-layer normalization, and a final LayerNorm. Patch-grid coordinates are used only to determine region membership, while neither RSA nor SSA introduces positional embeddings.

For protein-expression supervision, each target is assigned an independent linear regression head mapping the WSI representation to a scalar expression value. The feature projections, RSA and SSA encoders, learnable `[CLS]` tokens, and regression heads are jointly trained, while the CONCH v1.5 patch encoder remains frozen.

Table S7: **Implementation configuration of Prot2Path.**

| Parameter | Configuration |
|---|---|
| Proteomic teacher | Frozen PFE |
| Teacher representation | Final `[CLS]` output after LayerNorm |
| Representation dimension | 768 |
| Candidate pool size | 12,695 real profiles |
| Sub-bank size | $K_{\mathrm{sub}} = 1{,}594$ |
| Paired references | All 1,338 paired training profiles |
| Standalone references | 256 randomly sampled standalone profiles |
| Representation normalization | $L_2$ normalization |
| Similarity measure | Cosine similarity |
| Teacher temperature | $\tau_{\mathrm{t}} = 0.05$ |
| Student temperature | $\tau_{\mathrm{s}} = 0.05$ |
| Distillation objective | Teacher-to-student KL divergence |

### C.2 Prot2Path Implementation Details

To establish a shared molecular reference space for distillation, we first encode the candidate protein profiles using the frozen PFE. For each profile, we extract the 768-dimensional `[CLS]` representation after the final LayerNorm, without applying the pretraining projection head. The candidate reference pool contains 12,695 real protein profiles: 1,338 paired training profiles and 11,357 standalone profiles. Table S7 summarizes the implementation settings. A linear alignment head with bias maps the ProSlide representation into the 768-dimensional PFE space. Teacher representations, projected WSI representations, and bank entries are $L_2$-normalized before computing cosine similarities. Both teacher and student temperatures are fixed at 0.05. We minimize the teacher-to-student KL divergence, summing over reference entries and averaging over paired samples in each mini-batch. The loss has unit weight, with no temperature-squared scaling or auxiliary training losses.

## D Downstream Evaluation and Additional Analyses

To complement the downstream evaluation in the main text, we provide dataset and task descriptions, baseline configurations, and detailed results. We also present additional analyses of pretraining data efficiency, GPU memory usage during inference, and model interpretability.

### D.1 Datasets and Task Definitions

To assess the downstream utility of proteomics-guided pathology representations, we consider classification endpoints that characterize both molecular states and tissue morphology. The benchmark includes mutation prediction, receptor-status prediction, and histological subtyping, with cross-cohort settings incorporated to assess transferability across data sources. Below, we detail the 12 WSI-level tasks, grouped into renal (Tasks 1–4), breast (Tasks 5–8), and lung (Tasks 9–12) cancer benchmarks.

- **Task 1 (MUT-SETD2).** This task predicts *SETD2* mutation status in the MUT-HETRCC cohort. It evaluates morphology associated with this chromatin-regulation gene alteration.
- **Task 2 (MUT-BAP1).** This task predicts *BAP1* mutation status in the MUT-HETRCC cohort. It provides an additional renal-cancer mutation benchmark outside the CPTAC cohort.
- **Task 3 (CPTAC-CCRCC-BAP1).** This CPTAC clear-cell renal-cell carcinoma task predicts *BAP1* mutation status. It tests the relationship between WSI morphology and a key renal-cancer molecular alteration.
- **Task 4 (Cross-MUT-BAP1).** This task predicts *BAP1* mutation status under cross-cohort transfer. Patients from MUT-HETRCC are split into training and validation sets, while

Table S8: Task-specific cohort sizes for downstream evaluation.

| Task | Cohort(s) | Patients ($N_p$) | WSIs ($N_s$) |
|---|---|---|---|
| Task 1 | MUT-HETRCC | 1292 | 1292 |
| Task 2 | MUT-HETRCC | 1292 | 1292 |
| Task 3 | CPTAC-CCRCC | 103 | 245 |
| Task 4 | MUT-HETRCC + CPTAC-CCRCC | 1395 | 1537 |
| Task 5 | CPTAC-BRCA | 97 | 106 |
| Task 6 | CPTAC-BRCA | 82 | 89 |
| Task 7 | CPTAC-BRCA | 103 | 112 |
| Task 8 | BCNB | 1028 | 1028 |
| Task 9 | CPTAC-LUAD | 108 | 324 |
| Task 10 | Local-LUNG + DHMC-LUNG | 395 | 410 |
| Task 11 | Local-LUNG + CPTAC-(LUAD & LSCC) + DHMC-LUNG | 1003 | 2819 |
| Task 12 | Local-LUNG + DHMC-LUNG | 395 | 410 |

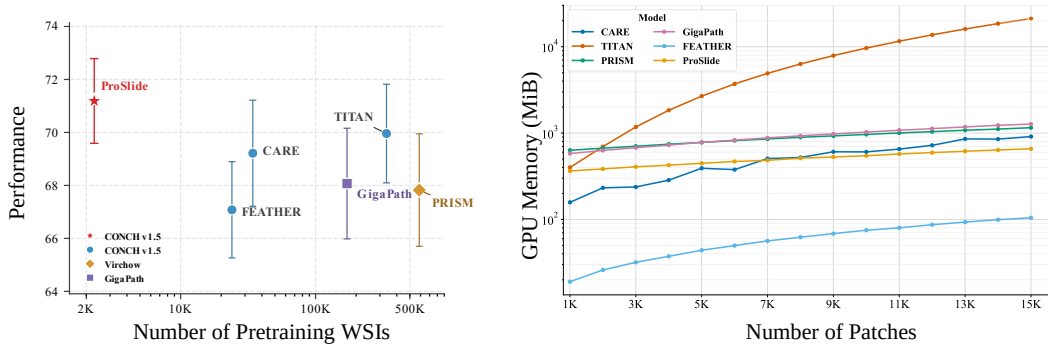
CPTAC-CCRCC serves as the fixed external test cohort. This setting evaluates whether a classifier trained on MUT-HETRCC transfers to CPTAC-CCRCC.

- **Task 5 (CPTAC-BRCA-PR).** This task uses the CPTAC breast cancer cohort to predict progesterone-receptor (PR) status from whole-slide images. It evaluates whether histomorphological patterns are informative for this clinically relevant receptor phenotype.
- **Task 6 (CPTAC-BRCA-HER2).** This CPTAC breast cancer task predicts *HER2* status. It assesses whether slide-level representations capture histomorphological patterns associated with HER2-positive breast cancer.
- **Task 7 (CPTAC-BRCA-TP53).** This task predicts *TP53* mutation status in the CPTAC breast cancer cohort. It serves as a molecular-phenotype benchmark with potentially heterogeneous mutation-associated morphology.
- **Task 8 (BCNB-ER).** This task predicts estrogen-receptor (*ER*) status in the BCNB breast cancer cohort. It complements the CPTAC breast cancer tasks with an independent data source.
- **Task 9 (CPTAC-LUAD-EGFR).** This CPTAC lung adenocarcinoma task predicts *EGFR* mutation status from whole-slide images. It evaluates sensitivity to a clinically actionable genomic alteration in lung cancer.
- **Task 10 (Combine-LUNG fine subtype).** This task performs three-class classification of the predominant histological patterns of lung adenocarcinoma: acinar, lepidic, and solid. Samples from the Local-LUNG and DHMC-LUNG cohorts are pooled to evaluate fine-grained histological subtype discrimination using multi-source data.
- **Task 11 (Cross-LUNG subtyping).** This task distinguishes lung adenocarcinoma (LUAD) from lung squamous cell carcinoma (LSCC) under cross-cohort transfer. Local-LUNG provides the training and validation patients, while DHMC-LUNG, CPTAC-LUAD, and CPTAC-LSCC jointly form the fixed external test set.
- **Task 12 (Cross-LUNG fine subtype).** This task performs three-class classification of the predominant histological patterns of lung adenocarcinoma: acinar, lepidic, and solid. The Local-LUNG cohort is used for training and validation, while DHMC-LUNG serves as the fixed test cohort. This setting evaluates cross-cohort generalization in fine-grained histological subtype discrimination.

The benchmark comprises nine molecular prediction tasks (Tasks 1–9), covering gene mutation and receptor-status prediction, and three histological subtyping tasks (Tasks 10–12). Task 10 uses pooled data from Local-LUNG and DHMC-LUNG, while Tasks 4, 11, and 12 assess cross-cohort transfer across different data sources. We use repeated patient-level holdout splits (Table S8). All WSIs from the same patient were assigned to the same partition within each split. Tasks 1, 2, 8, and 10 used five repeated patient-level splits, with target allocations of 60% for training, 20% for validation, and 20% for testing. Tasks 3, 5, 6, 7, and 9 used 50 repeated patient-level splits, allocating 80% of

Table S9: **Comparison of representative slide-level pathology foundation models.** We compare the patch encoder, whole-slide aggregation strategy, pretraining scale, and supervision of the pathology foundation models considered in this work.

| Model | Patch Encoder | Slide Aggregation | Pretraining Scale | Pretrain Sup. |
|---|---|---|---|---|
| CARE | CONCH v1.5 | CARE | 34,277 WSIs | iBoT & Clip |
| TITAN | CONCH v1.5 | ViT | 335,645 WSIs | iBoT & Coca |
| PRISM | Virchow | Perceiver | 587,196 WSIs | Coca |
| GigaPath | GigaPath | LongNet | 171,189 WSIs | MAE |
| FEATHER | CONCH v1.5 | ABMIL | 24,000 WSIs | Supervised classification |
| ProSlide | CONCH v1.5 | Region ViT | 12,695 protein profiles & 2,229 WSIs | Protein supervision & Proteomic knowledge distill |




(a) Performance vs Pre-training WSI Scale

(b) GPU Memory Analysis

Figure S1: **Data and memory efficiency.** ProSlide achieves strong performance with limited slide-level pretraining data (a) and low GPU memory usage during inference (b).

patients to training and 20% to testing, without a separate validation set. For Tasks 4, 11, and 12, the test patients were fixed across all five repetitions, while the development pool was repartitioned into training and validation sets at a 3:1 ratio. This ratio applies only to the development pool and excludes the fixed test set. The split proportions refer to patients rather than WSIs. Because patients can contribute different numbers of WSIs, slide counts need not follow the same proportions. Integer partition sizes can also differ slightly from the nominal patient-level proportions.

### D.2 Evaluation Protocols and Baselines

To assess the transferability of slide representations learned through different pretraining strategies, we evaluate all models using the frozen-feature linear-probing protocol described in the main text. Table S9 summarizes their patch encoders, slide aggregation architectures, and slide-level pretraining settings. The multimodal baselines incorporate molecular or textual supervision. CARE (Zhang et al., 2026a) combines adaptive region modeling with iBOT pretraining and subsequent contrastive alignment with transcriptomic and protein profiles. TITAN (Ding et al., 2025) uses a slide-level Vision Transformer pretrained through iBOT and CoCa-based alignment with synthetic captions and pathology reports. PRISM (Shaikovski et al., 2024) aggregates Virchow patch features using a Perceiver and jointly optimizes contrastive alignment and report generation. The remaining baselines represent visual self-supervision and supervised pretraining. GigaPath (Xu et al., 2024) uses a LongNet-based slide encoder pretrained through masked autoencoding. FEATHER (Shao et al., 2025) learns transferable slide representations through attention-based multiple instance learning and pan-cancer classification. We consider its CONCH v1.5-based FEATHER-24K variant, pretrained on approximately 24,000 slides across 108 morphological categories.

Table S10: Logistic-regression results on the selected 12-task set. Values are mean±standard deviation. The best and second-best means within each task-metric row are shown in bold and underlined, respectively.

| Task | Metric | Mean-pooling | FEATHER | GigaPath | PRISM | CARE | TITAN | ProSlide |
|---|---|---|---|---|---|---|---|---|
| Task 1 | ACC | 54.52±0.04 | 54.88±0.07 | 55.98±0.04 | 55.93±0.11 | 56.72±0.02 | 56.29±0.07 | **61.65±0.20** |
| | AUC | 70.32±0.13 | 73.41±0.10 | 72.71±0.05 | 72.29±0.10 | 73.54±0.08 | 72.56±0.06 | **76.09±0.07** |
| Task 2 | ACC | 60.95±0.04 | 54.74±0.03 | 61.44±0.13 | 57.55±0.15 | 61.44±0.07 | 59.83±0.14 | **62.87±0.16** |
| | AUC | 84.84±0.06 | 84.28±0.05 | 86.57±0.01 | 86.49±0.01 | **88.89±0.02** | 85.98±0.01 | 86.89±0.03 |
| Task 3 | ACC | 57.50±0.83 | 58.14±1.61 | 55.18±0.65 | 53.98±1.13 | 58.70±1.03 | **61.20±1.43** | 60.39±1.14 |
| | AUC | 62.61±2.92 | 58.23±3.68 | 60.08±2.79 | 59.52±2.78 | 65.83±2.61 | **73.16±2.08** | 67.69±2.31 |
| Task 4 | ACC | 58.30±0.08 | 52.03±0.03 | 51.57±0.16 | 57.64±0.05 | 56.87±0.10 | 57.35±0.16 | **58.48±0.32** |
| | AUC | 70.87±0.08 | 67.90±0.19 | **81.96±0.02** | 72.58±0.01 | 72.68±0.05 | 71.24±0.01 | 76.00±0.03 |
| Task 5 | ACC | 63.06±0.97 | 65.95±1.34 | **68.57±1.03** | 63.10±1.20 | 57.51±1.07 | 59.82±0.89 | 65.63±0.73 |
| | AUC | 69.57±1.27 | 73.05±1.27 | **75.65±1.72** | 72.17±1.15 | 62.10±1.56 | 68.68±1.27 | 73.82±1.15 |
| Task 6 | ACC | 63.11±1.51 | 59.65±1.25 | 62.04±0.86 | 54.77±1.21 | 64.52±0.76 | 62.79±1.00 | **64.54±1.29** |
| | AUC | 69.14±1.51 | 63.85±1.21 | **72.16±1.39** | 59.28±1.51 | 66.60±1.13 | 68.14±1.09 | 68.38±1.43 |
| Task 7 | ACC | 74.63±0.89 | 67.54±0.85 | 67.05±0.85 | 73.18±0.96 | 74.08±0.74 | **78.22±0.67** | 75.39±0.90 |
| | AUC | 81.23±1.02 | 75.60±0.65 | 74.68±0.98 | 79.84±0.85 | 83.19±0.56 | **87.70±0.44** | 84.43±0.79 |
| Task 8 | ACC | 70.16±0.12 | 74.02±0.23 | 71.11±0.61 | 74.50±0.16 | 74.39±0.12 | **74.52±0.09** | **74.52±0.21** |
| | AUC | 82.85±0.23 | 86.69±0.16 | 87.51±0.21 | 88.24±0.10 | **88.40±0.08** | 88.22±0.08 | 87.50±0.10 |
| Task 9 | ACC | 69.97±0.54 | 69.46±0.49 | **74.17±0.73** | 69.34±0.53 | 72.25±0.50 | 68.94±0.64 | 72.93±0.65 |
| | AUC | 76.81±0.69 | 76.47±0.59 | **81.14±0.62** | 77.32±0.56 | 79.75±0.48 | 76.58±0.60 | 79.55±0.70 |
| Task 10 | ACC | 52.40±0.16 | 59.90±0.07 | 49.41±0.44 | 58.17±0.01 | 57.53±0.28 | **60.90±0.21** | 60.53±0.15 |
| | AUC | 72.65±0.06 | **77.39±0.06** | 69.59±0.36 | 75.66±0.01 | 76.83±0.13 | 77.24±0.13 | 77.28±0.14 |
| Task 11 | ACC | 52.50±0.00 | 55.28±0.06 | 57.91±0.10 | **61.45±0.01** | 52.46±0.00 | 55.12±0.10 | 53.51±0.01 |
| | AUC | 78.06±0.01 | 76.76±0.12 | **80.83±0.01** | 78.47±0.01 | 77.00±0.01 | 76.60±0.02 | 79.10±0.11 |
| Task 12 | ACC | 63.56±0.08 | 56.22±0.26 | 54.22±0.14 | 58.34±0.09 | 65.53±0.03 | 63.75±0.07 | **65.69±0.01** |
| | AUC | 73.70±0.03 | 68.48±0.28 | 62.16±0.07 | 67.98±0.10 | 74.37±0.00 | 74.18±0.02 | **75.65±0.01** |

### D.3 Detailed Downstream Results

To examine performance beyond the aggregate comparisons, we provide task-level classification results.

**Task-level performance.** Table S10 reports the complete linear-probing results across the 12 downstream tasks. ProSlide has achieved competitive performance across molecular prediction and histological subtyping, with the highest ACC and AUC on both SETD2 mutation prediction and cross-cohort fine-grained lung subtyping. Its relative performance has nevertheless varied across endpoints, with other encoders leading on several CPTAC molecular tasks. Cross-cohort binary lung subtyping has also revealed a discrepancy between its competitive AUC and weaker ACC, highlighting the importance of considering both metrics when assessing transferability.

### D.4 Data Efficiency and Inference Memory Usage

**Pretraining data efficiency.** To examine downstream performance in relation to slide-level pretraining scale, we compare the evaluated models in Figure S1a. ProSlide achieves the highest aggregate performance among these models with 2,229 WSIs for slide-level pretraining. This comparison highlights the potential of protein-derived supervision to support slide representation learning with limited paired data. The reported WSI counts refer specifically to slide-level pretraining; ProSlide also draws on pretrained CONCH v1.5 features and PFE-derived proteomic knowledge.

**Inference memory usage.** To examine inference memory requirements across slide sizes, we measure GPU memory usage as the number of input patches increases (Figure S1b). ProSlide exhibits modest memory growth over the evaluated range. Its memory usage remains below that of TITAN, PRISM, and GigaPath, and falls below CARE at larger patch counts. FEATHER has the

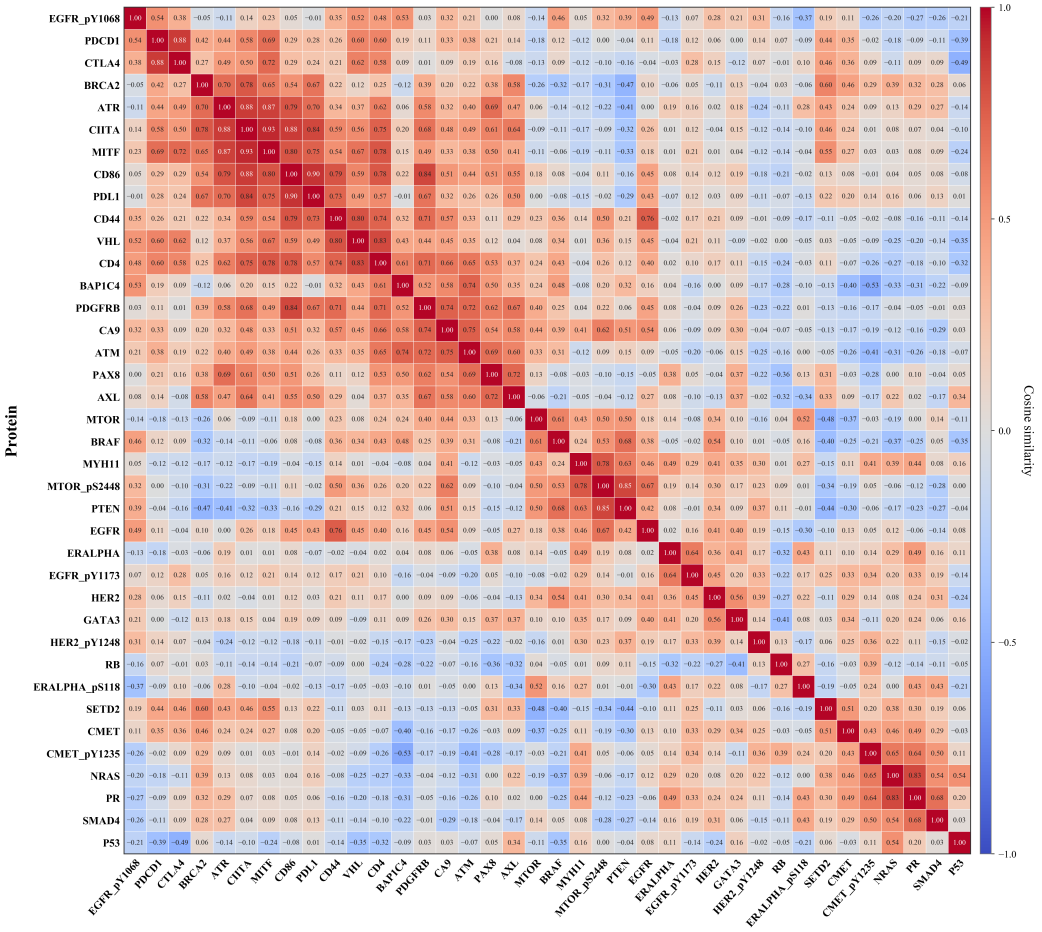



Figure S2: **PFE protein identity embeddings.** Pairwise cosine similarities among 38 learned protein identity embeddings. Both axes follow the same clustering order, with warmer colors indicating higher similarity.

lowest memory usage among the evaluated models. ProSlide's memory behavior is consistent with its hierarchical aggregation, which summarizes patch features into region representations before slide-level attention.

### D.5 Analysis of PFE Protein Identity Embeddings

To assess whether the learned protein identity embeddings reflect known biological relationships, we compute pairwise cosine similarities among the 38 target embeddings and visualize them in a clustered heatmap (Figure S2). These learned input embeddings are shared across samples and provide the identity component of each protein token. Both axes follow the same clustering order to facilitate inspection of the resulting similarity structure.

The heatmap has shown high similarity between PDCD1 and CTLA4, both of which encode inhibitory immune-checkpoint receptors involved in T-cell regulation (Pardoll, 2012). BRCA2 and ATR also exhibit relatively high similarity. Their functional connection involves DNA damage signaling and repair, with ATM/ATR-mediated PALB2 phosphorylation promoting RAD51 function within the PALB2–BRCA2 repair pathway (Ahlskog et al., 2016). These examples illustrate a qualitative correspondence between selected embedding similarities and known biological relationships.