Generalized cell phenotyping for spatial proteomics with language-informed vision models

Xuefei (Julie) Wang[1], Rohit Dilip[2], Yuval Bussi[4,5], Caitlin Brown[1], Elora Pradhan[1], Yashvardhan Jain[3], Kevin Yu[1], Shenyi Li[1], Martin Abt[1], Katy Börner[3], Leeat Keren[5], Yisong Yue[2], Ross Barnowski[1], and David Van Valen[1,6*]

[1] *Division of Biology and Biological Engineering, Caltech, Pasadena, CA*

[2] *Division of Computing and Mathematical Science, Caltech, Pasadena, CA*

[3] *Department of Intelligent Systems Engineering, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN*

[4] *Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel*

[5] *Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel*

[6] *Howard Hughes Medical Institute, Chevy Chase, MD*

November 2, 2024

*Correspondence: vanvalen@caltech.edu

Abstract

We present a novel approach to cell phenotyping for spatial proteomics that addresses the challenge of generalization across diverse datasets with varying marker panels. Our approach utilizes a transformer with channel-wise attention to create a language-informed vision model; this model's semantic understanding of the underlying marker panel enables it to learn from and adapt to heterogeneous datasets. Leveraging a curated, diverse dataset with cell type labels spanning the literature and the NIH Human BioMolecular Atlas Program (HuBMAP) consortium, our model demonstrates robust performance across various cell types, tissues, and imaging modalities. Comprehensive benchmarking shows superior accuracy and generalizability of our method compared to existing methods. This work significantly advances automated spatial proteomics analysis, offering a generalizable and scalable solution for cell phenotyping that meets the demands of multiplexed imaging data.

1 Introduction

Understanding the structural and functional relationships present in tissues is a challenge at the forefront of basic and translational research. Recent advances in multiplexed imaging have expanded the number of transcripts and proteins that can be quantified simultaneously[1–10], opening new avenues for large-scale analysis of human tissue samples. Concurrently, advances in deep learning have shown immense potential in integrating information from both image and natural language to build foundation models, and these approaches have also shown to be promising for various biomedical imaging applications[11,12]. However, a critical question persists: how can these innovative methods be harnessed to transform the vast amounts of data generated by multiplexed imaging into meaningful biological insights?

This paper proposes a novel language-informed vision model to solve the problem of generalized cell phenotyping in spatial proteomic data. While the new data generated by modern spatial proteomic platforms are exciting, significant challenges in analyzing and interpreting these datasets at scale remain. Unlike flow cytometry or single-cell RNA sequencing, tissue imaging is performed with intact specimens. Thus, to extract single-cell data, individual cells must be identified - a task known as cell segmentation - and the resulting cells must be examined to determine their cell type and which markers they express - a task known as cell phenotyping. A general solution for cell phenotyping has proven more challenging for several reasons. First, it requires scalable, automated, and accurate cell segmentation, which has only recently become available[13–17]. Second, imaging artifacts, including staining noise, marker spillover, and cellular projections, pose a formidable challenge to phenotyping algorithms[18–20]. Third, general phenotyping algorithms must handle the substantial differences in marker panels, cell types, and tissue architectures across experiments. Each new dataset often has a different number of markers, each with its own distinct meaning. Existing approaches to meet this challenge range from conventional methods that require manual gating and clustering[21–23] to more recent machine learning-based solutions[18–20,24–26]. While representing significant breakthroughs, these tools have failed to scale for two reasons. First, many of them rely on human intervention, which places a fundamental limit on their ability to scale to big data. Second, these methods cannot handle the wide variability in marker panels that exist across experiments. Some methods require labeling and re-training for new datasets; even when transfer learning is possible[18–20], the target dataset must share similarities in the marker panel with the source dataset. Bridging this gap requires a versatile model that can be trained on multiple datasets and performance inference on new data with unseen markers.

To this end, we developed an end-to-end cell phenotyping model capable of learning from and generalizing to diverse datasets, regardless of their specific marker panels. Our approach was twofold. First, we began by curating and integrating

a large, diverse set of spatial proteomics data that includes a substantial amount of publicly available datasets in addition to all the datasets generated by the NIH HuBMAP consortium to date. Human experts generated labels with a human-in-the-loop framework for representative fields of view (FOV) for each dataset member; we term the resulting labeled dataset Expanded TissueNet. Second, we developed a new deep learning method, DeepCellTypes, that is capable of learning how to perform cell phenotyping on these diverse data. This architecture incorporates language and vision encoders, enabling it to leverage information from raw marker images and the semantic information associated with language describing different markers and cell types. We employed a transformer architecture with the channel-wise attention mechanism to integrate this visual and linguistic information, thereby eliminating the dependence on specific marker panels. When trained on Expanded TissueNet, our method demonstrates state-of-the-art accuracy across a wide spectrum of cell types, tissue types, and spatial proteomics platforms. Moreover, we demonstrate that DeepCellTypes has superior zero-shot cell phenotyping performance compared to existing methods and can generalize to new datasets with unseen markers. Both Expanded TissueNet and DeepCellTypes are made available through the DeepCell software library with permissive open-source licensing.

2 Results

Here, we describe three key aspects of our work - constructing Expanded TissueNet, the deep learning architecture of DeepCellTypes, and our model training strategy designed to improve generalization.

2.1 DeepCell Label enables scalable construction of ExpandedTissueNet

Training data quality, diversity, and scale are at the foundation of robust and generalizable deep learning models. To create a dataset that captured the diversity of marker panels, cellular morphologies, tissue heterogeneity, and technical artifacts present in the field, we first compiled data from published sources[19,27–40], as well as unpublished data deposited in the HuBMAP data portal. For each dataset, we collected raw images, corresponding channel names, and cell type labels (when available). Each dataset was resized to a standard resolution of 0.5 microns per pixel. A key step in our process was standardizing marker names and cell types across all datasets to enable cross-dataset comparisons, integration, and analysis. Cell types were organized by lineage as shown in Fig. 1b. We then performed whole-cell segmentation with Mesmer[13] and mapped any existing cell type labels to the resulting cell masks. When quality labels did not exist, we generated them through a human-in-the-loop labeling framework that leveraged expert labelers. Marker positivity labels were generated by manually gating the mean signal intensity for each marker for each dataset. To accelerate this step, we extended DeepCell Label, our cloud-based software for distributed image labeling, to the cell type labeling task (Fig. S3). DeepCell Label allows users to visualize images, analyze marker intensities, refine segmentation masks, and annotate cell types.

The resulting dataset, Expanded TissueNet, consists of 10.5 million cells, spanning six imaging platforms: Imaging Mass Cytometry (IMC)[3], CO-Detection by indEXing (CODEX)[2], Multiplex Ion Beam Imaging (MIBI)[1], Iterative Bleaching Extends Multiplexity (IBEX)[5], MICS (MACSima Imaging Cyclic Staining)[4], and Multiplexed immunofluorescence (MxIF) with Cell DIVE^{TM} technology (Leica Microsystems, Wetzlar, Germany), with the majority of data coming from the first three. The dataset covers 13 common tissue types and 28 specific cell types across 7 broad cell lineages, providing a diverse representation of human biology. Across all datasets, we cataloged 177 unique protein markers with an average of 27

markers per dataset.

To generate single-cell images to train cell phenotyping models, we extracted a 64x64 patch from each marker image centered on each cell to capture an image of that cell and its surrounding neighborhood. These images were augmented with two binary masks: a self-mask delineating the central cell (1's for pixels inside the cell and 0's for pixels outside) and a neighbor-mask capturing surrounding cells within the patch. This approach preserves cell morphology, marker information, and spatial context for model training.

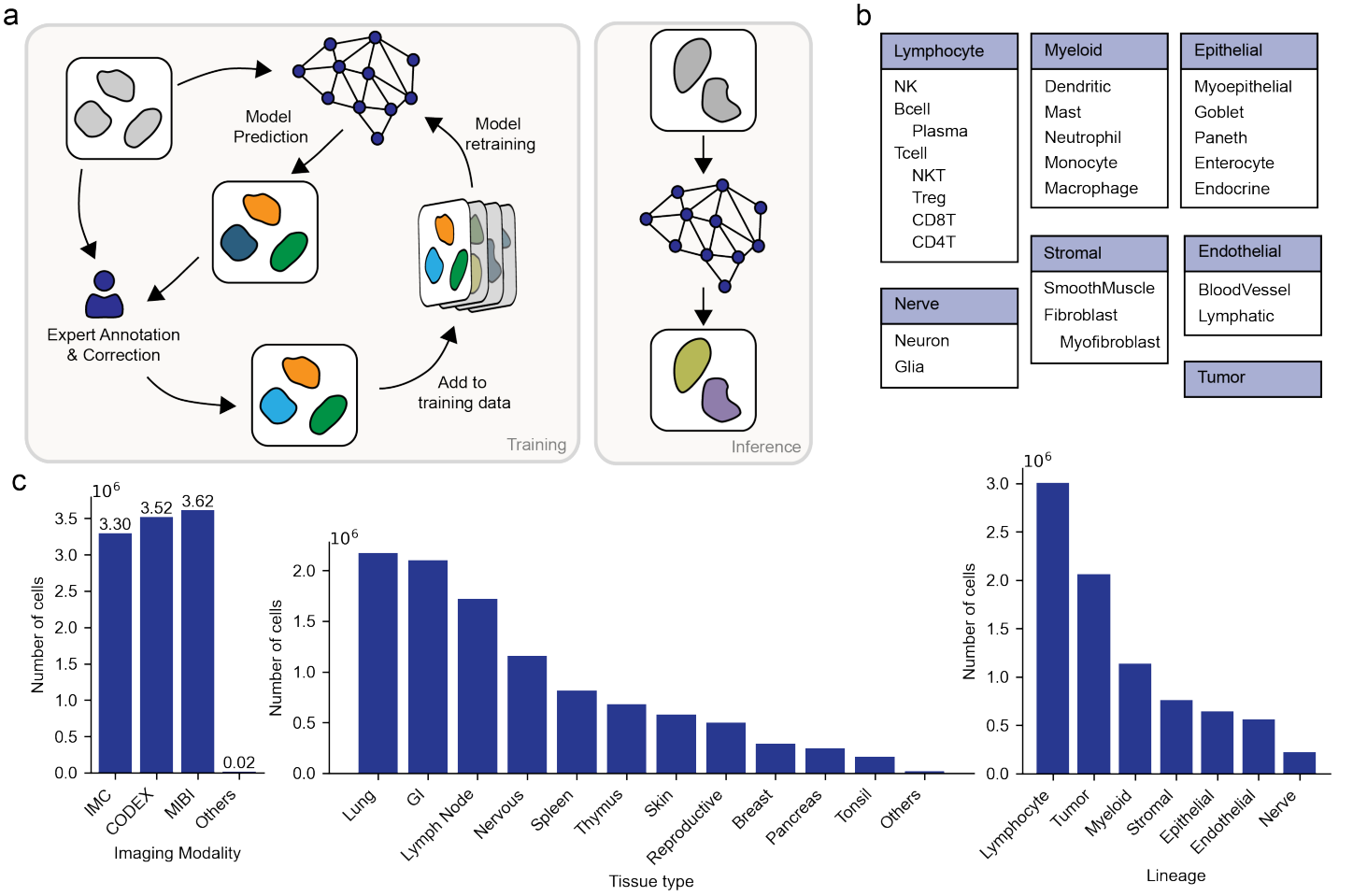


Figure 1: **Construction of Expanded TissueNet with Expert-in-the-loop labeling.** a) We constructed Expanded TissueNet by integrating human experts into a human-in-the-loop framework. Cell phenotype labels are generated by having an expert either label fields of view from scratch or by correcting model errors. We adapted our image labeling software DeepCell Label to facilitate visualization, inspection, and phenotype labeling of spatial proteomic datasets. Newly labeled data were added to the training dataset to enable continuous model improvement. b) Expanded TissueNet covers 7 major cell lineages, each containing a hierarchy of cell types. The labels for these 28 specific cell types enable the training of our cell phenotype model. c) Expanded TissueNet contains over 10 million cells; here, we show the number of labeled cells across 6 imaging platforms, 13 tissues, and 7 lineages.

2.2 DeepCellTypes is a language-informed vision model with a channel-wise transformer

We developed DeepCellTypes, a cell phenotyping model with the unique ability to learn from and adapt to diverse datasets with different marker panels. Our model consists of three main components: a visual encoder, a language encoder, and a channel-wise transformer (Fig. 2a).

Visual Encoder: The visual encoder processes 64x64 cell patches from each channel, along with corresponding self-masks and neighbor-masks. This module employs a Convolutional Neural Network (CNN) to condense the image patches into embedding vectors, capturing spatial information about staining patterns, cell morphology, marker expression levels, and neighborhood context.

Language Encoder: To incorporate semantic understanding of markers and cell types, we employed a language encoder that uses a large language model (LLM) to generate semantically rich embedding vectors in a two-step process (Fig. 2b): First, a frozen LLM explainer is prompted to extract comprehensive knowledge about a marker or cell type, capturing general information, marker-cell type relationships, and alternative names. Next, a frozen LLM embedder converts this semantic information into an information-rich embedding vector. This approach leverages knowledge about the biology of markers and cell types, enabling our model to understand the meaning behind each channel.

Channel-wise Transformer: To allow our model to generalize across marker panels, we use a transformer module with channel-wise attention. This module adds a marker's image embedding with its corresponding language embedding and applies self-attention to the combined representation. Crucially, this self-attention operation is applied across all the markers, similar to how a human might look across the marker images to interpret a stain. While transformers for sequence data typically include a positional encoding, we did not include one. This design choice preserved the length-and-order invariance of self-attention[41] and allowed our model to process inputs from diverse marker panels without modification. We appended a learnable `[CLS]` token to represent the overall cell and to aggregate information across all channels. The normalized attention weights between this `[CLS]` token and the marker embeddings in the final layer provide interpretable marker positivity scores. This architecture serves as a fusion point for visual and linguistic information, enabling our model to discern cross-channel correlations while also understanding the biological significance of each marker.

2.3 Improved cross-platform generalization through contrastive and adversarial learning

Our training strategy extends the joint vision-language approach from the model architecture to loss function design. Instead of relying on a standard classification loss, we employ the Contrastive Language-Image Pretraining (CLIP) loss, which has demonstrated remarkable efficacy in integrating image and text information across various tasks[42]. We use the same language encoder to embed markers and cell types. During training, our language-informed vision model was tasked with aligning each cell's image-marker embedding (e.g., the cell's `[CLS]` token embedding taken from the final transformer layer) with the embedding of its corresponding cell type name (Fig. 2a). It was also tasked with minimizing the similarity between the cell's image-marker embedding with embeddings of incorrect cell type names. This contrastive training approach unifies the visual (`[CLS]` token embeddings) and textual (cell-type-name embeddings) representations in a shared latent space, enhancing the model's ability to capture nuanced semantic meanings between different cell types. We used a focal-enhanced[43] of the CLIP loss to deal with class imbalance by assigning higher importance to difficult examples. Furthermore, we applied a binary cross entropy loss to the normalized attention weights to force alignment between the attention weights and marker positivity. For this loss, we used label smoothing[44,45] to prevent the model from becoming over-confident.

During the development of Expanded TissueNet, we noted significant class imbalances with respect to imaging modalities. To prevent the model from learning representations that are overfit to any single modality, we implemented an auxiliary task that encourages the model to learn invariant representations across imaging platforms. To do so, we added a

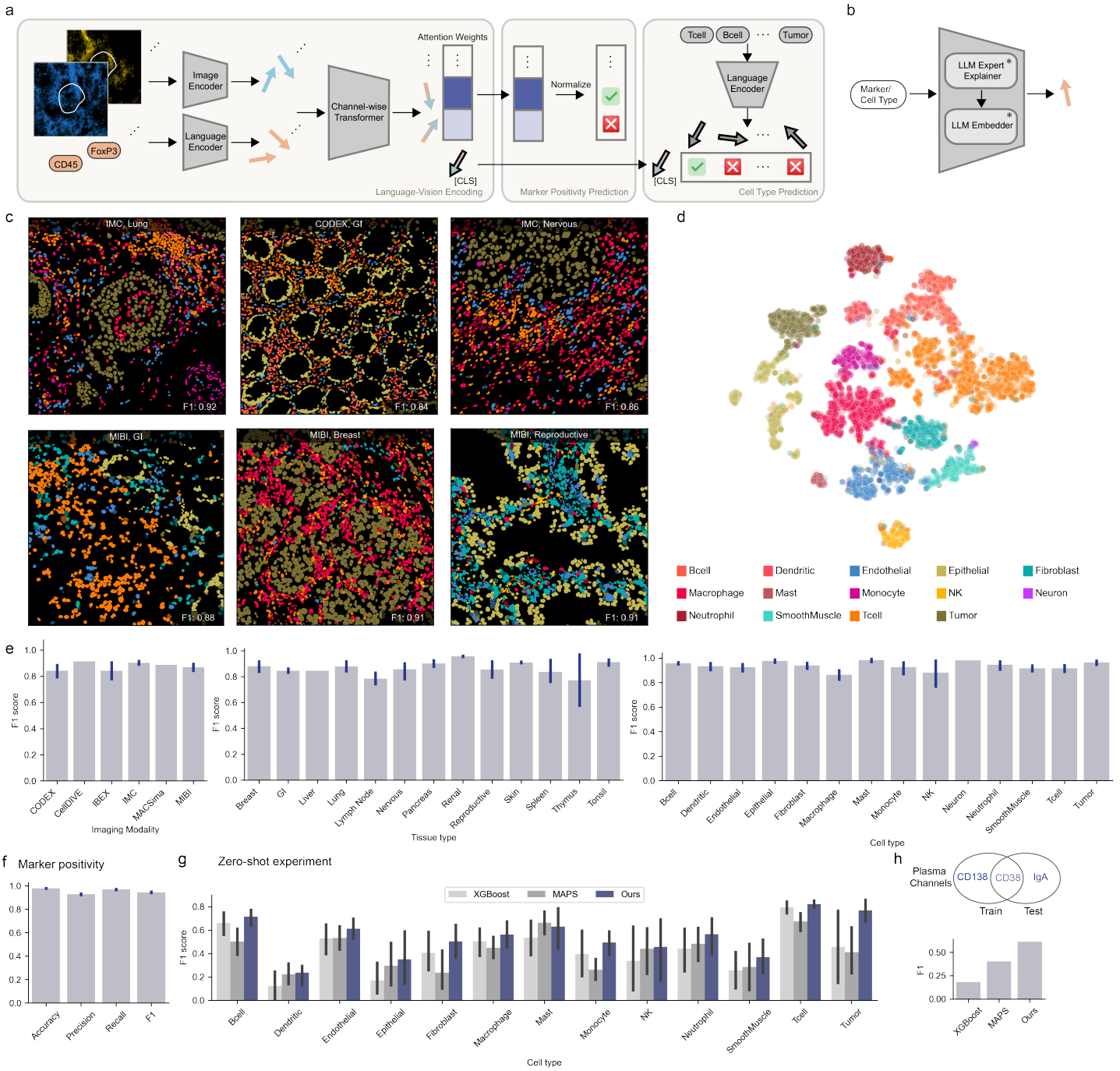


Figure 2: **DeepCell Types enables generalized cell phenotyping.** a) Model Design: Image patches and marker names are processed by image and language encoders to produce image embeddings (blue arrows) and text embeddings (orange arrows), respectively. A channel-wise transformer module combines these embeddings, generating marker (blue-orange blended arrows) and cell representations (`[CLS]` arrow). Attention weights predict marker positivity, while the `[CLS]` token's embedding is used for contrastive cell type prediction, enabling flexible processing of varied marker panels. b) Language Encoder: An LLM expert explainer retrieves relevant knowledge by prompting an LLM to generate detailed descriptions of queried markers or cell types. An LLM embedder then converts these descriptions into vector representations. c) Example Field of Views (FOVs) with cell types colorized, error predictions are masked by crossed lines. d) Latent Space Visualization: Cell types form distinct clusters, demonstrating our model's ability to learn biologically relevant features independent of imaging modalities or dataset origins. e) Classification performance analyzed by imaging modalities, tissue types and cell types. DeepCell Type's model architecture and diverse dataset facilitate generalization. f) Marker positivity performance. g-h) Comparison of zero-shot generalization performance against baselines. Models were evaluated by holding out a single dataset before training and evaluating performance on the held-out dataset. Each dataset was held out once; we report the average performance across all models. DeepCellTypes outperforms existing methods and can generalize well even when alternative markers are used to identify a cell type.

classification head that takes [CLS] token embeddings and predicts imaging modalities and coupled it with a gradient reversal layer[46]. During backpropagation, the gradients were reversed, thus teaching the model to 'unlearn' modality-specific differences. We found the resulting model developed more robust and general features and focused on the underlying biological characteristics of cells rather than platform-specific features (Fig. 2d, Fig. S2c).

2.4 Benchmarking

We sought to evaluate our model's performance across diverse datasets against other state-of-art approaches. Our analysis encompassed various imaging modalities, tissue types, and cell types, providing a comprehensive view of the model's effectiveness. As illustrated in Fig. 2e, our model demonstrates robust performance across all these dimensions. Notably, the model's performance remained strong even for underrepresented imaging modalities like MACSima, highlighting its ability to generalize beyond the dominant data sources. Fig. 2c showcases example images, visually demonstrating the model's efficacy across diverse inputs. Additionally, the model also performs well in marker positivity prediction, as evidenced in Fig. 2f.

To assess the modality-invariance of learned features, we applied a two-step dimensionality reduction technique to the cell embeddings: Neighborhood Components Analysis (NCA)[47] followed by t-SNE[48]. This approach, previously shown to be effective in revealing latent space structure without overfitting[49], allowed us to visualize the organization of our model's latent space. The resulting visualization, presented in Fig. 2d, reveals that the embedding space is primarily organized by cell types rather than imaging modalities (Fig. S2c). This organization demonstrates the model's focus on biologically relevant features over platform-specific characteristics.

We conducted a series of hold-out experiments to evaluate our model's zero-shot generalization capabilities. In each experiment, we excluded one dataset from training, trained the model on the remaining data, and then tested its performance on the held-out dataset. We benchmarked our model against two baselines: XGBoost[50], a tree-boosting algorithm, and MAPS[20], a neural network-based approach. We note that in this evaluation, differences in marker panels are effectively flagged as missing data for XGBoost and set to zero for MAPS. As shown in Fig. 2g, our model demonstrates favorable performance compared to the baselines across most cell types. This demonstrates that our language-informed vision model effectively leverages semantic understanding to achieve superior generalization across marker panels compared to previous approaches. Moreover, it underscores the utility of integrating language in vision information to meet the challenges of heterogeneous spatial proteomics data.

Another key advantage of our language-informed approach is the model's ability to generalize to unseen markers. We demonstrated this capability using the GVHD dataset[40], where plasma cells are identified using a unique marker (IgA) instead of the more common CD138 marker. Despite sharing only the CD38 marker with other datasets (which is not exclusive to plasma cells), our model robustly recognizes plasma cells in the GVHD dataset when trained on all other datasets (Fig. 2g). This further showcases the model's ability to understand the semantic information inherent in the names of markers and cell types, enabling effective generalization to new marker panels.

3 Discussion

This work addresses a critical challenge in spatial proteomics: the need for automated, generalizable, and scalable cell phenotyping tools. Our approach directly tackled the fundamental issue of variability in marker panels across different ex-

periments, enabling learning from multiple data sources. Trained on a diverse dataset, our model demonstrated robust and accurate performance across various experimental conditions, outperforming existing methods in zero-shot generalization tasks.

Our model's use of raw image data as input represents a significant advancement over traditional cell phenotyping approaches originated from non-spatial single-cell technologies like scRNA-Seq and flow cytometry. These methods rely on mean intensity values extracted from segmented cells and often fall short in coping with the artifacts of spatial proteomic data. Our work adds to the growing body of evidence that image-centric approaches are more robust to segmentation errors, noises, and signal spillovers[18,26].

A key innovation in our model is the integration of language components to enhance generalization. By incorporating textual information about cell types and markers, we leverage broader biological knowledge that extends beyond the training data. The synergy between visual and linguistic information allows our language-informed vision model to integrate information across a broad set of experimental data, achieving superior performance compared to traditional machine-learning approaches. A key advantage of our approach is its ability to train on multiple datasets simultaneously, leveraging marker correlations across datasets, a capability lacking in existing methods. This unified framework allows continuous model improvement as new data emerges, consolidating the field's knowledge into a single, increasingly comprehensive model. Given the success of language-informed vision models in handling the heterogeneous marker panels here, applying this methodology to image-based spatial transcriptomics or marker-aware cell segmentation would be a natural extension of this work.

Despite these advances, our method has limitations that merit discussion. First, our zero-shot experiment demonstrates a common characteristic of machine learning systems: while the performance is optimal within the domain of training data, it inevitably degrades when encountering samples too far out-of-domain. Despite our best efforts, Expanded TissueNet does not exhaust all imaging modalities and tissue types. Hence, even though we showed promising improvement compared to existing methods, datasets that substantially diverge from ExpandedTissueNet may require additional labeling and model finetuning to achieve adequate performance. We anticipate our labeling software and human-in-the-loop approaches to labeling will enable these efforts and facilitate the collection and labeling of increasingly diverse data, further improving model performance. As demonstrated here and previously[13], integrated data and model development is viable for consortium-scale deep learning in the life sciences. Second, while our work enables generalization across marker panels, generalization across cell types remains an open challenge, given the wide array of tissue-specific cell types. Addressing this limitation is a crucial direction for future investigation. Last, our model is trained on cell patches and cannot access the full image. While constraining the model's effective receptive field can help generalization, it can limit accuracy as features at longer length scales like functional tissue units and anatomical structures can provide context to aid cell type determination. Multi-scale language-informed vision models that integrate local and global features may offer a novel avenue for enhanced performance.

In conclusion, DeepCellTypes marks a significant advancement in spatial proteomics data analysis. By directly addressing the challenge of generalization across marker panels, we enabled accurate cell-type labeling for spatial proteomics data generated throughout the cellular imaging community and laid the groundwork for future innovations in cellular image analysis.

4 Code and data availability

Source code for model inference is available at `https://github.com/vanvalenlab/deepcell-types`. Instruction for downloading the pretrained model weights and a subset of ExpandedTissueNet that includes all data sourced from public datasets (5.2 million cells) is available at `https://vanvalenlab.github.io/deepcell-types`. The remaining datasets were made available to our lab before their publication to improve model performance. These are available upon reasonable request and will be made publicly available upon publication of the corresponding manuscripts. The code to reproduce the figures included in this paper is available at `https://github.com/vanvalenlab/DeepCellTypes-2024_Wang_et_al`. Note that the reproduced figures are generated using only the public datasets and may show minor variations from the manuscript figures, which incorporate the complete dataset.

5 Acknowledgements

We thank Noah Greenwald, Michael Angelo, and Sean Bendall, Georgia Gkioxari, Edward Pao, Uriah Israel, Ellen Emerson and the other members of the Van Valen lab for helpful feedback and interesting discussions. We thank John Hickey and Jean Fan for contributing novel datasets. This work was supported by awards from the National Institutes of Health awards OT2OD033756 (to KB, subaward to DVV), OT2OD033759 (to KB), DP2-GM149556 (to DVV); the Enoch foundation research fund (to LK); the Abisch-Frenkel foundation (LK); the Rising Tide foundation (LK); the Sharon Levine Foundation (to LK); the Schwartz/Reisman Collaborative Science Program (to DVV and LK); the European Research Council (948811) (to LK); the Israel Science Foundation (2481/20, 3830/21) (to LK); and the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center (to LK); the Shurl and Kay Curci Foundation (to DVV); the Rita Allen Foundation (to DVV), the Susan E Riley Foundation (to DVV); the Pew-Stewart Cancer Scholars program (to DVV); the Gordon and Betty Moore Foundation (to DVV); the Schmidt Academy for Software Engineering (to KY, SL); the Michael J Fox Foundation through the Aligning Science Across Parkinsons consortium (to DVV); the Heritage Medical Research Institute (to DVV); and the HHMI Freeman Hrabowski Scholar Program (to DVV).

6 Author Contributions

XW and DVV conceived the project; XW, YY, and DVV developed the DeepCellTypes model architecture; XW, RD, and EP developed data engineering infrastructure; YB, YJ, KB, LK contributed novel datasets and provided insight on image labeling; YB and CB performed image labeling; KY, SL, and RB developed cell phenotyping capabilities for DeepCell Label; RB oversaw software engineering and deployment within HuBMAP; DVV supervised the project.

7 Declaration of Interests

DVV is a co-founder of Aizen Therapeutics and holds equity in the company.

8 Methods

8.1 Data preprocessing and standardization

We implemented several preprocessing steps that were applied to each marker image independently to ensure consistent and high-quality input for our model. All images were resized to a resolution of 0.5 microns per pixel (mpp). We normalized the images by scaling pixel values based on the 99th percentile of all non-zero values across FOVs for each channel and each dataset. The resulting images were subsequently clipped between 0 and 5 to mitigate the impact of extremely bright pixels. Processed images were saved in zarr format, enabling chunked, compressed storage and fast loading. We also standardized marker names and cell type labels across datasets, identifying and combining alternative names to ensure consistency.

Cell segmentation was performed using the Mesmer algorithm[13]. We identified nuclear channels (e.g., DAPI, Histone H3) and cytoplasm/membrane channels (e.g., Pan-Keratin, CD45) to generate accurate whole-cell masks. For each segmented cell, we extracted a 64x64 pixel patch centered on the cell. For cells near image boundaries, we added padding to ensure complete 64x64 patches. To capture segmentation information, we appended a self-mask and a neighbor-mask to each channel's raw image. This resulted in images of shape $(C, 3, 64, 64)$, where C is the number of marker images varying by dataset. We zero-padded the tensor to $C_{max} = 87$ to allow processing by regular neural networks. A binary padding mask of $(C_{max},)$ is also generated for later use in the transformer module. Marker positivity labels were generated by manually gating the mean signal intensity for each marker for each dataset.

8.2 Model architecture

Our model employs a hybrid architecture, combining Convolutional Neural Network (CNN) and Transformer components to effectively process both image and textual data. Our model architecture has four components:

- *Image encoder*: The image encoder consists of an 11-layer CNN with 2D convolution layers, followed by SiLU activation and batch normalization. We reshape the input tensors from $(B, C_{max}, 3, 64, 64)$ to $(B * C_{max}, 3, 64, 64)$, allowing a single CNN to process all channels regardless of their marker correspondence. This design ensures the model remains agnostic to specific marker representations. The CNN output is then reshaped into embeddings of size $(B, C_{max}, 256)$.
- *Text encoder*: For the text encoder, we utilized OpenAI's GPT-4 model[51] as the expert explainer and the GPT-3[52] text-embedding-3-large model as the embedder. We prompt the explainer to provide detailed descriptions of each marker type including its general description, alternative names, and cell type correspondence. The resulting 1024-dimensional embeddings are linearly mapped to 256 dimensions to match the image embeddings.
- *Channel-wise transformer*: To merge information contained in the image and text embeddings, we fed them into a transformer module that uses channel-wise attention. A learnable `[CLS]` token was appended to the embedding tensor to represent the entire cell. The transformer module comprises 5 encoder layers with a hidden dimension of 256 and a feed-forward dimension of 512. We employ a padding mask to exclude dummy channels, ensuring the model focuses only on real channels. This flexible mechanism accommodates varying numbers of input channels.
- *Gradient reversal*: We employed a gradient reversal technique to mitigate systematic bias induced by different imaging modalities. The `[CLS]` token is fed into a 3-layer MLP classification head for predicting imaging modalities. During backpropagation, we reverse the gradient of the first layer. This approach minimizes the loss on cell type classification

while maximizing the loss on imaging modality classification, effectively removing modality-specific information from the learned representations.

To determine marker positivity, the attention weights between the `[CLS]` token and all other tokens were extracted, normalized to the range [0, 1], and interpreted as marker positivity values. For cell type classification, we first extracted the `[CLS]` token embedding from the output layer and reused the text encoder to convert cell type names into embeddings. We trained the `[CLS]` token and cell type name tokens using the contrastive (e.g., CLIP) loss [42], calculating cosine similarities with a ground truth similarity of 1 for the correct cell type and 0 for all others.

8.3 Model training

Our training strategy incorporates three distinct losses: cell type classification, marker positivity prediction, and reverse imaging modality classification. We assign constant weights to the first two components while employing a ramping weight for the third, proportional to the quartic root of the number of epochs. This approach allows the model to initially learn representations best suited for cell type classification and marker positivity prediction before gradually moving towards modality-invariant features to enhance generalization performance.

For cell type classification, we implement a focal version of the contrastive loss [42,43] with gamma = 2.0. This adaptation addresses class imbalance and improves performance on more challenging categories. The marker positivity prediction utilizes binary cross-entropy loss with label smoothing = 0.2, a technique known to prevent overfitting [44,45].

Algorithm 1 Focal CLIP Contrastive Loss

procedure FocalCLIPLoss(I, T, γ, τ)
  Input: Image embeddings I
  Input: Text embeddings T
  Input: Focusing parameter γ
  Input: Learnable scaling parameter τ
  $I \leftarrow I/\|I\|_2, T \leftarrow T/\|T\|_2$ ▷ Normalize embeddings
  $S \leftarrow IT^\top/\tau$ ▷ Similarity matrix
  $P_{img} \leftarrow \text{softmax}(S)$
  $P_{txt} \leftarrow \text{softmax}(S^\top)$
  $Y \leftarrow [0, 1, \ldots, N-1]$ ▷ Groundtruth indices
  procedure FocalLoss(P, Y)
    $ce \leftarrow \text{NLLLoss}(P, Y)$
    $P_t \leftarrow P[i, Y[i]]$ for all i ▷ Probabilities of all true classes
    $loss \leftarrow (1 - P_t)^\gamma \cdot ce$ ▷ Focal-weighted loss
    return mean(loss)
  end procedure
  $L_{img} \leftarrow \text{FocalLoss}(L_{img})$
  $L_{txt} \leftarrow \text{FocalLoss}(L_{txt})$
  $L \leftarrow (L_{img} + L_{txt})/2$ ▷ Symmetric loss
  return L
end procedure

We train the model for 15 epochs using the RAdam optimizer with a learning rate of 10^{-4}. We employ several data augmentation techniques to boost generalization, including random flipping, rotation, and image resizing. Further, we randomly drop out 8 marker channels during training to encourage the model to learn robust features that are less dependent on specific markers. We also added random gaussian noises to marker name embeddings and cell type embeddings to prevent overfitting and to improve robustness.

8.4 Baselines

To benchmark our model's performance, we implemented two established approaches: XGBoost [50] and MAPS [20]. We began by extracting features from each dataset and calculating the mean intensity value for each available marker channel in every cell. Next, we compiled a list of all unique markers present across all datasets, which served as a universal reference for data alignment. We then harmonized the data by aligning each dataset to this universal marker list. For markers present in a dataset, we used the calculated mean intensity values, while for absent markers, we inserted a placeholder value ('nan' for XGBoost, 0 for MAPS).

We implemented XGBoost using the Python XGBoost package (`https://github.com/dmlc/xgboost`). Placeholder values were explicitly indicated through the data matrix interface. We trained the XGBoost model for 200 epochs using default parameters.

For MAPS, we used the implementation available at `https://github.com/mahmoodlab/MAPS`. In addition to the marker expression data, we appended a column representing cell size to the input matrix, as per the MAPS protocol. The MAPS model was trained using default parameters for 500 epochs.

References

[1] Keren, L.; Bosse, M.; Thompson, S.; Risom, T.; Vijayaragavan, K.; McCaffrey, E.; Marquez, D.; Angoshtari, R.; Greenwald, N. F.; Fienberg, H.; others MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances* **2019**, *5*, eaax5851.

[2] Goltsev, Y.; Samusik, N.; Kennedy-Darling, J.; Bhate, S.; Hale, M.; Vazquez, G.; Black, S.; Nolan, G. P. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **2018**, *174*, 968–981.

[3] Giesen, C.; Wang, H. A.; Schapiro, D.; Zivanovic, N.; Jacobs, A.; Hattendorf, B.; Schüffler, P. J.; Grolimund, D.; Buhmann, J. M.; Brandt, S.; others Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **2014**, *11*, 417–422.

[4] Kinkhabwala, A.; Herbel, C.; Pankratz, J.; Yushchenko, D. A.; Rüberg, S.; Praveen, P.; Reiß, S.; Rodriguez, F. C.; Schäfer, D.; Kollet, J.; others MACSima imaging cyclic staining (MICS) technology reveals combinatorial target pairs for CAR T cell treatment of solid tumors. *Scientific reports* **2022**, *12*, 1911.

[5] Radtke, A. J.; Kandov, E.; Lowekamp, B.; Speranza, E.; Chu, C. J.; Gola, A.; Thakur, N.; Shih, R.; Yao, L.; Yaniv, Z. R.; others IBEX: A versatile multiplex optical imaging approach for deep phenotyping and spatial analysis of cells in complex tissues. *Proceedings of the National Academy of Sciences* **2020**, *117*, 33455–33465.

[6] Gerdes, M. J.; Sevinsky, C. J.; Sood, A.; Adak, S.; Bello, M. O.; Bordwell, A.; Can, A.; Corwin, A.; Dinn, S.; Filkins, R. J.; others Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences* **2013**, *110*, 11982–11987.

[7] Chen, K. H.; Boettiger, A. N.; Moffitt, J. R.; Wang, S.; Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **2015**, *348*, aaa6090.

[8] Moffitt, J. R.; Hao, J.; Wang, G.; Chen, K. H.; Babcock, H. P.; Zhuang, X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* **2016**, *113*, 11046–11051.

[9] Lubeck, E.; Coskun, A. F.; Zhiyentayev, T.; Ahmad, M.; Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature methods* **2014**, *11*, 360–361.

[10] Eng, C.-H. L.; Lawson, M.; Zhu, Q.; Dries, R.; Koulena, N.; Takei, Y.; Yun, J.; Cronin, C.; Karp, C.; Yuan, G.-C.; others Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **2019**, *568*, 235–239.

[11] Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; others Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **2024**, *30*, 850–862.

[12] Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; others Towards a general-purpose foundation model for computational pathology. *Nature Medicine* **2024**, *30*, 850–862.

[13] Greenwald, N. F.; Miller, G.; Moen, E.; Kong, A.; Kagel, A.; Dougherty, T.; Fullaway, C. C.; McIntosh, B. J.; Leow, K. X.; Schwartz, M. S.; others Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology* **2022**, *40*, 555–565.

[14] Israel, U.; Marks, M.; Dilip, R.; Li, Q.; Yu, C.; Laubscher, E.; Li, S.; Schwartz, M.; Pradhan, E.; Ates, A.; others A foundation model for cell segmentation. *bioRxiv* **2023**,

[15] Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods* **2021**, *18*, 100–106.

[16] Berg, S.; Kutra, D.; Kroeger, T.; Straehle, C. N.; Kausler, B. X.; Haubold, C.; Schiegg, M.; Ales, J.; Beier, T.; Rudy, M.; others Ilastik: interactive machine learning for (bio) image analysis. *Nature methods* **2019**, *16*, 1226–1232.

[17] Schmidt, U.; Weigert, M.; Broaddus, C.; Myers, G. Cell detection with star-convex polygons. Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. 2018; pp 265–273.

[18] Amitay, Y.; Bussi, Y.; Feinstein, B.; Bagon, S.; Milo, I.; Keren, L. CellSighter: a neural network to classify cells in highly multiplexed images. *Nature communications* **2023**, *14*, 4302.

[19] Brbić, M.; Cao, K.; Hickey, J. W.; Tan, Y.; Snyder, M. P.; Nolan, G. P.; Leskovec, J. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods* **2022**, 1–8.

[20] Shaban, M.; Bai, Y.; Qiu, H.; Mao, S.; Yeung, J.; Yeo, Y. Y.; Shanmugam, V.; Chen, H.; Zhu, B.; Weirather, J. L.; others MAPS: Pathologist-level cell type annotation from tissue images through machine learning. *Nature Communications* **2024**, *15*, 28.

[21] Van Gassen, S.; Callebaut, B.; Van Helden, M. J.; Lambrecht, B. N.; Demeester, P.; Dhaene, T.; Saeys, Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **2015**, *87*, 636–645.

[22] Schapiro, D.; Jackson, H. W.; Raghuraman, S.; Fischer, J. R.; Zanotelli, V. R.; Schulz, D.; Giesen, C.; Catena, R.; Varga, Z.; Bodenmiller, B. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods* **2017**, *14*, 873–876.

[23] Bankhead, P.; Loughrey, M. B.; Fernández, J. A.; Dombrowski, Y.; McArt, D. G.; Dunne, P. D.; McQuaid, S.; Gray, R. T.; Murray, L. J.; Coleman, H. G.; others QuPath: Open source software for digital pathology image analysis. *Scientific reports* **2017**, *7*, 1–7.

[24] Zhang, W.; Li, I.; Reticker-Flynn, N. E.; Good, Z.; Chang, S.; Samusik, N.; Saumyaa, S.; Li, Y.; Zhou, X.; Liang, R.; others Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nature Methods* **2022**, *19*, 759–769.

[25] Geuenich, M. J.; Hou, J.; Lee, S.; Ayub, S.; Jackson, H. W.; Campbell, K. R. Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data. *Cell Systems* **2021**, *12*, 1173–1186.

[26] Rumberger, L.; Greenwald, N. F.; Ranek, J.; Boonrat, P.; Walker, C.; Franzen, J.; Varra, S.; Kong, A.; Sowers, C.; Liu, C. C.; others Automated classification of cellular expression in multiplexed imaging data with Nimbus. *bioRxiv* **2024**, 2024–06.

[27] Hickey, J. W.; Becker, W. R.; Nevins, S. A.; Horning, A.; Perez, A. E.; Zhu, C.; Zhu, B.; Wei, B.; Chiu, R.; Chen, D. C.; others Organization of the human intestine at single-cell resolution. *Nature* **2023**, *619*, 572–584.

[28] Hartmann, F. J.; Mrdjen, D.; McCaffrey, E.; Glass, D. R.; Greenwald, N. F.; Bharadwaj, A.; Khair, Z.; Verberk, S. G.; Baranski, A.; Baskar, R.; others Single-cell metabolic profiling of human cytotoxic T cells. *Nature biotechnology* **2021**, *39*, 186–197.

[29] Liu, C. C.; Bosse, M.; Kong, A.; Kagel, A.; Kinders, R.; Hewitt, S. M.; Varma, S.; van de Rijn, M.; Nowak, S. H.; Bendall, S. C.; others Reproducible, high-dimensional imaging in archival human tissue by multiplexed ion beam imaging by time-of-flight (MIBI-TOF). *Laboratory Investigation* **2022**, *102*, 762–770.

[30] Aleynick, N.; Li, Y.; Xie, Y.; Zhang, M.; Posner, A.; Roshal, L.; Pe'er, D.; Vanguri, R. S.; Hollmann, T. J. Cross-platform dataset of multiplex fluorescent cellular object image annotations. *Scientific Data* **2023**, *10*, 193.

[31] Keren, L.; Bosse, M.; Marquez, D.; Angoshtari, R.; Jain, S.; Varma, S.; Yang, S.-R.; Kurian, A.; Van Valen, D.; West, R.; others A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **2018**, *174*, 1373–1387.

[32] Risom, T.; Glass, D. R.; Averbukh, I.; Liu, C. C.; Baranski, A.; Kagel, A.; McCaffrey, E. F.; Greenwald, N. F.; Rivero-Gutiérrez, B.; Strand, S. H.; others Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* **2022**, *185*, 299–310.

[33] Sorin, M.; Rezanejad, M.; Karimi, E.; Fiset, B.; Desharnais, L.; Perus, L. J.; Milette, S.; Yu, M. W.; Maritan, S. M.; Doré, S.; others Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature* **2023**, *614*, 548–554.

[34] McCaffrey, E. F.; Donato, M.; Keren, L.; Chen, Z.; Delmastro, A.; Fitzpatrick, M. B.; Gupta, S.; Greenwald, N. F.; Baranski, A.; Graf, W.; others The immunoregulatory landscape of human tuberculosis granulomas. *Nature immunology* **2022**, *23*, 318–329.

[35] Karimi, E.; Yu, M. W.; Maritan, S. M.; Perus, L. J.; Rezanejad, M.; Sorin, M.; Dankner, M.; Fallah, P.; Doré, S.; Zuo, D.; others Single-cell spatial immune landscapes of primary and metastatic brain tumours. *Nature* **2023**, *614*, 555–563.

[36] Greenbaum, S.; Averbukh, I.; Soon, E.; Rizzuto, G.; Baranski, A.; Greenwald, N. F.; Kagel, A.; Bosse, M.; Jaswa, E. G.; Khair, Z.; others A spatially resolved timeline of the human maternal–fetal interface. *Nature* **2023**, *619*, 595–605.

[37] Ghose, S.; Ju, Y.; McDonough, E.; Ho, J.; Karunamurthy, A.; Chadwick, C.; Cho, S.; Rose, R.; Corwin, A.; Surrette, C.; others 3D reconstruction of skin and spatial mapping of immune cell density, vascular distance and effects of sun exposure and aging. *Communications Biology* **2023**, *6*, 718.

[38] Müller, W.; Rüberg, S.; Bosio, A. OMAP-10: Multiplexed antibody-based imaging of human Palatine Tonsil with MACSima v1.0. 2023.

[39] Radtke, A. J.; Chu, C. J.; Yaniv, Z.; Yao, L.; Marr, J.; Beuschel, R. T.; Ichise, H.; Gola, A.; Kabat, J.; Lowekamp, B.; others IBEX: an iterative immunolabeling and chemical bleaching method for high-content imaging of diverse tissues. *Nature protocols* **2022**, *17*, 378–401.

[40] Azulay, N.; Milo, I.; Bussi, Y.; Ben Uri, R.; Keidar Haran, T.; Eldar, M.; Elhanani, O.; Harnik, Y.; Yakubovsky, O.; Nachmany, I.; others A spatial atlas of human gastro-intestinal acute GVHD reveals epithelial and immune dynamics underlying disease pathophysiology. *bioRxiv* **2024**, 2024–09.

[41] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

[42] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others Learning transferable visual models from natural language supervision. International conference on machine learning. 2021; pp 8748–8763.

[43] Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2017; pp 2980–2988.

[44] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 2818–2826.

[45] Müller, R.; Kornblith, S.; Hinton, G. E. When does label smoothing help? *Advances in neural information processing systems* **2019**, *32*.

[46] Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. International conference on machine learning. 2015; pp 1180–1189.

[47] Goldberger, J.; Hinton, G. E.; Roweis, S.; Salakhutdinov, R. R. Neighbourhood components analysis. *Advances in neural information processing systems* **2004**, *17*.

[48] Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.

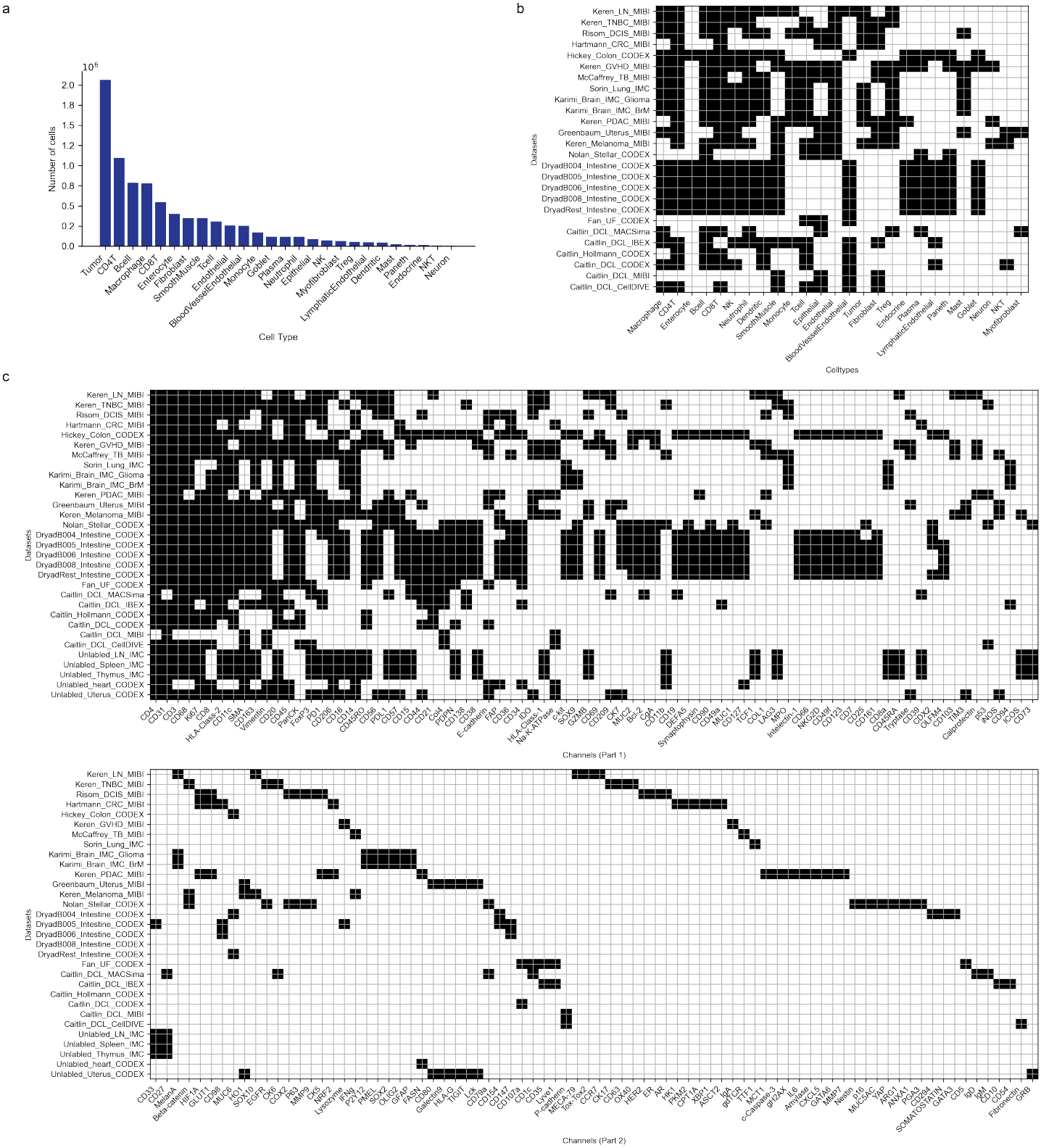
[49] Booeshaghi, A. S.; Yao, Z.; van Velthoven, C.; Smith, K.; Tasic, B.; Zeng, H.; Pachter, L. Isoform cell-type specificity in the mouse primary motor cortex. *Nature* **2021**, *598*, 195–199.

[50] Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

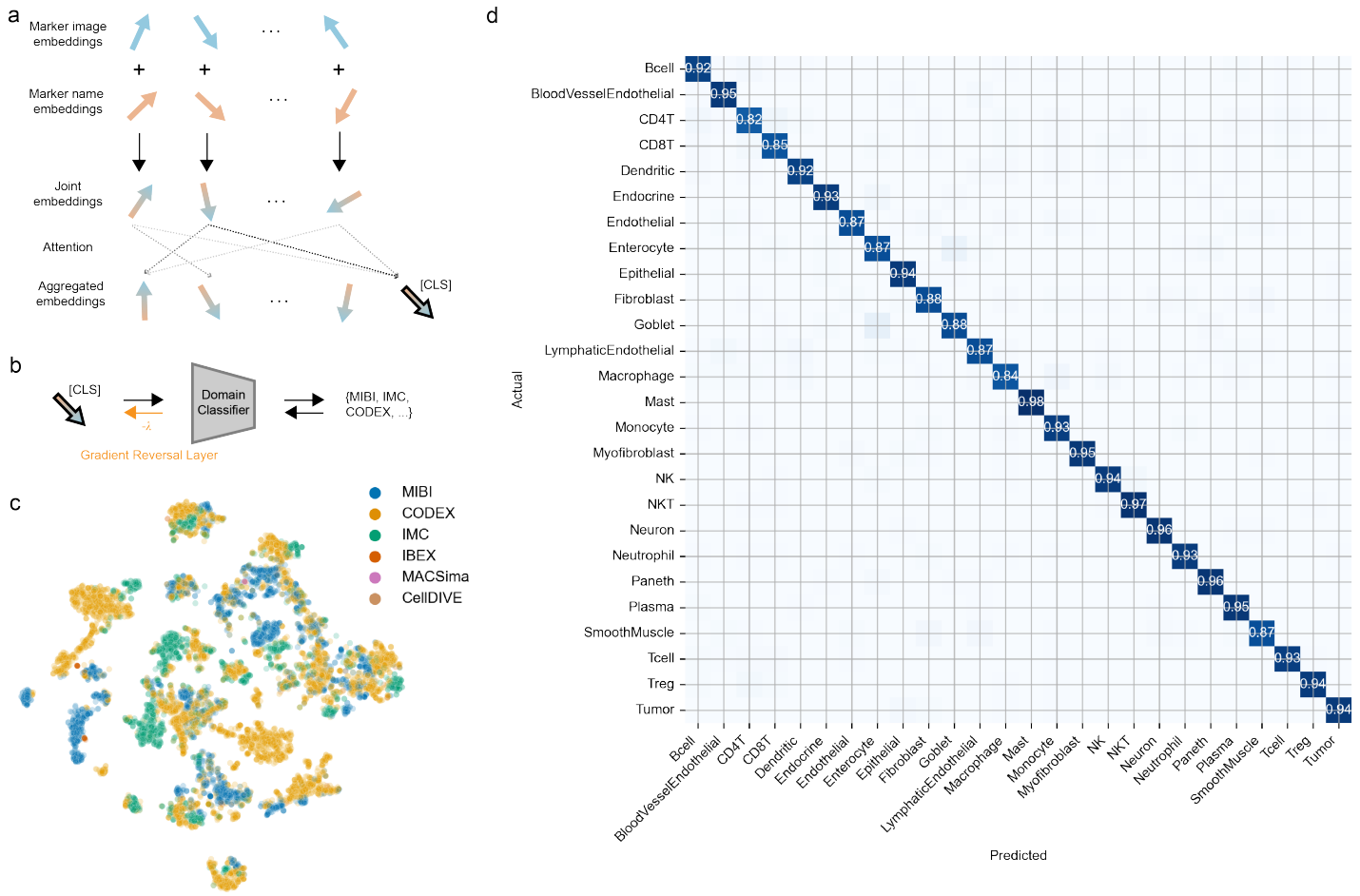
[51] Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; others Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**,

[52] Brown, T. B. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165* **2020**,

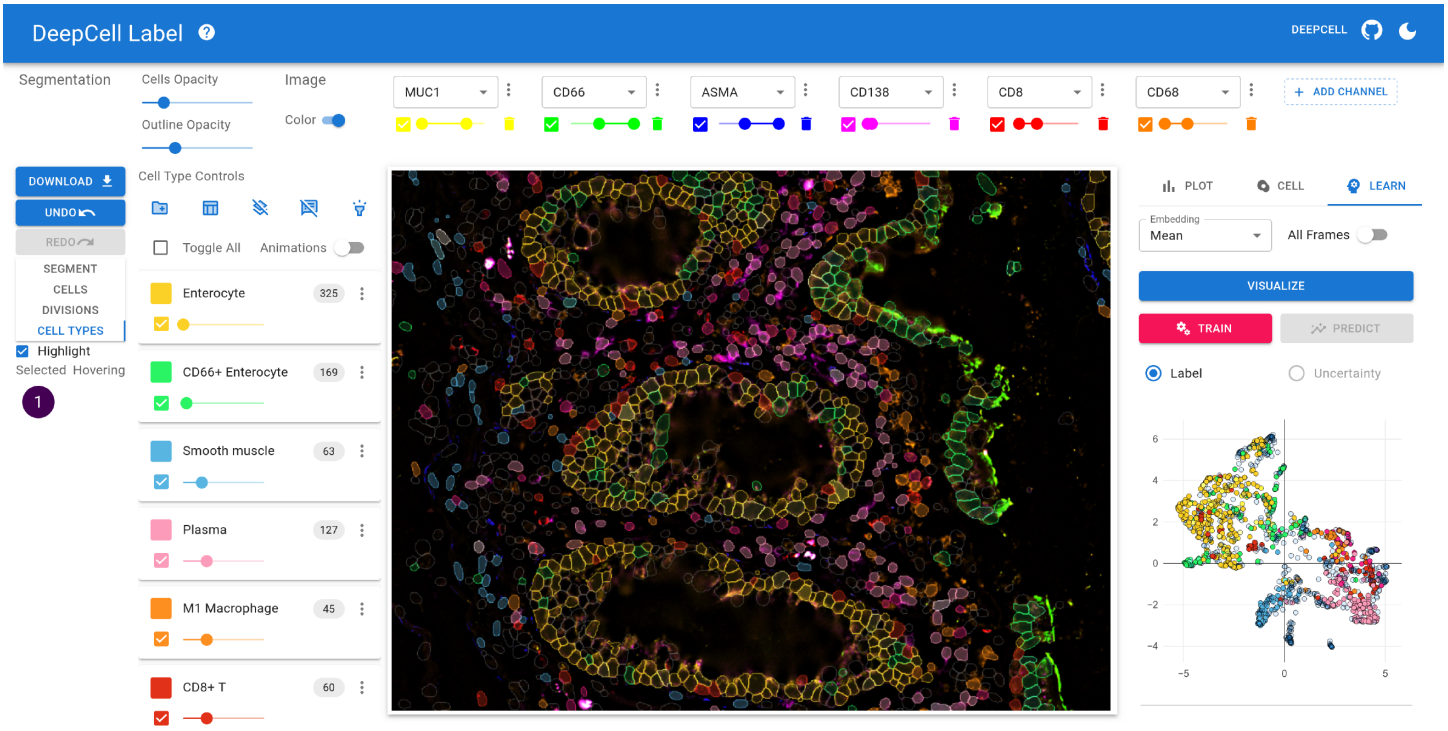
Supplementary Information



Supplementary Figure S1: **Expanded TissueNet statistics.** a) Distribution of cell type abundances within Expanded TissueNet. b, c) Heatmaps illustrating the presence (dark) or absence (light) of specific cell types (b) and markers (c) across all datasets in Expanded TissueNet.



Supplementary Figure S2: **Model architecture and performance.** a) A detailed schematic of the channel-wise attention mechanism. Marker image and name embeddings are added together to form joint embeddings. These serve as tokens for the attention mechanism, modeling inter-channel correlations. A `[CLS]` token that attends to all tokens represents all the information in the images and semantic content of the markers. b) Domain adversarial training architecture: A Gradient Reversal Layer between the encoder and domain classifier is employed. While the classifier attempts to identify imaging modalities, gradient reversal during backpropagation encourages the encoder to learn modality-invariant features. c) Latent space visualization: an NCA/t-SNE plot of the learned representations, color-coded by imaging modalities (MIBI, CODEX, IMC, IBEX, MACSima, CellDIVE). The intermixing of colors demonstrates the model achieved invariance to imaging modalities. Recoloring the plot by cell type (Fig. 2d) demonstrates that the latent space is organized by cell type instead. d) A comprehensive confusion matrix for cell type classification across all cell types.



Supplementary Figure S3: **An improved DeepCell Label Interface for spatial proteomic data.** The DeepCell Label interface facilitates expert-in-the-loop development of cell classification models. This browser-native software integrates multichannel visualization, cell-type annotation tools, statistical analysis, and in-browser classifier training based on cell embeddings. With DeepCell Label, we can perform iterative refinement of cell type and marker positivity labels.

Our Category	Ontology ID	Ontology Name
Tcell	CL:0000084	T cell
Treg	CL:0000815	regulatory T cell
CD4T	CL:0000624	CD4-positive, alpha-beta T cell
CD8T	CL:0000625	CD8-positive, alpha-beta T cell
NKT	CL:0000814	mature NK T cell
Bcell	CL:0000236	B cell
Plasma	CL:0000786	plasma cell
NK	CL:0000623	natural killer cell
Dendritic	CL:0000451	dendritic cell
Mast	CL:0000097	mast cell
Neutrophil	CL:0000775	neutrophil
Macrophage	CL:0000235	macrophage
Monocyte	CL:0000576	monocyte
Epithelial	CL:0000066	epithelial cell
Goblet	CL:0000160	goblet cell
Paneth	CL:0000510	paneth cell
Enterocyte	CL:0000584	enterocyte
Endocrine	CL:0000163	endocrine cell
Endothelial	CL:0000115	endothelial cell
LymphaticEndothelial	CL:0002138	endothelial cell of lymphatic vessel
BloodVesselEndothelial	CL:0000071	blood vessel endothelial cell
Fibroblast	CL:0000057	fibroblast
Myofibroblast	CL:0000186	myofibroblast
SmoothMuscle	CL:0000192	smooth muscle
Nerve	CL:0002319	neural cell
Neuron	CL:0000540	neuron
Glia	CL:0000125	glial cell
Tumor	NA	NA

Supplementary Table S1: Mapping of our cell types to HuBMAP categories cased on Cell Ontology (CL).