Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The STP Open Challenge: A Community Effort to Define and Benchmark Spatial Transcriptome-to-Proteome Prediction

Jiguang Wang

jgwang@ust.hk

The Hong Kong University of Science and Technology https://orcid.org/0000-0002-6923-4097

Yimeng Qiao

The Hong Kong University of Science and Technology

Zijie Fang

The Hong Kong University of Science and Technology

Yixuan Li

Southern University of Science and Technology

Lingzhe Xu

Tsinghua University

Fuhao Guo

Nankai University

Mingyue Wei

Shandong University

Lang Wu

Hunan Normal University

Jihong Tang

The Hong Kong University of Science and Technology

Ran Liu

The Hong Kong University of Science and Technology

Yuyan Ruan

The Hong Kong University of Science and Technology

Zhidong Yang

The Hong Kong University of Science and Technology

Hongxin Wei

Southern University of Science and Technology

Shimin Shuai

Southern University of Science and Technology https://orcid.org/0000-0002-9527-8018

Yongbing Zhang

Harbin Institute of Technology (Shenzhen) https://orcid.org/0000-0003-3320-2904

Kai Zhang

Tsinghua University

Gang Hu

Nankai University

Zhiping Liu

Shandong University https://orcid.org/0000-0001-7742-9161

JianCheng Zhong

Hunan Normal University

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10262522/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

The STP Open Challenge: A Community Effort to Define and Benchmark Spatial Transcriptome-to-Proteome Prediction

Yimeng Qiao[1], Zijie Fang[1], Yixuan Li[2], Lingzhe Xu[3], Fuhao Guo[4], Mingyue Wei[5], Lang Wu[6], Jihong Tang[1], Ran Liu[1], Yuyan Ruan[1], Zhidong Yang[1], Hongxin Wei[2], Shimin Shuai[7], Yongbing Zhang[8], Kai Zhang[3], Gang Hu[4], Zhiping Liu[5], Jiancheng Zhong[6], Jiguang Wang[1,9*]

[1]Division of Life Science, Department of Chemical and Biological Engineering, State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[2]Department of Statistics and Data Science, College of Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China

[3]Open FIESTA Center, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

[4]NITFID, School of Statistics and Data Science, AAIS, LPMC, and KLMDASR, Nankai University, Tianjin 300071, China

[5]Department of Biomedical Engineering, Shandong University, Jinan, Shandong 250061, China

[6]College of Information Science and Engineering, Hunan Normal University, Changsha 410081, Hunan, China

[7]Department of Human Cell Biology and Genetics, School of Medicine, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China

[8]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China

[9]SIAT-HKUST Joint Laboratory of Cell Evolution and Digital Health, HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China

* Corresponding to jgwang@ust.hk

Abstract

While spatial transcriptomics technologies are increasingly accessible, spatial proteomics remains constrained by high costs and limited throughput. Furthermore, poor mRNA-protein correlation, driven by post-transcriptional regulation, necessitates multi-omics integration for complementary biological insights. Current RNA-to-protein prediction models, primarily trained on bulk or single-cell data, often overlook critical spatial context. Consequently, it remains unclear which model architecture, spatial representation, or histopathological features best facilitate spatial protein abundance inference. To address these gaps, we defined the STP (Spatial Transcriptome-to-Proteome) task and launched the STP Open Challenge, a competition for predicting spatial proteomics from spatial transcriptomics of adjacent tissue sections. Utilizing a curated paired spatial dataset and a standardized evaluation framework, the challenge attracted 101 teams and yielded 896 valid submissions. Top-performing methods achieved strong predictive performance while preserving spatial organization. The landscape of submitted models revealed diverse modeling strategies and motivated us to develop a consensus-based ensemble framework that leverages various participant methods to enhance prediction robustness. Extended evaluation on unseen datasets demonstrated that the top-performing models outperformed existing baselines, highlighting the effectiveness of the community effort for this task. By formally defining the STP task and providing a standardized benchmarking platform, this work establishes a foundation for scalable, cost-effective spatial proteomics prediction and future multi-omics integration.

Introduction

Recent advances in spatial transcriptomics (ST) have enabled high-resolution mapping of gene expression within intact tissues. While ST captures gene expression within its spatial context, spatial proteomics (SP) often provides a more direct representation of cellular function. Due to post-transcriptional regulation, translation efficiency, and protein degradation, RNA expression is often poorly correlated with protein levels[1,2]. This discrepancy motivates the integration of ST and SP to obtain a more comprehensive view of tissue organization, providing a complementary link to cellular phenotype and function.

Technologies that jointly profile RNA and protein expression within the same tissue section have begun to emerge, including Stereo-CITE-seq[3], spatial-CITE-seq[4], and spatial-Mux-seq[5]. However, these platforms remain constrained by throughput, cost, molecular coverage and technical variability. As a result, most studies still rely on separate measurements from adjacent tissue sections[6,7], creating a practical need for computational approaches that can infer spatial proteomics based on spatial transcriptomic data. From a biological perspective, constructing a predictive model that captures the nonlinear relationship between ST and SP could help elucidate RNA–protein correlations, overcome proteomic bottlenecks, and mark a major step toward the Virtual Cell challenge. From an economic perspective, such approaches could substantially reduce the cost of spatial profiling, particularly given the high expense associated with SP measurements.

At the single-cell level, RNA-to-protein prediction has recently been studied, with methods such as cTP-net[8], sciPENN[9] and scTranslator[10] demonstrating the feasibility of inferring protein abundance from transcriptomic profiles. However, these methods were primarily developed for dissociated single-cell data and do not explicitly model spatial context. In intact tissues, protein abundance may be shaped not only by intracellular transcriptional and translational programs, but also by intercellular interactions and microenvironmental factors. More recently, spatially aware models such as NicheTrans[11] and DGAT[12] have begun to incorporate neighborhood expression, spatial graph structure, or histological images for ST-to-SP prediction, based on the hypothesis that spatial context and morphological information provide complementary signals beyond local gene expression in determining protein profiles.

Despite growing interest in ST-to-SP prediction, the field lacks a standardized framework for evaluating prediction methods. Existing studies differ in datasets, preprocessing strategies, target protein panels, evaluation metrics, and validation settings, making it difficult to determine whether performance gains arise from transcriptomic signal, spatial context, histological morphology, or model architecture. As a result, several fundamental questions remain unresolved: which modeling strategies are most effective, which proteins are predictable, and how well models generalize across samples, diseases and experimental platforms.

Community-driven challenges provide a powerful paradigm for advancing

computational biology, enabling standardized task definition, rigorous results evaluation, and rapid method development across diverse research groups. While such efforts have accelerated progress in fields such as protein structure prediction[13] and single-cell analysis[14], no comparable community benchmark currently exists for ST-to-SP prediction, an emerging task at the intersection of spatial omics and multimodal/ multiomics learning.

To address this gap, we introduced the STP (Spatial Transcriptome-to-Proteome) prediction task and organized the STP Open Challenge, a community effort to systematically evaluate computational approaches for ST-to-SP prediction. We generated paired spatial transcriptomic–proteomic datasets from adjacent tissue sections and established a standardized evaluation framework on Codabench[15] platform to ensure fair comparison across methods. The challenge attracted more than 100 teams and yielded 896 valid submissions in 2025. Top-performing methods achieved strong predictive performance while preserving spatial organization, demonstrating the feasibility of STP prediction and also highlighting the remaining challenges. Analysis of the submitted methods highlighted spatial context as a key component of successful prediction, while also showing the benefits of histological images and pretrained representations. At the marker level, we further showed that protein predictability was mainly associated with protein abundance and spatial organization, motivating a consensus-based ensemble strategy that improves robustness across markers. Finally, evaluations on unseen datasets revealed robust performance under matched experimental settings. Together, this work provides a valuable benchmark resource and conceptual framework for developing robust spatial multi-omics integration methods.

Results

The STP Open Challenge

We organized the STP Open Challenge to systematically benchmark computational approaches for predicting spatial proteomics from spatial transcriptomics data. Beyond identifying models with optimal predictive accuracy, the challenge was also designed to evaluate how spatial context and additional modalities, such as histological images, contribute to model performance.

The challenge was hosted on the Codabench platform and conducted in two phases (**Fig. 1c**). During the development phase, participants were provided with 166,184 spatial bins containing paired transcriptomic and proteomic measurements for model training, along with an additional 8,786 RNA-only bins from distinct spatial regions of the same sample for validation and model selection. In the testing phase, submitted models were evaluated on an independent held-out sample from the same patient, comprising 84,376 bins generated under the same experimental conditions. This design enabled standardized and unbiased evaluation of model performance on unseen tissue data.

To enable evaluation under realistic experimental conditions, we generated paired spatial multi-omics datasets using an adjacent-section design (**Fig. 1a**). Spatial transcriptomics data were acquired using Visium HD[16], capturing 18,085 genes, while spatial proteomics measurements were obtained from adjacent tissue sections using CODEX[17], targeting 44 protein markers. The two modalities were aligned using an AI-based deformable image registration framework (**Methods, Extended Fig. 1a**), allowing gene expression and protein abundance to be mapped into a shared coordinate system. We then formulated the STP prediction task as a regression problem, in which protein abundance at each spatial location is inferred from binned transcriptomic profiles (**Fig. 1b, Methods**). Participants were invited to develop models to address this task.

For each spatial bin, the released input features included expression profiles of 18,085 genes, spatial coordinates, and the corresponding Hematoxylin and Eosin (H&E) stain image on the same slide acquired before Visium HD profiling (**Fig. 2a**). Participants were tasked to predict the abundance of 44 target protein markers at matched spatial locations in the adjacent slide. The registration of the two slides are in good quality, supported by the observation that the close alignment of the overall tissue contours and major anatomical structures between the Visium HD and CODEX images after registration (**Fig. 2b, Extended Fig. 1b**). As a representative example, we visualized a registered perivascular lymphocytic cuffing (PLC) region from the independent testing sample. This structure is characterized by lymphocyte enrichment surrounding a blood vessel and is feasible to identify in both H&E and CODEX images. The concordant spatial organization observed across modalities demonstrates the high-quality alignment between the Visium HD and CODEX datasets (**Fig. 2c**). We further

characterized the challenge datasets by examining transcriptomic input quality, protein target distributions, and consistency **Extended Fig. 2** and **Supplementary Table 1**.

Model performance was assessed using Spearman correlation coefficient (SCC) between predicted and measured protein abundance (**Methods**). For each submission, the leaderboard score was defined as the mean SCC across the ten best-predicted protein markers, thereby reducing the impact of low-signal markers on ranking. To ensure a fair and reproducible comparison, the challenge organizers executed all submitted code on the held-out testing sample using the same computational server. Based on final leaderboard performance and complementary strengths across protein markers, we selected the top six teams for further analysis.

Leaderboard performance reveals recurrent principles of effective model design

The performance of the top six teams in the STP Open Challenge is summarized in **Fig. 2d-g**. The best-performing algorithm achieved a mean SCC of 0.58 across its 10 best-predicted protein markers, indicating that spatial protein abundance can be inferred from spatial transcriptomics data with relatively high accuracy and supporting the feasibility of the STP prediction task. Beyond the primary leaderboard score (**Fig. 2d**), we further examined the distribution of SCC scores across all 44 protein markers for each team (**Fig. 2e**). These results were broadly consistent with the final leaderboard rankings, while also revealing substantial marker-specific variability in prediction performance. To assess spatial consistency, we computed the bivariate Moran's Index to quantify the spatial correlation between predicted and measured protein abundance (**Fig. 2f-g**). This analysis showed that the top-performing models were also able to reconstruct spatial organization, largely agreeing with the SCC-based ranking, except for teams 5 and 6, which exhibited slightly different relative performance.

We next examined the modeling strategies used by the top-performing teams (**Table 1**). Despite substantial variation in model architecture, preprocessing, post-processing, and training procedures, several recurring design principles emerged.

(1) Spatial context is critical for STP prediction . All top-performing teams explicitly incorporated spatial information, indicating that neighborhood context plays an important role in predicting spatial proteomics. We identified four major strategies for modeling spatial information. First, some methods directed encoded spatial coordinates within Transformer architectures, enabling models to learn positional dependencies. Second, several approaches constructed neighborhood graphs to propagate information across spatially adjacent bins. Third, K-nearest neighbor (KNN)-based methods were used to impose spatial smoothing constraints during training. Fourth, some methods directly incorporated neighboring transcriptomic profiles as additional inputs. Despite methodological differences, spatially aware models consistently outperformed expression-only baselines on the held-out test sample, suggesting that protein expression is shaped not only by intracellular signals but also by the microenvironment.

(2) Histological image information provided substantial benefits. In addition to transcriptomic and spatial features, four of the top six teams incorporated histological image features into their modeling pipelines. These approaches adopted pretrained pathology foundation models, including UNI[18], UNI-v2, CONCH[19,] and Virchow[20], aiming to capture tissue morphological features that are not directly encoded by gene expression, under the assumption that image-derived features provide complementary information for STP prediction. However, the benefit of image integration varied across submission. Notably, Team 2 used a relatively simple convolutional architecture (ResNet-50[21]), yet achieved better performance than other methods relying on large pretrained foundation models. One possible explanation is that current pathology foundation models are often trained for classification or general representation learning tasks, which may not align optimally with molecular prediction objectives. In contrast, lightweight models trained end-to-end on task-specific data may better capture features relevant to STP prediction while reducing overfitting or domain mismatch.

(3) Pretrained transcriptomic representations enhanced feature learning. Several teams leveraged pretrained single-cell or spatial transcriptomics foundation models, such as SToFM[22] and scConcept[23] as feature encoders. These approaches achieved strong performance, suggesting that large-scale pretraining on transcriptomic data can capture transferable biological prior knowledge that benefits downstream cross-modal prediction tasks. This finding highlights the potential value of representation learning in spatial multi-omics integration.

(4) Model complexity is not the sole determinant of performance. Interestingly, not all top-performing approaches relied on deep learning. Team 6 used classical machine learning methods, and achieved superior performance for specific protein markers, demonstrating that simpler models can remain competitive under certain conditions (**Extended Fig. 3**). For example, when the target marker has favorable signal properties or when the input features are sufficiently informative.

Together, these observations underscore the importance of spatial information in STP prediction, while also suggesting that the benefits of histological image features, pretrained representations, and more complex model architectures depend on the specific dataset and modeling strategy. The diversity of successful methods further indicates that different models may capture complementary predictive signals for the STP task, providing a rationale for combining their predictions.

Protein predictability is associated with marker abundance and spatial organization

We next analyzed model performance at the level of individual protein markers and sought to identify factors associated with marker-specific predictability in the STP task. Protein-specific SCC scores from the top six teams, together with RNA-protein correlations, were used to cluster markers, revealing heterogeneity in predictability across protein markers (**Fig. 3a**). Overall, RNA-protein correlations were lower than predicted-measured protein correlations, further highlighting the strong performance of

the top-six leaderboard models. Notably, the relative rankings of these six models varied across markers, indicating that the prediction difficulty is protein-dependent rather than uniform across the task.

To better understand this variability, we examined several potential determinants of protein predictability, including RNA-protein correlation, protein abundance, detection prevalence and measured protein spatial autocorrelation. Spatial autocorrelation was defined as the extent to which protein abundance at nearby spatial locations exhibits similar patterns and was quantified using bivariate Moran's I. Contrary to the expectation that proteins with stronger RNA-protein coupling would be easier to infer, RNA-protein correlation and marker detection prevalence showed only limited association with prediction performance (**Extended Fig. 4a-b**). In contrast, protein abundance and spatial autocorrelation were strongly positively associated with marker predictability (**Fig. 3b-c**).

A possible explanation of this observation is that protein abundance reflects the strength and reliability of the measured proteomic signals, while spatial autocorrelation represents the degree to which the signals are organized into coherent tissue-level patterns. Low-abundance markers are more likely to be affected by sparse detection and measurement noise, whereas weakly structured markers provide limited spatial patterns for models to learn. In contrast, abundant markers provide denser and more reliable supervision signals, and spatially organized markers represent coherent tissue-level patterns that align with the inductive biases of spatially aware models that explicitly incorporated neighborhood aggregation, graph propagation and coordinate-based regularization.

The above associations motivated us to define a marker-level signal abundance-structure axis associated with STP predictability (**Fig. 3d**). We stratified protein markers into four abundance-structure regimes using protein abundance levels and measured protein Moran's I, and observed that proteins with higher abundance levels and high Moran's I are almost consistently better predicted. For example, IDH1 showed enriched protein expression and coherent spatial structure, providing both reliable target supervision signal and clear spatial autocorrelation patterns for model prediction (**Extended Fig. 5a**). By comparison, markers with high abundance but weak spatial autocorrelation showed intermediate predictability (**Extended Fig. 5b**). Although weak spatial organization limits the ability of models to leverage neighborhood information, their abundant protein signal might support estimation of overall expression levels. Meanwhile, markers with low abundance but strong spatial autocorrelation also present a variable predictability regime (**Extended Fig. 5c**). The local spatial expression enrichment could provide learnable patterns for spatially aware models, but the sparse detection and weak signal intensity make these proteins harder to infer. The least favorable markers typically exhibit low abundance and minimal spatial patterns, leading to consistently poor performance across all methods (**Extended Fig. 5d**). Such patterns reflect a combination of low data quality, high noise levels and limited

biological signal. Thus, the abundance-structure axis provides an interpretable framework for understanding the diversity of marker-level predictability.

Top-performing methods generalize across unseen matched-platform datasets and motivate marker-aware ensemble strategy

To assess whether the top-performing models could generalize beyond the original challenge test sample, we further evaluated the top-six approaches on three independent datasets that were not used during the STP Open Challenge (**Methods**). These datasets were generated using the same experimental protocol as the challenge data but represented distinct biological conditions, including both samples from the same disease and samples from a different disease. This design allowed us to evaluate model robustness and generalization under matched technical conditions while introducing biological variation. For comparison, we further evaluated these submissions with existing baselines, including spatially aware STP models, expression-only single-cell RNA-to-protein models, as well as a customized traditional regression model (**Methods**).

All the compared methods are trained on the challenge training set, and transferred to the evaluation datasets. As shown in **Fig. 4a**, top-performing submissions consistently outperformed state-of-the-art (SOTA) STP prediction baselines across the challenge test sample and all three unseen samples, indicating that the modeling strategies identified through the challenge were not restricted to a single dataset or experimental instance. Notably, performance improvement was observed in both same-disease (unseen sample A) and cross-disease (unseen samples B and C) settings, suggesting that the models were not merely optimized for the original challenge sample but could transfer to related yet unseen biological contexts under the same experimental platform.

Nevertheless, model performance remained marker- and dataset-dependent, and no single model performs optimally across all protein markers. While the top-ranked methods achieved strong performance on average, different models exhibit distinct strengths across markers. Lower-ranked models even outperform top methods for specific proteins, highlighting complementary capabilities among approaches. Motivated by these observations, we constructed a consensus-based ensemble model that integrated predictions from multiple submissions (**Fig. 4b, Methods**) and evaluated it on the challenge datasets and unseen datasets.

Encouragingly, the ensemble model achieved the best overall performance across all unseen datasets as well as the challenge test dataset, supporting the idea that different top-performing methods capture complementary predictive signals. By aggregating predictions from diverse models, the ensemble likely reduced method-specific variability and improved robustness across samples. These results suggest that model diversity captures distinct aspects of the STP prediction task, and that ensemble strategies provide a principled approach to addressing heterogeneity in protein predictability.

To further examine marker-level instability, we analyzed proteins with the largest performance decreases from the developmental phase to the final test phase and compared their abundance distributions between the training and test samples. Several markers with substantial validation-to-test performance drops showed noticeable distributional differences across samples, suggesting that marker-specific dataset shift may contribute to unstable prediction performance even under matched technical conditions (**Extended Fig. 6**).

Together, these results show that top-performing challenge submissions transferred well across unseen matched-platform datasets, while marker-level complementarity and distributional shifts motivated ensemble strategies and highlighted remaining barriers to robust STP prediction.

Discussion

Here we introduced the STP Open Challenge as a community-driven effort to define and search for optimal methods for the task to predict spatial proteomics from spatial transcriptomics. We generated paired spatial transcriptomic-proteomic datasets, established a standardized evaluation protocol, and hosted an open competition framework, providing a benchmark resource for systematic comparison of computational approaches to address the task. Beyond ranking the submitted methods, the challenge enabled us to examine the model design strategies, marker-level predictability and robustness under controlled evaluation settings.

The challenge leaderboard demonstrates that STP prediction is feasible but remains challenging. Top-performing models outperformed existing baselines and achieved the best average Spearman correlation of the top 10 predicted protein abundance at 0.58. These results suggest that spatial transcriptomic profiles could serve as informative, although incomplete proxies for targeting spatial proteomics. Meanwhile, substantial variation across protein markers and methods highlights the incomplete correspondence between RNA and protein abundance, as well as the marker-dependent nature of STP prediction.

Our results highlight several key observations of STP modeling strategies. First, spatial context was widely used by top-performing methods, suggesting that neighborhood information and tissue organization are important for predicting protein abundance beyond local transcriptional states alone. Second, multimodal integration with histological images provides additional information, although its benefits remain sensitive to model design and data distribution. Third, pretrained transcriptomic feature representations might improve the model prediction accuracy by utilizing prior knowledge contained in large-scale datasets. Lastly, model complexity is not the sole factor that determines the model performance. These findings serve as practical guidelines for the STP modeling design.

Diving into the protein-wise prediction results of each team, we observe substantial heterogeneity in protein markers' predictability, which is positively correlated with protein abundance and measured spatial autocorrelation. These factors define an abundance-structure axis that reflects both the reliability of the measured protein signal and the spatial organization patterns. This perspective indicates that STP prediction is constrained by both biological regulation and technical signal quality, rather than being a uniform regression problem. Instead, each protein marker may present a distinct level of predictability.

Evaluation on unseen datasets further showed that top-performing methods achieved strong generalization performance across related biological conditions. These datasets were generated under the same experimental protocol as the challenge data but included both same-disease and cross-disease samples, allowing us to assess the models' generalizability. The top-performing methods trained on the challenge dataset could transfer well to the unseen samples and were consistently superior to the state-of-the-

art baselines, indicating that challenge-derived models are not limited to the original sample but could be generalized even to unseen biological contexts.

Importantly, no single modeling paradigm uniformly outperforms others across all proteins or datasets. Instead, different methods exhibited complementary strengths, with individual models performing better for different markers and data regimes. The consensus-based ensemble model further provided a practical strategy to improve robustness and accuracy. These findings underscore the value of methodological diversity and suggest that future advances might benefit from combining multiple modeling strategies together.

This study has several limitations. First, the primary challenge data remain limited in sample size and experimental platforms, Expanding the benchmark to larger cohorts, more tissue types, and additional disease contexts and platforms will be essential for future benchmarking efforts. Second, the current evaluation also focuses primarily on prediction accuracy and spatial consistency, whereas future benchmarks should further assess downstream biological utility, such as cell state annotation and spatial domain identification. In addition, although the challenge provides a standardized evaluation framework, further work is needed to understand how different preprocessing choices, feature representations and training strategies affect model performance across platforms.

Despite these limitations, the STP Challenge establishes a standardized framework for evaluating spatial transcriptome-to-proteome prediction and provides a comprehensive resource for the community. Beyond ranking individual submissions, the challenge reveals key factors associated with marker predictability, model design and generalization in this task. We anticipate that this effort will facilitate the development of more robust and generalizable STP prediction methods and contribute to the broader goal of integrating spatial multi-omics data for understanding tissue organization and cellular function.

Methods

Spatial transcriptome-to-proteome dataset generation and preprocessing

Paired spatial transcriptome-to-proteome datasets were generated using an adjacent-section spatial multi-omics workflow (**Extended Fig. 1a**). Spatial transcriptomic profiles were acquired using the Visium HD platform, capturing 18,085 genes at 2 μm-bin resolution, whereas spatial proteomic measurements were obtained from adjacent tissue sections using CODEX, targeting 44 protein markers with subcellular spatial resolution.

To enable cross-modality learning, deformable image registration with a deep learning model named DeepHistReg[24] was performed between the two modalities. Briefly, haematoxylin and eosin (H&E)-stained images corresponding to the spatial transcriptomic sections were registered to the DAPI channel of the spatial proteomic data following image preprocessing steps including intensity normalization, image rescaling, and data augmentation. This procedure accounted for nonlinear tissue deformation introduced during tissue sectioning, staining and imaging. After registration, spatial coordinates from both modalities were mapped into a shared coordinate system.

Gene expression and protein abundance matrices were subsequently extracted from aligned spatial locations. Raw transcriptomic counts were normalized and log-transformed while protein expression values were log-transformed prior to downstream analysis. To harmonize spatial resolution across modalities, transcriptomic and proteomic profiles were aggregated into fixed-size spatial bins (16 μm bins in the challenge datasets and 26 μm bins in the unseen datasets).

The resulting aligned datasets consisted of paired transcriptomic and proteomic profiles at corresponding spatial locations, providing the foundation for supervised learning in the STP prediction task.

Definition of the STP prediction task

The STP prediction task was defined as predicting spatial proteomic abundance profiles from spatial transcriptomic measurements at aligned spatial locations. Formally, for each spatial location i, the input consists of a gene expression vector $x_i \in \mathbb{R}^G$, where G denotes the number of measured genes. The target is a protein abundance vector $y_i \in \mathbb{R}^P$, where P denotes the number of protein markers. In the challenge dataset, $G = 18{,}085$ and $P = 44$.

This formulation corresponds to a supervised regression problem at the spatial bin level. Participants were tasked to infer protein abundance from transcriptomic features alone or incorporating additional modalities, including spatial context and histological image

information. The predicted protein profiles were then evaluated against the measured CODEX-derived protein profiles at the matched spatial locations.

Challenge design and evaluation protocol

The STP Open Challenge was designed to simulate realistic model development settings while ensuring standardized and unbiased evaluation across participating teams. The challenge was hosted on the online evaluation platform Codabench, where participants were provided with paired spatial transcriptomic and proteomic profiles from a development dataset for model training and optimization. A separate validation dataset containing spatial transcriptomic profiles with masked proteomic measurements was used to provide leaderboard feedback during the challenge period, enabling iterative model refinement without exposing ground-truth protein measurements.

For each submission, participants were required to predict all target protein markers for every spatial bin in the validation or testing datasets. Submitted predictions were automatically evaluated against held-out CODEX-derived measurements using predefined quantitative metrics. All teams were evaluated on the same held-out spatial locations, target protein markers and evaluation criteria under a unified evaluation pipeline. Diverse modelling strategies were permitted, including expression-only, spatially aware and morphology-integrated approaches.

Final rankings were determined based on performance on the held-out evaluation dataset. In addition to the primary leaderboard metric, post-challenge analyses were performed using complementary evaluation metrics to assess both expression-level prediction accuracy and preservation of spatial organization. Together, this framework enabled a community-wide comparison of diverse STP prediction strategies under a unified and reproducible benchmarking setting.

Top-performing submitted methods

The best-performing algorithm developed a three-branch multimodal framework that integrated single-cell foundation model fine-tuning, spatial neighborhood aggregation and histology-guided representation learning. The first branch fine-tuned scConcept on rank-encoded gene tokens from neighboring spots, with UNI-v2-derived H&E features projected into the RNA feature space as morphology prompts. The second branch used an LSTM to model absolute RNA expression patterns that may be lost during rank-based encoding, whereas the third branch directly incorporated histological features as complementary morphological signals. The three representations were aligned using an InfoNCE-based contrastive objective, fused for protein prediction and further improved through ensembling of model variants trained under different data usage schemes.

The second-performing algorithm proposed a multimodal graph neural network for spatial protein expression prediction that integrates spatial transcriptomics (RNA) and whole-slide histology images (H&E). RNA profiles are normalized and projected into a latent space, while H&E patches are encoded using a pretrained ResNet-50. In

addition, cell-type abundance vectors estimated from RNA provide complementary biological context. These modality-specific features are fused as node representations on a spatial graph constructed via a radius-constrained k-nearest neighbor scheme and processed with GATv2[25] layers. To enhance robustness, it incorporates biological priors into the training objective by enforcing smoothness among cells with similar compositions and consistency among co-expressed proteins.

The third-performing algorithm developed a multimodal framework for spatial protein prediction by integrating RNA expression, H&E image features, and spatial neighborhood information. RNA expression served as the primary predictive signal and was standardized before combined with a spatial context representation derived from k-nearest-neighbor statistics and low-dimensional projection. For histology, multi-scale H&E patches were encoded using an ImageNet-pretrained ResNet18 backbone to capture complementary morphological information. The model backbone was built as an attention-augmented residual MLP, where RNA and spatial features formed the main prediction branch, while H&E features were incorporated through conditional modulation and gated refinement. A correlation-aware output head was further used to model inter-protein dependencies. The model was trained with a combined regression objective, and the final predictions were refined by seed ensembling and spatial smoothing.

Ensemble learning strategy

To investigate whether predictions from independently developed models captured complementary signals, we constructed an unsupervised consensus-based ensemble model using the top-performing submissions. Before learning ensemble, predicted protein expression values from each team were standardized for each protein marker, yielding $z_{i,t,p}$, where i denotes the spatial bin, t denotes the submitted model, and p denotes the protein marker.

For each protein marker, we first computed a model-level consensus score based on the agreement between one model and all other selected models. Specifically, let $\rho_{p,t,s}$ denote the Spearman correlation coefficient (SCC) between the predictions of models t and s for protein p across all spatial bins. The consensus score for model t on protein p was defined as

$$a_{p,t} = \frac{1}{T-1}\sum_{s \neq t} \rho_{p,t,s},$$

where T is the total number of selected models. A higher $a_{p,t}$ indicates stronger agreement with the remaining models.

The consensus scores were then converted into protein-specific ensemble weights using a SoftMax transformation:

$$w_{p,t} = \frac{\exp(a_{p,t}/\tau)}{\sum_{u=1}^{T} \exp(a_{p,u}/\tau)},$$

where τ is a temperature parameter controlling the sharpness of the weight distribution. Smaller values of τ assign larger weights to models with higher consensus scores, whereas larger values produce more uniform weights.

The consensus-weighted ensemble prediction for each spatial bin and protein marker was then computed as

$$\tilde{y}_{i,p} = \sum_{t=1}^{T} w_{p,t}\, z_{i,t,p}.$$

We further applied a post hoc spatial smoothing step to examine whether local spatial regularization could improve prediction stability. Let $N_k(i)$ denote the set of k-nearest spatial neighbors of bin i, including itself. The smoothed prediction was calculated as

$$\hat{y}_{i,p} = \frac{1}{|\, N_k(i)\,|} \sum_{j \in N_k(i)} \tilde{y}_{j,p}.$$

In this study, we used $k = 15$ for spatial smoothing analysis. Thus, the final ensemble strategy first aggregated standardized model predictions using SCC-based consensus weights and then smoothed the resulting predictions over local spatial neighborhoods.

Baseline methods

To provide reference performance, we compared challenge submissions against representative transcriptome-to-proteome prediction baselines, including spatially aware models, single-cell RNA-to-protein expression-only prediction models and a classical regularized regression model. All baseline methods were implemented or adapted according to their original descriptions and evaluated under the same preprocessing and evaluation pipeline as the challenge submissions whenever applicable. We used the default parameters reported in the original publications or provided in the corresponding open-source implementations. For methods originally designed for single-cell data, spatial bins were treated as individual observations, with gene expression profiles as input and matched protein measurements as targets.

DGAT. DGAT is a dual-graph attention network designed to infer spatial protein expressions from spatial transcriptomic data. It incorporates both transcriptomic profiles and spatial neighborhood information through graph attention mechanisms.

NicheTrans. NicheTrans is a spatially aware cross-omics translation framework based on a Transformer architecture. It models local cellular neighborhoods to improve prediction of spatial protein abundance from spatial transcriptomic measurements. NicheTrans* is the H&E image-integrated version of NicheTrans.

scTranslator. scTranslator is a pretrained generative model for translating single-cell transcriptomic profiles into proteomic profiles. In this study, it was adapted to the spatial setting by treating each spatial bin as an individual observation. We evaluated both the zeroshot performance of scTranslator (scTranslator-zs) and finetuned version (scTranslator-ft) following the official instructions.

SPIDER[26]. SPIDER is a context-agnostic deep ensemble model developed for predicting cell-surface protein abundance from transcriptomic profiles. It provides a reference for protein prediction using transcriptomic features without explicit spatial modeling.

sciPENN. sciPENN is a recurrent neural network framework for integrating CITE-seq and scRNA-seq data and imputing cell-surface protein expression. It was included as a representative single-cell multi-omics imputation baseline.

cTP-net. cTP-net is the first deep learning-based framework that predicts cell-surface protein abundance from single-cell transcriptomic profiles using paired single-cell RNA and protein measurements from CITE-seq-like datasets.

ElasticNet. We also implemented an ElasticNet as the classical machine learning baseline, which is a regularized linear regression model combining L1 and L2 penalties.

Evaluation metrics

Model performance in the STP Open Challenge was primarily evaluated using the Spearman correlation coefficient between predicted and measured protein abundance. For each protein marker, SCC was computed across all spatial locations. The final leaderboard score was defined as the average SCC across the ten best-predicted protein markers:

$$\text{Score} = \frac{1}{10} \sum_{p \in \mathcal{P}_{10}} \text{SCC}_p,$$

where SCC_p denotes the Spearman correlation coefficient for protein marker p, and $\mathcal{P}_{10}$ denotes the set of ten protein markers with the highest SCC values for a given submission.

Spearman correlation was selected as the primary metric because it is robust to scale differences across protein markers and captures monotonic agreement between predicted and measured protein abundance. The top 10 averaging strategy was used

because protein predictability varies substantially across markers owing to differences in expression prevalence, technical signal quality and biological coupling between RNA and protein levels. This metric therefore reduces the influence of low-signal or poorly detectable markers while emphasizing proteins for which meaningful spatial transcriptome-to-proteome prediction is feasible.

Beyond the primary challenge score, we performed post-challenge analyses using complementary metrics to assess both expression-level prediction accuracy and spatial structure consistency. To evaluate whether models preserved spatial organization, we further computed bivariate Moran's Index between predicted and measured protein distributions for each marker. This spatial metric assesses whether predicted protein signals are spatially concordant with the corresponding measured protein signals, thereby complementing location-wise correlation metrics.

Extensive unseen datasets

To evaluate model robustness beyond the challenge evaluation set, we curated additional in-house unseen datasets. The datasets were generated using the same adjacent-section experimental paradigm as the challenge data and included both same-disease and cross-disease tissue samples. These datasets were used to assess model robustness to biological variability while keeping the overall data generation workflow consistent.

All submitted models and baselines were trained on the challenge training set, and evaluated on the unseen datasets using the same evaluation metrics as in the challenge. Preprocessing procedures were standardized whenever applicable, with platform-specific adaptations made to account for differences in molecular features, spatial resolution and data representation.

Data availability

The challenge-derived paired spatial transcriptomic-proteomic datasets were released on the official website (https://www.codabench.org/competitions/10696/) via Google Drive (https://drive.google.com/drive/folders/1eq6sbTUaWCCOKcnkei6B65rozx-VX70K?usp=sharing).

Code availability

The code of spatial multi-omics data alignment framework and ensemble model is available on GitHub (https://github.com/YimengQiao/STP-Open-Challenge-2025).

Acknowledgments

This work was partially supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0525300: 2024ZD0525303 and 2024ZD0525304) and NSFC/RGC Collaborative Research Scheme (No. 82261160578 to Jiang lab and No. CRS_HKUST605/22 to Wang lab), the RGC grants (R6003-22, 16102724, C4008-23W, C6040-24G, T12-705/24R, T12-611/25-N), the ITC grant (ITCPD/17-9), the Guangdong Science and Technology Department (No. 2025A0505000039, VP2020S25SC07). J.W. is also supported by Padma Harilela Professorship, CAS-Croucher Funding Scheme for Joint Laboratories (CAS24SC03), and Sinovac Fellowship Program (Z1541).

Author Contributions

J.W. supervised the entire project. J.T., R.L. and Y.Q. collected and preprocessed the paired spatial transcriptomic and spatial proteomic data. Y.Q. developed the spatial multi-omics data alignment framework, generated the challenge datasets, initialized the STP Open Challenge, reproduced the submitted codes and performed extensive evaluation. Y.R. helped design the spatial multi-omics data alignment framework. Z.F. developed models and contributed to extensive evaluation. Z.Y. facilitated access to the baseline method code. Y.L., L.X., F.G., M.W., and L.W. developed models and helped with extensive evaluation. H.W., S.S., Y.Z., K.Z., G.H., Z.L., and J.Z. supervised model development. Y.Q. and J.W. wrote the manuscript and all authors read and approved the final manuscript.

Competing Interests

The authors declare no potential conflict of interests.

References

1. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell.* 165, 535–550 (2016).
2. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973 (2009). https://doi.org/10.1016/j.febslet.2009.10.036
3. Liao, S. et al. Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq. Preprint at https://doi.org/10.1101/2023.04.28.538364 (2023).
4. Liu, Y. et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. *Nat Biotechnol* 41, 1405–1409 (2023).
5. Guo, P., Mao, L., Chen, Y. *et al.* Multiplexed spatial mapping of chromatin features, transcriptome and proteins in tissues. *Nat Methods* **22**, 520–529 (2025). https://doi.org/10.1038/s41592-024-02576-0
6. Greenwald, A. C. et al. Integrative spatial analysis reveals a multi-layered organization of glioblastoma. *Cell* 187, 2485-2501.e26 (2024).
7. Tang, J. et al. Protein-based classification reveals an immune-hot subtype in IDH mutant astrocytoma with worse prognosis. *Cancer Cell* 43, 2136-2155.e14 (2025).
8. Zhou, Z., Ye, C., Wang, J. & Zhang, N. R. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat. Commun.* 11, 651 (2020).
9. Lakkis, J. et al. A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation. *Nat. Mach. Intell.* 4, 940–952 (2022).
10. Liu, L. et al. A pre-trained large generative model for translating single-cell transcriptomes to proteomes. *Nat. Biomed. Eng* https://doi.org/10.1038/s41551-025-01528-z (2025) doi:10.1038/s41551-025-01528-z.
11. Wang, Z. et al. NicheTrans: Spatial-aware Cross-omics Translation. Preprint at https://doi.org/10.1101/2024.12.05.626986 (2024).
12. Wang, H. et al. DGAT: a dual-graph attention network for inferring spatial protein landscapes from transcriptomics. *Nat Commun* https://doi.org/10.1038/s41467-026-73114-z (2026) doi:10.1038/s41467-026-73114-z.
13. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins* 89, 1607–1617 (2021).
14. Pham, V. V. H. et al. The winning methods for predicting cellular position in the DREAM single-cell transcriptomics challenge. *Briefings in Bioinformatics* 22, bbaa181 (2021).
15. Xu, Z. et al. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns* 3, 100543 (2022).
16. Oliveira, M.F.d., Romero, J.P., Chung, M. *et al.* High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nat Genet* **57**, 1512–1523 (2025). https://doi.org/10.1038/s41588-025-02193-3
17. Goltsev, Y. *et al.* Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981.e15 (2018). https://doi.org/10.1016/j.cell.2018.07.010
18. Chen, R. J. et al. Towards a general-purpose foundation model for computational

pathology. *Nat Med* 30, 850–862 (2024).
19. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* 30, 863–874 (2024).
20. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* 30, 2924–2935 (2024).
21. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, USA, 2016). doi:10.1109/CVPR.2016.90.
22. Zhao, S. et al. SToFM: a Multi-scale Foundation Model for Spatial Transcriptomics. *In Forty-second International Conference on Machine Learning* (2025).
23. Bahrami, M., Tejada-Lapuerta, A., Becker, S., Hashemi G., F. S. & Theis, F. J. scConcept: Contrastive pretraining for technology-agnostic single-cell representations beyond reconstruction. Preprint at https://doi.org/10.1101/2025.10.14.682419 (2025).
24. Wodzinski, M., Marini, N., Atzori, M. & Müller, H. RegWSI: Whole slide image registration using combined deep feature- and intensity-based methods: Winner of the ACROBAT 2023 challenge. *Computer Methods and Programs in Biomedicine* 250, 108187 (2024).
25. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. & Bengio, Y. Graph attention networks. In *International Conference on Learning Representations* (2018).
26. Chen, R., Zhou, J. & Chen, B. Imputing abundance of over 2,500 surface proteins from single-cell transcriptomes with context-agnostic zero-shot deep ensembles. *Cell Syst.* **15**, 869–884.e6 (2024).

Table 1. Characteristics of the top six best-performing algorithms in preprocessing, feature encoder, model backbone, protein decoder and loss function.

Team	Preprocessing		Feature Encoder			Model Backbone	Protein Decoder	Loss
	RNA	Protein	RNA	Spatial Location	H&E Image			
❶ Team 1	Normalize total to 1e4 + log1p + HVG (LSTM); Raw RNA expression data → Rank encoding (scConcept)	Log1p, z-score	scConcept (finetuned) + LSTM	kNN	UNI-v2	Three Branches: LSTM + scConcept + MLP Late fusion with MLP	MLP	MSE + InfoNCE
❷ Team 2	Normalization, log1p	Log1p, z-score	MLP	Spatial KNN graph	ResNet50	GATv2	Linear projection	Huber loss + biological regularization
❸ Team 3	Z-score	Z-score	MLP	kNN neighborhood statistics	ResNet18 on multi-scale patches	residual MLP with conditional H&E refinement and gated fusion	Correlation-aware prediction head	Combined regression loss
❹ Team 4	CPM, log1p, HVGs	Z-score	/	/	ViT	MoE + Evoformer	MLP	MSE loss
❺ Team 5	CPM, log1p	QuantileTransformer, z-score	SToFM	/	UNI-v2 + Virchow2	Diffusion model	Reverse diffusion + inverse pipeline	MSE loss
❻ Team 6	Normalize total to 1e4, log1p, select top 1000 correlated RNA	Z-score	/	kNN graph (K=8)	/	GAT + residual MLP	MLP + linear bypass	MSE + spatial smooth loss

Figures

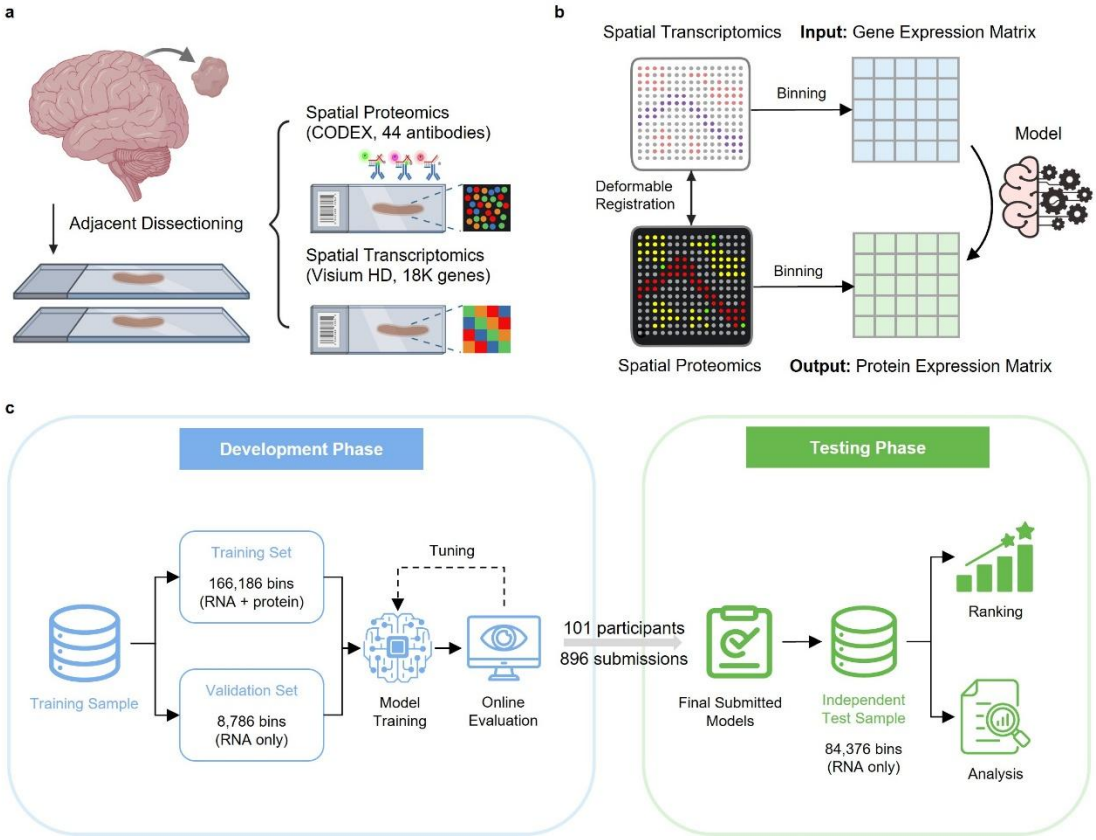


Figure 1. Overview of the STP Open Challenge task and pipeline

(**a**) Schema of spatial transcriptomic and spatial proteomic data generation process. For one patient sample, two adjacent slides were dissected serially, one for spatial proteomics (CODEX with 44 antibodies) and another for spatial transcriptomics (Visium HD with $2\mu m$ resolution and 18k gene panel). (**b**) Spatial Transcriptome-to-Proteome (STP) task definition. For paired spatial transcriptomics and spatial proteomics data, gene expression matrix and protein abundance levels were extracted after deformable registration of the H&E staining image of the spatial transcriptomics data and the DAPI channel of the spatial proteomics data at 16 μm bin resolution. (**c**) The challenge contained two phases. During the development phase, participants developed STP prediction models based on 166,186 bins with paired RNA and protein expression values. The models could be evaluated on a validation set generated from the same sample with the training set consisting of 8,786 held-out bins with RNA data only. The Codabench online evaluation platform automatically returned back the quantitative performance for model tuning. During the testing phase, each team could submit one model to the platform as the final solution, which was independently evaluated on an independent testing sample with 84,376 bins to obtain ranking results and perform qualitative analysis.

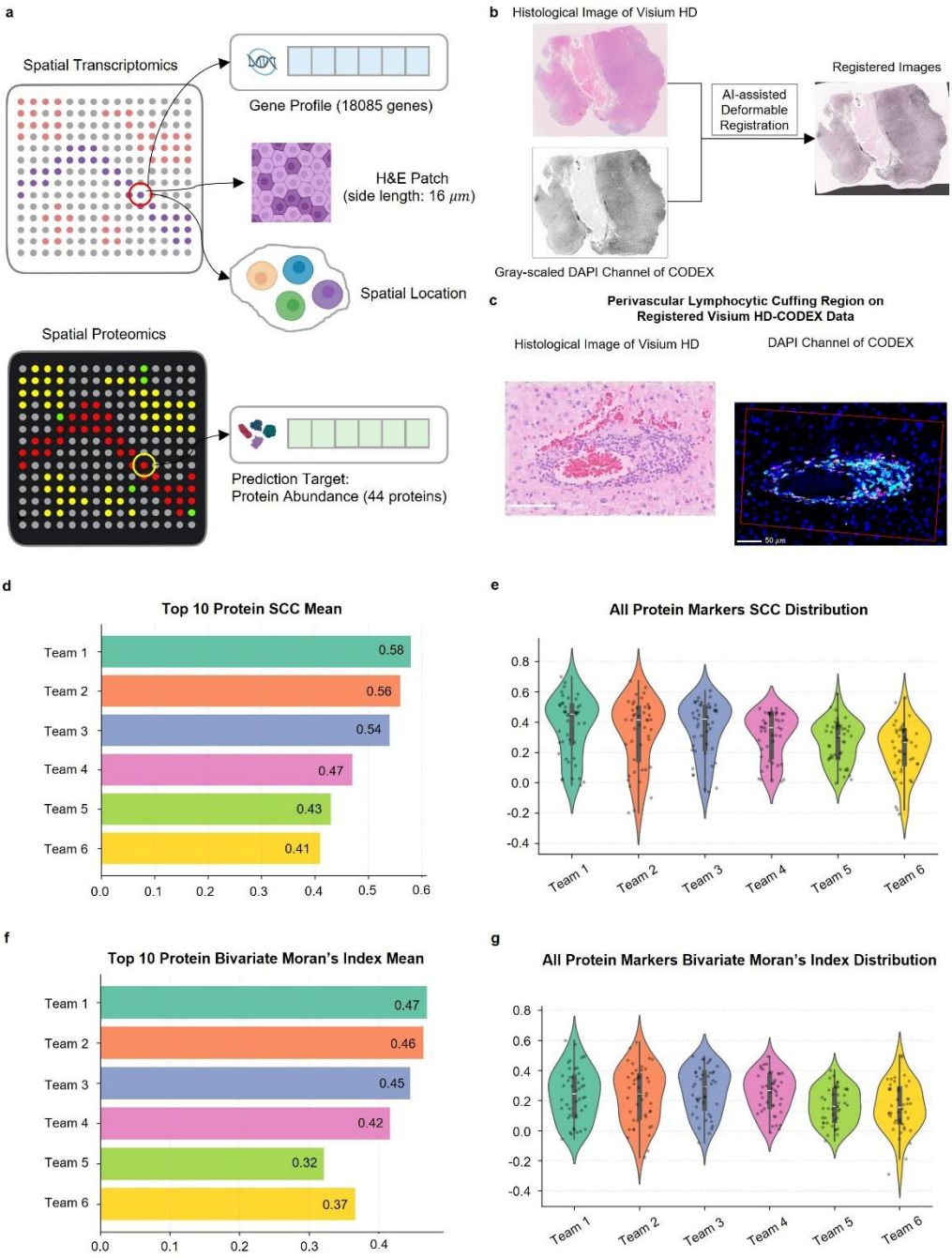


Figure 2. STP Open Challenge dataset description and performance summary

(a) Overview of the input and output of the challenge. For each spatial location, participants were provided with the 18,085 gene profile, spatial coordinates, and matched histological image patch derived from the Visium HD data, and were tasked to develop models to predict the abundance of 44 target protein markers measured by CODEX in the corresponding tissue region. **(b)** AI-assisted spatial multi-omics alignment. The high-resolution histological image of Visium HD and gray-scaled DAPI channel image of CODEX were extracted for deep learning-based deformable registration. The registered histological image of Visium HD was overlayed on the DAPI channel image of CODEX shown in the right figure, which demonstrated excellent

overlap in both global or local morphological structures. **(c)** Representative example of a registered Visium HD-CODEX tissue region showing a perivascular lymphocytic cuffing (PLC) structure. The left shows the histological image of PLC derived from Visium HD data, while the right shows corresponding multiplexed immunofluorescence staining region of CODEX data. This example demonstrates the spatial correspondence between transcriptomic and proteomic measurements after image registration. **(d)** Mean Spearman Correlation Coefficients (SCC) of the top 10 best-predicted proteins of the top six teams on the held-out test set, which is same as the final challenge leaderboard. **(e)** Distribution of SCC scores across all 44 protein markers for each team, highlighting substantial marker-specific variability in prediction performance. **(f)** Mean bivariate Moran's Index values of the top 10 best-predicted proteins of the six teams on the held-out test set, measuring the preservation of spatial expression patterns in model predictions. **(g)** Distribution of bivariate Moran's Index values across all 44 protein markers for each team, illustrating the differences in the ability of top-performing submissions to preserve spatial protein organization.

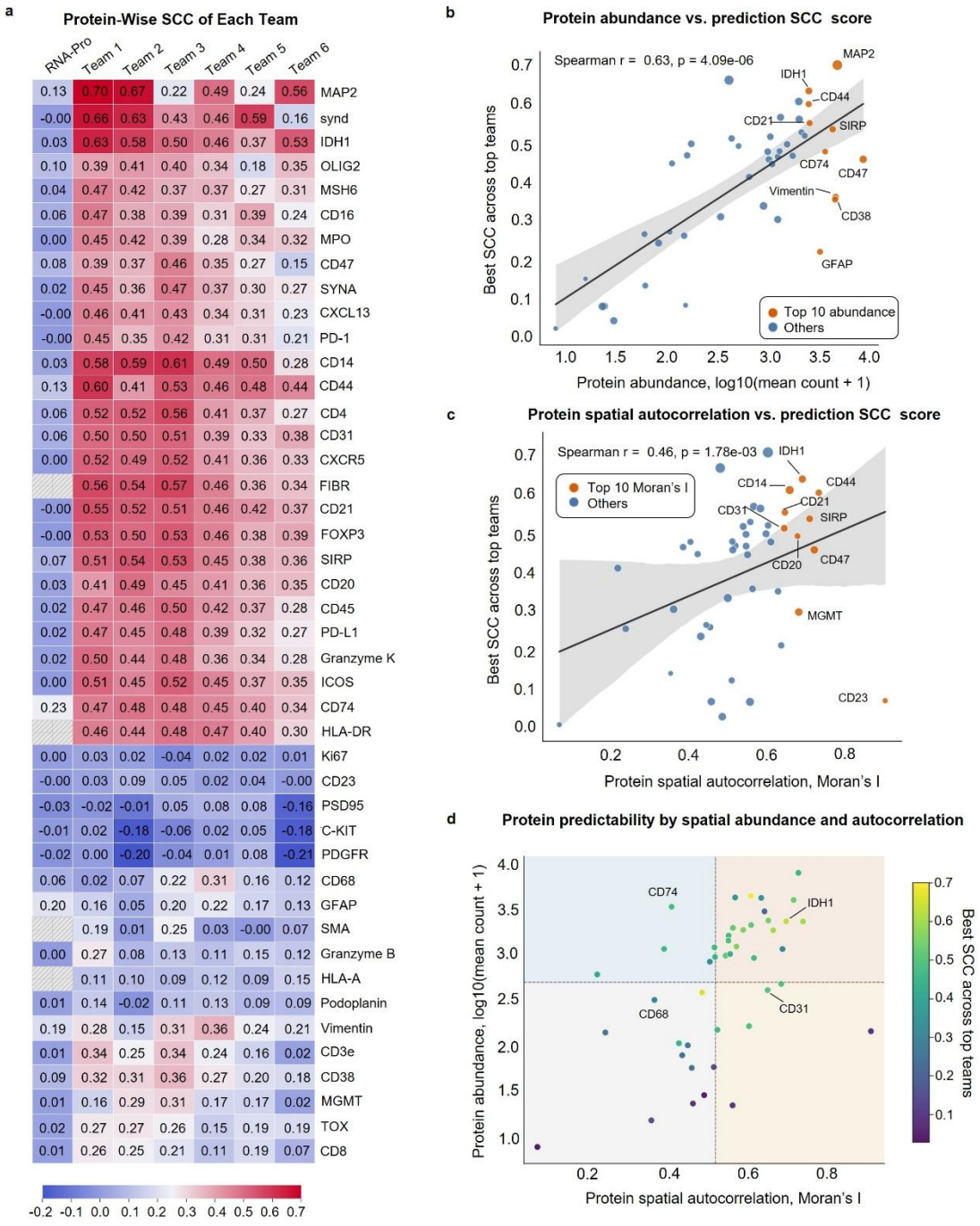


Figure 3. Detailed analysis of the predicted proteins by each team

(a) Heatmap of the SCC between predicted protein abundance levels and the ground truth of each marker by each model, revealing substantial heterogeneity in both marker predictability and model performance. **(b)** Scatter plot of the correlation between protein abundance and best predicted SCC scores across teams. Each dot represents a protein marker. **(c)** Scatter plot of the correlation between protein spatial autocorrelation and best predicted SCC scores across teams. Each dot represents a protein marker. **(d)** Scatter plot of the protein predictability defined by the spatial abundance-structure axis. Each dot represents a protein marker, illustrating how

protein abundance and spatial organization jointly relate to the upper-bound prediction performance achieved in this challenge.

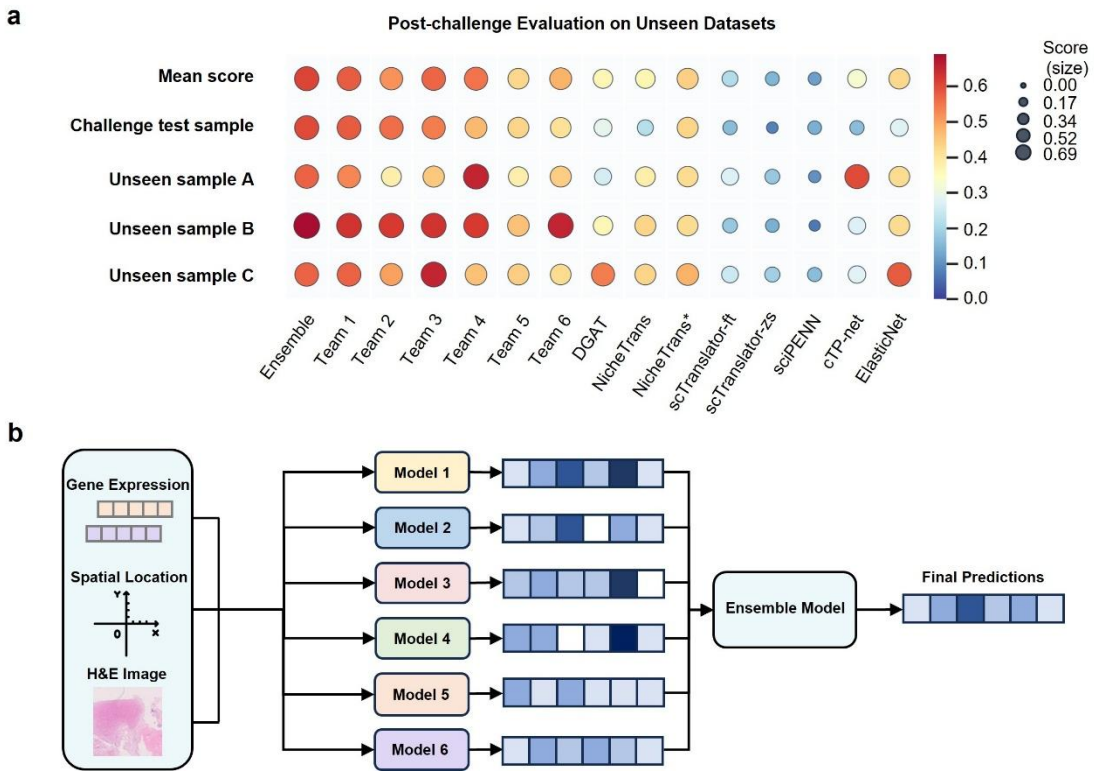
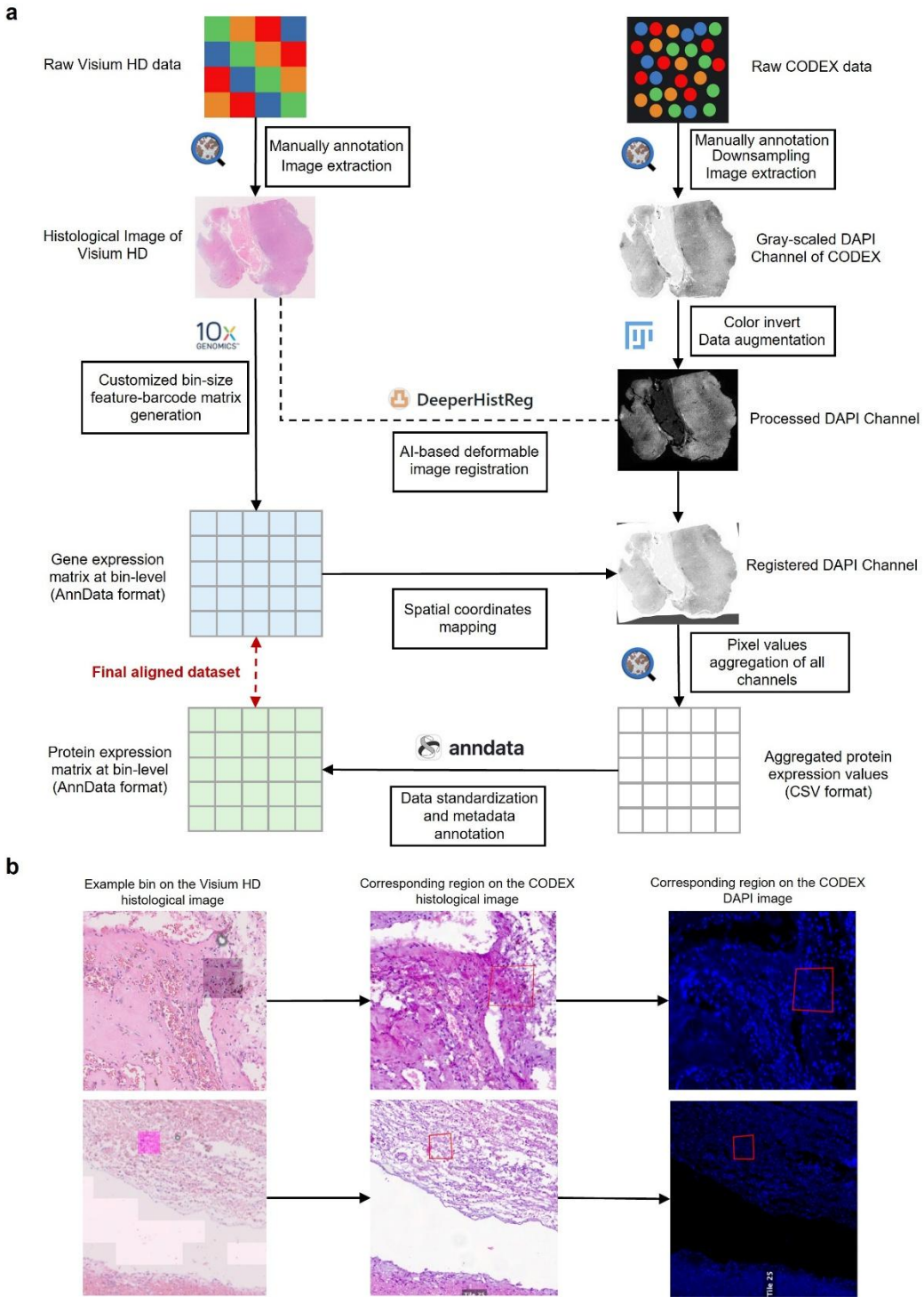


Figure 4. Post-challenge extensive evaluation on unseen datasets and ensemble model

(a) Performance comparison of the ensemble model, the top six submitted models and existing baseline methods on the challenge test sample and three unseen samples. Dot size and color represent the mean SCC of the 10 best-predicted protein markers. **(b)** Schematic overview of the ensemble model integrating predictions from the top six submitted models. Predictions generated by the top six challenge submissions were integrated through a consensus-based ensemble strategy to produce final protein abundance predictions.

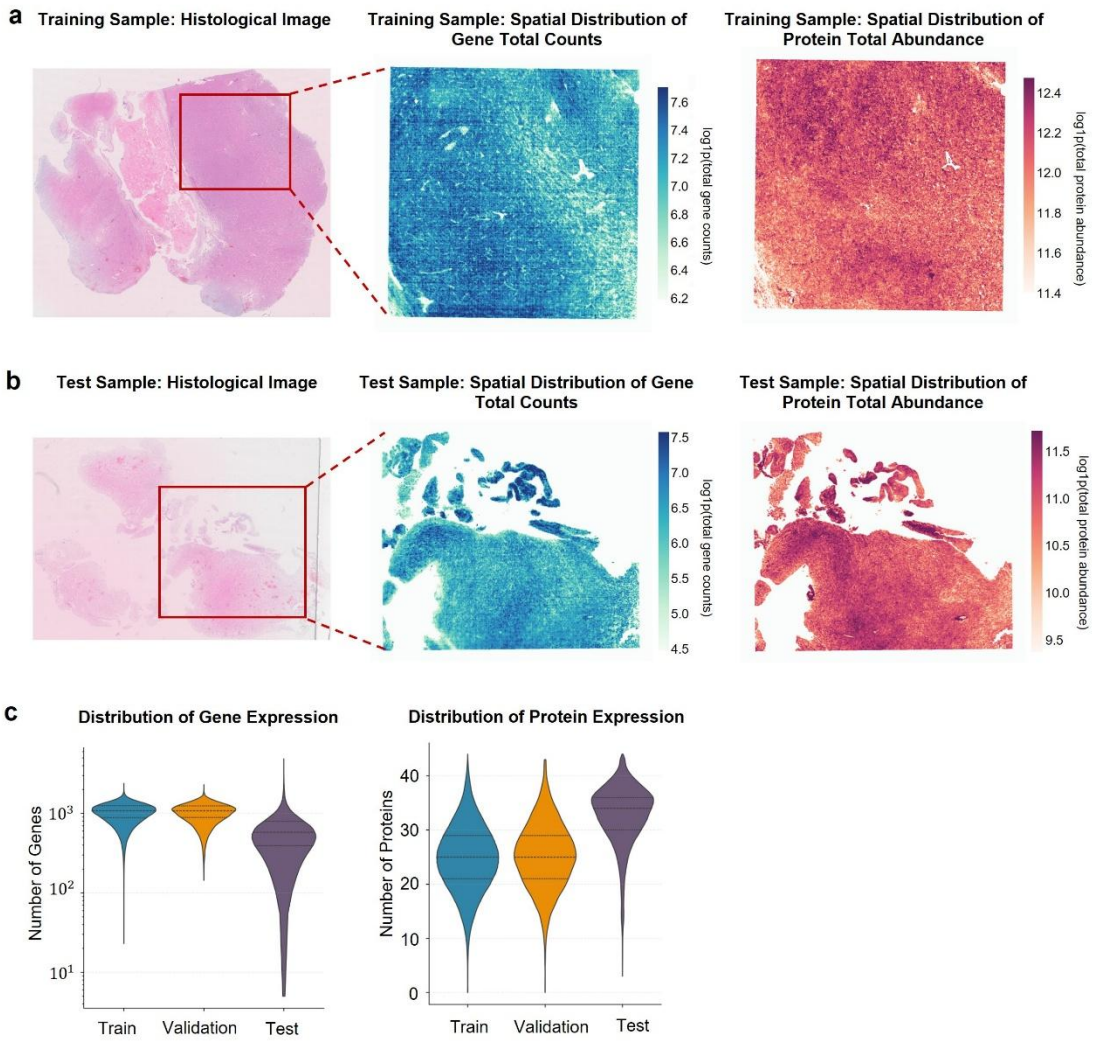
Extended Figures



Extended Figure 1. Workflow for Visium HD-CODEX data alignment and examples of aligned Visium HD-CODEX bins

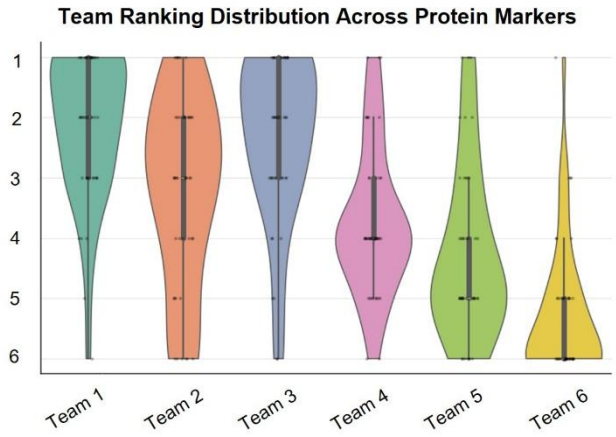
(a) Workflow for Visium HD-CODEX data alignment. Histological images from the raw Visium HD data were manually annotated and extracted using QuPath. DAPI- channel images from raw CODEX data were manually annotated, downsampled, and gray-scaled in QuPath as well. The preprocessed DAPI-channel images were then loaded into Fiji for color inversion and data augmentation. Next, the processed DAPI-channel images were registered to the Visium

HD histological images with an AI-based deformable image registration model DeeperHistReg. The bin-level spatial coordinates from the Visium HD gene expression matrix were then mapped onto the registered CODEX DAPI images. Finally, pixel-level protein signals were aggregated at the same bin resolution, followed by data standardization and metadata annotation, to generate the paired bin-level protein expression matrix. **(b)** Representative examples of aligned Visium HD–CODEX regions. The first column shows example bins on the Visium HD histological images. The middle column shows the corresponding regions on the adjacent CODEX images, and the right column shows the same regions on the registered CODEX DAPI images. These examples demonstrate high-quality cross-modality alignment even at a fine bin-level resolution of 26 μm.



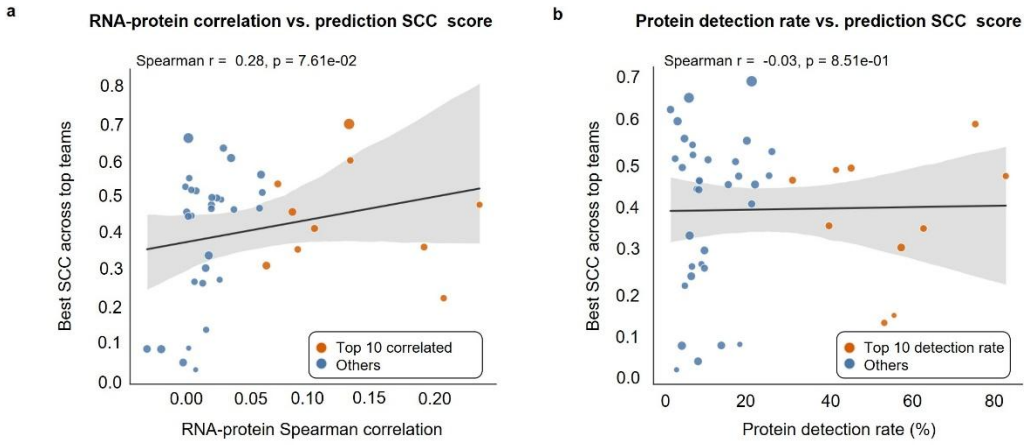
Extended Figure 2. Data characteristics of challenge datasets

(a) Data characteristics of the training sample. From left to right, the panels show the histological image, the spatial distribution of total gene counts in the Visium HD data, and the spatial distribution of total protein abundance in the CODEX data. **(b)** Data characteristics of the test sample. From left to right, the panels show the histological image, the spatial distribution of total gene counts in the Visium HD data, and the spatial distribution of total protein abundance in the CODEX data. (**c**) Distributions of expressed genes and proteins in the training, validation, and test sets. These visualizations provide an overview of tissue morphology, spatial molecular coverage, and expressed molecular features across the challenge datasets.



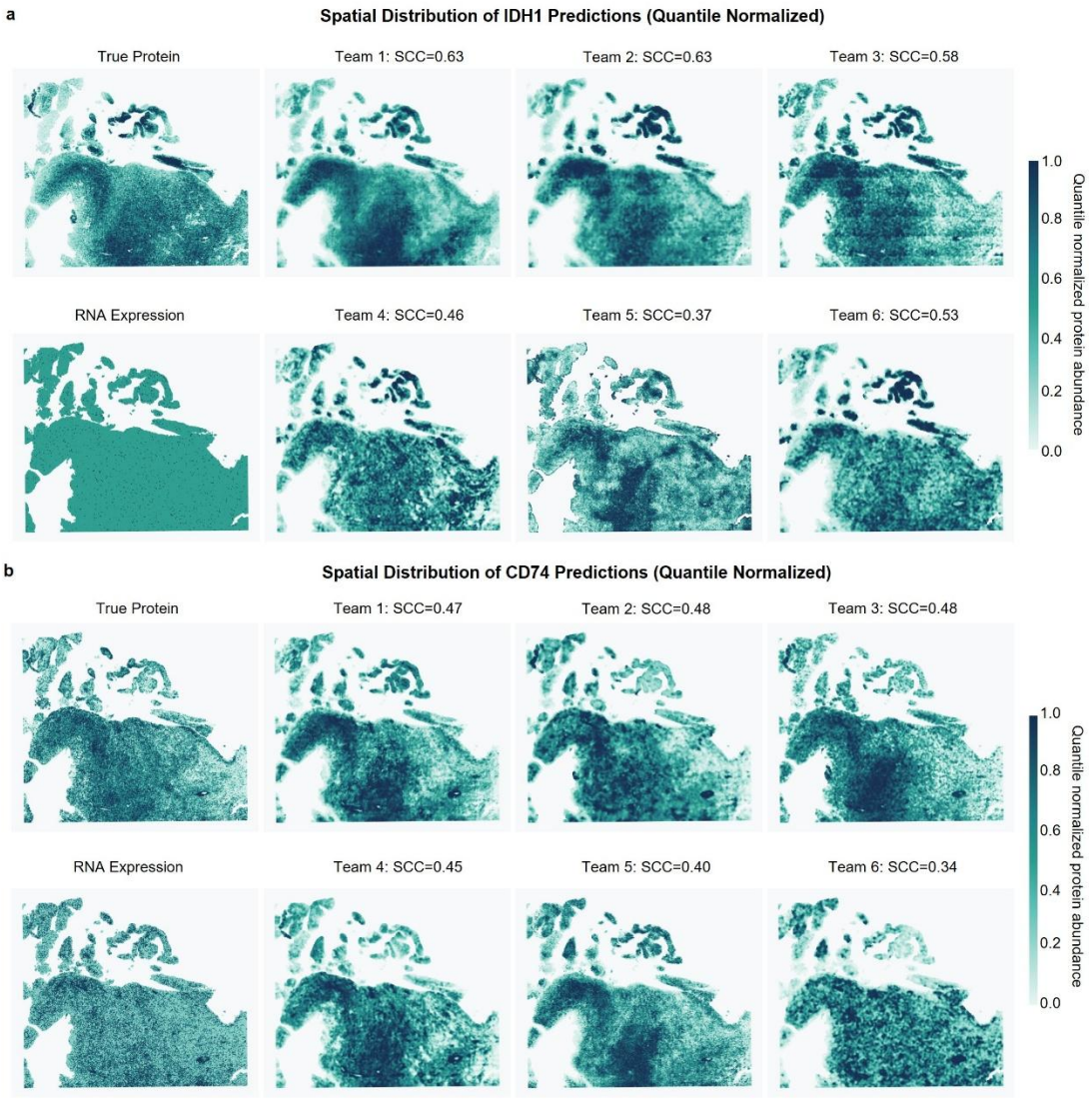
Extended Figure 3. Team ranking distribution across protein markers

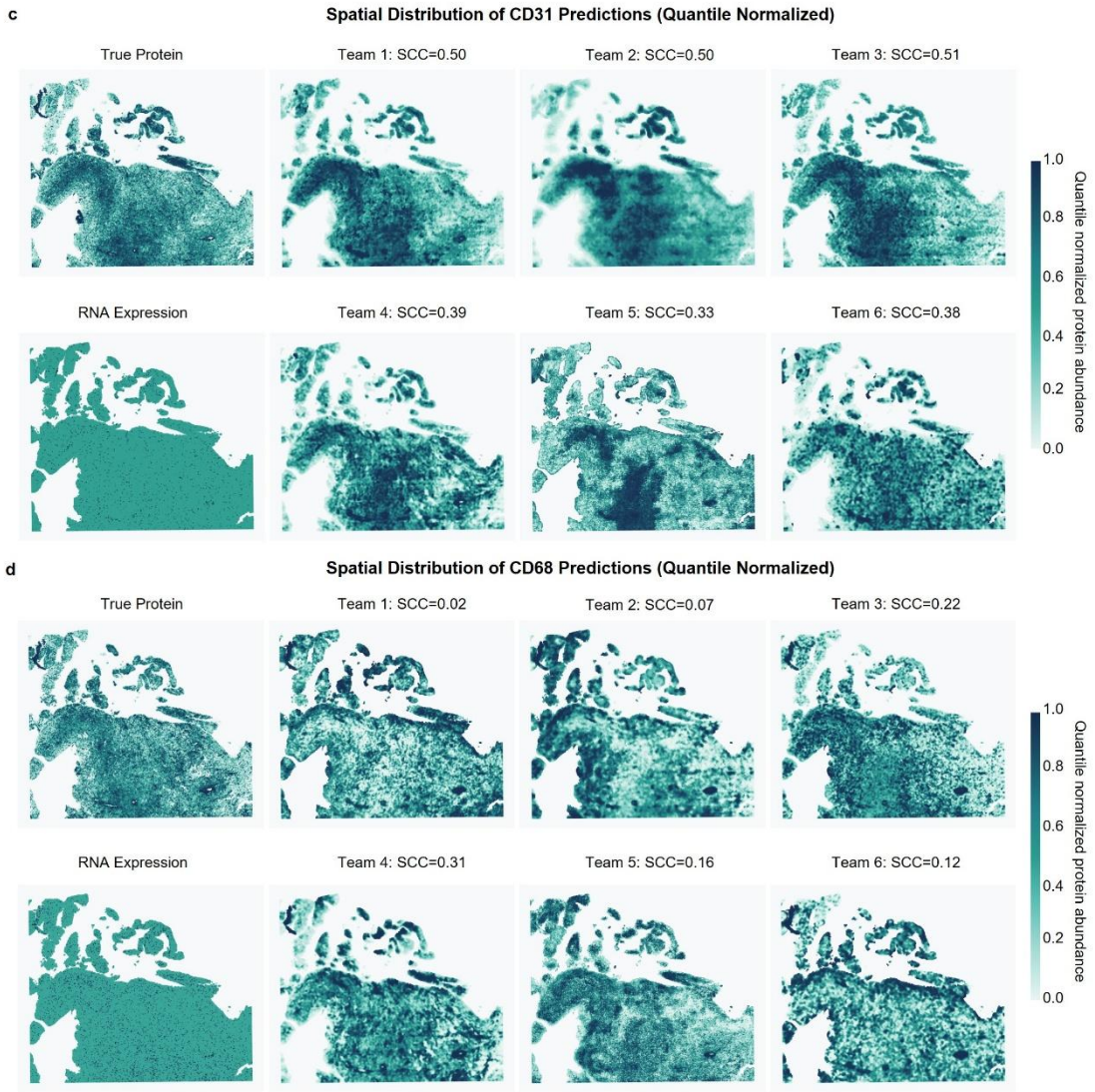
Violon plot of team ranking distribution across all 44 protein markers on the challenge testing sample. The rankings are calculated based on SCC scores. Each dot represents a protein marker. The distribution highlights substantial marker-level prediction variation in model performance, with no single method consistently achieving the top rank across all proteins, suggesting complementary strengths among prediction methods and motivating the development of ensemble strategies.



Extended Figure 4. Association between protein predictability and protein properties.

(a) Association between RNA-protein Spearman correlation and marker-level predictability. Each dot represents one protein marker, and predictability is defined as the best SCC across teams for that marker. Orange dots show the top 10 markers with the highest RNA-protein correlation. **(b)** Association between protein detection rate and marker-level predictability. Each dot represents one protein marker, and predictability is defined as the best SCC across teams for that marker. Orange dots show the top 10 markers with the highest detection rates.

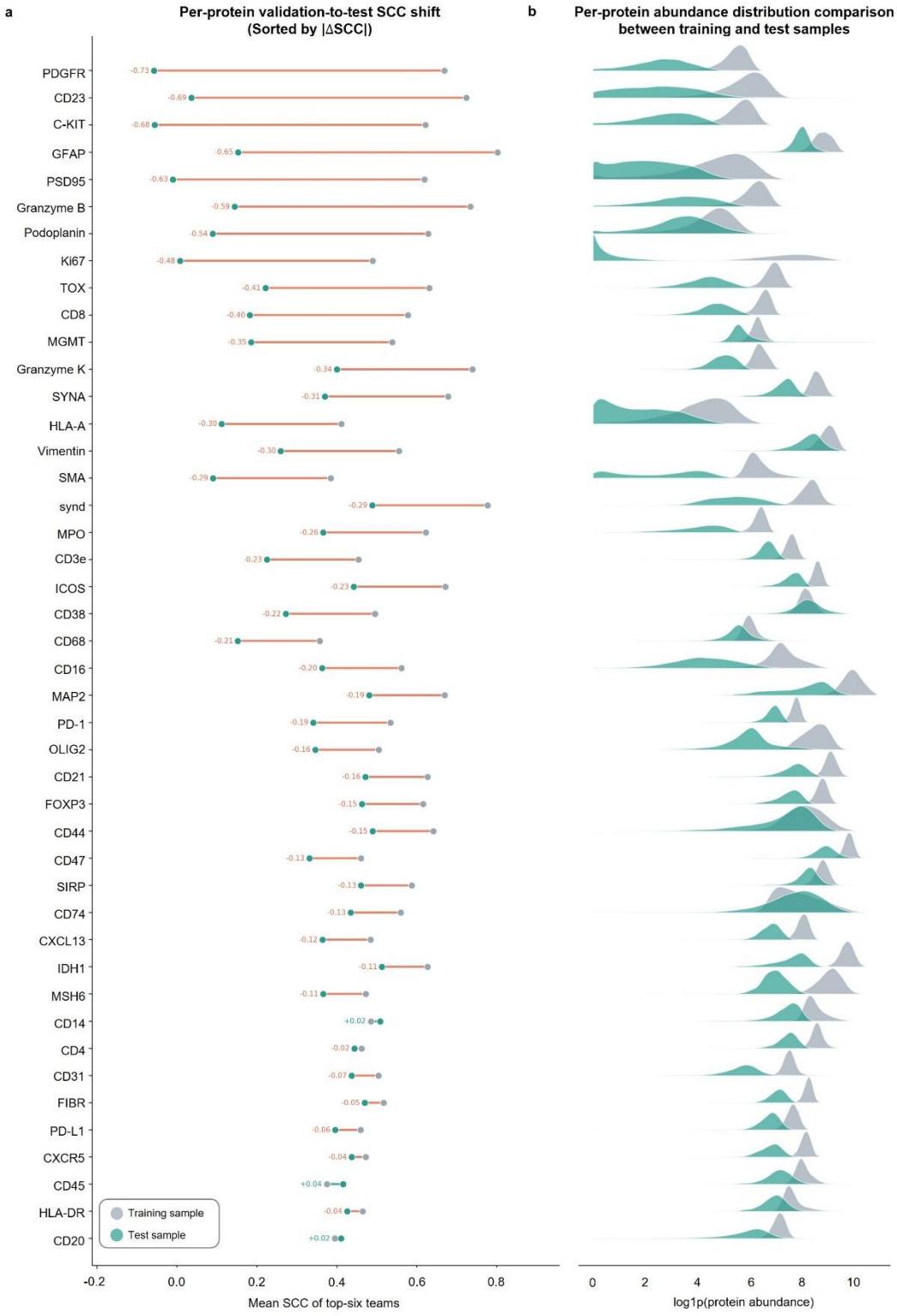

a
Spatial Distribution of IDH1 Predictions (Quantile Normalized)
True Protein
Team 1: SCC=0.63
Team 2: SCC=0.63
Team 3: SCC=0.58
RNA Expression
Team 4: SCC=0.46
Team 5: SCC=0.37
Team 6: SCC=0.53
Quantile normalized protein abundance
1.0
0.8
0.6
0.4
0.2
0.0
b
Spatial Distribution of CD74 Predictions (Quantile Normalized)
True Protein
Team 1: SCC=0.47
Team 2: SCC=0.48
Team 3: SCC=0.48
RNA Expression
Team 4: SCC=0.45
Team 5: SCC=0.40
Team 6: SCC=0.34
Quantile normalized protein abundance
1.0
0.8
0.6
0.4
0.2
0.0



Extended Figure 5. Representative spatial distribution of protein abundance predicted by top-performing teams

(a) Spatial distribution of IDH1 predictions by top-performing teams. Predicted protein abundance values were quantile-normalized and rescaled to the range of 0–1 for visualization. IDH1 was broadly abundant across the tissue and exhibited clear spatial patterns. Teams with higher SCC scores more accurately reconstructed the IDH1-enriched regions, particularly in the upper and left regions of the tissue. **(b)** Spatial distribution of CD74 predictions by top-performing teams. Predicted protein abundance values were quantile-normalized and rescaled to the range of 0–1 for visualization. CD74 showed relatively high abundance at both the RNA and protein levels but lacked obvious spatial structure. Top-performing teams generally recovered the overall CD74 distribution, although some models overestimated local spatial enrichment, such as Team 3 and Team 5. **(c)** Spatial distribution of CD31 predictions by top-performing teams. Predicted protein abundance values were quantile-normalized and rescaled to the range of 0–1 for visualization. CD31 was not among the most abundant markers but displayed clear spatial enrichment, particularly in the upper and left regions of the tissue. Top-performing teams were able to capture these spatial patterns well. **(d)** Spatial distribution of CD68 predictions by top-performing teams. Predicted protein abundance values were quantile-

normalized and rescaled to the range of 0–1 for visualization. CD68 showed relatively low overall abundance and lacked clear spatial enrichment, making its spatial distribution difficult to infer. Even top-performing teams struggled to reconstruct CD68 patterns, likely due to the weak and spatially homogeneous target signal.



Extended Figure 6. Validation-to-test protein predictive accuracy changes and protein abundance distribution shift between training and test datasets

(a) Per-protein changes in validation-to-test SCC. Gray dots indicate the mean SCC scores of the top-six teams on the challenge validation set, whereas green dots indicate the corresponding mean SCC scores on the challenge test set. Connecting lines show the validation-to-test changes in protein-level predictive accuracy, with proteins sorted by the absolute SCC change, |ΔSCC|.

(b) Per-protein comparison of protein abundance distributions between the training and test samples. Each ridgeline panel shows the density distributions of one protein marker in the training and test samples, enabling visual assessment of marker-specific abundance distribution shifts across datasets.

Supplementary Table 1. Summary of the spatial transcriptomic-proteomic datasets

Dataset	Platform	Resolution	Num. of bins/cells	Num. of genes	Num. of proteins
Challenge train	Visium HD-CODEX	16 μm	174,972	18,085	44
Challenge test	Visium HD-CODEX	16 μm	84,376	18,085	44
Internal unseen sample A	Visium HD-CODEX	26 μm	60,796	18,085	44
Internal unseen sample B	Visium HD-CODEX	26 μm	49,273	18,085	44
Internal unseen sample C	Visium HD-CODEX	26 μm	65.032	18,085	44

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- STPChallengereportingsummary.pdf