# ImmuVis: Hyperconvolutional Foundation Models for Imaging Mass Cytometry

**Dawid Uchal**[1*] **Marcin Możejko**[1*] **Krzysztof Gogolewski**[1,2*] **Piotr Kupidura**[1] **Fabian Ozga**[1] **Szymon Łukasik**[1] **Jakub Giezgała**[1] **Tomasz Nocoń**[1] **Kacper Pietrzyk**[1] **Robert Pieniuta**[1] **Mateusz Sulimowicz**[1] **Michał Zmysłowski**[1] **Michał Orzyłowski**[1] **Tomasz Siłkowski**[1] **Karol Zagródka**[1] **Eike Staub**[3] **Ewa Szczurek**[1,4,5,6]
[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland, [2]Center of Trustworthy AI for Life Sciences - International Research Agendas Programme, University of Warsaw, Warsaw, Poland [3]Merck Healthcare KGaA, Darmstadt, Germany, [4]Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany, [5]Faculty of Mathematics, Informatics and Statistics, Ludwig-Maximilians-Universität München, Munich, Germany, [6]Munich Center for Machine Learning, *shared first author order determined by coin flips
em.szczurek@uw.edu.pl

## Abstract

We present ImmuVis, a family of **efficient foundation models** for imaging mass cytometry (IMC), a high-throughput multiplex imaging technology that handles molecular marker measurements as image channels and enables large-scale spatial tissue profiling. Unlike natural images, multiplex imaging lacks a fixed channel space, as real-world marker sets vary across studies, violating a core assumption of standard vision backbones. To address this, ImmuVis introduces **marker-adaptive hyperconvolutions** that generate convolutional kernels from learned marker embeddings, enabling a single model to operate on any subset of the markers seen during pretraining. We pretrain ImmuVis on **the largest dataset to date**, IMC17M (28 cohorts, 24,405 images, 265 markers, over 17M patches), using self-supervised masked reconstruction. ImmuVis achieves the best average cell-typing and strong virtual staining performance, while remaining competitive with state-of-the-art approaches on clinical prediction at substantially lower compute cost, and is the sole model that provides **calibrated uncertainty**. These results position ImmuVis as a practical framework for real-world IMC modeling.

## 1 Introduction

Imaging Mass Cytometry (IMC) profiles protein abundances in tissues at subcellular spatial resolution by measuring metal-tagged antibodies via mass spectrometry [1, 2]. IMC is increasingly central to spatially resolved single-cell analysis of healthy and disease tissues, and both datasets and preprocessing pipelines are rapidly maturing [3, 4]. Panel richness matters: larger marker panels better resolve cell types and functional states, enabling mapping of tissue spatial neighborhoods [5, 6].

Still, cohort-specific IMC panels are typically restricted to only tens of markers because of physical channel interference that reduces the number of concurrently usable metal isotopes for antibody tagging [7, 8]. In practice, as presented in Figure 1, cross-cohort marker overlap is limited, making **modelling and analysis across heterogeneous panels** a practical necessity [9]. This also motivates **virtual staining**, i.e. predicting cohort-dependent *missing-marker sets* from the *measured* panel, to enable denser tissue phenotyping without additional wet-lab work [10]. Basic approaches to this task suffer from a fixed input problem, i.e. are limited to a fixed, narrow marker panel [11, 12], which prevents panel flexibility and straightforward extension to diverse marker sets. VirTues [13]

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and Eva [14], recent Transformer-based models, support variable marker inputs by concatenating marker expression-specific spatial tokens augmented with marker-identity encodings derived from protein language models. However, since the spatial tokens provide only a global, flattened view per marker image patch, these models effectively ignore crucial local marker dependencies. In contrast, variable input learning in other vision domains is often implemented via conditional operators, e.g. hypernetworks [15] and dynamic convolutions [16], that modulate the feature extractor conditioned on the input in a computationally efficient manner; yet this approach remains underused in multiplex imaging.

Virtual staining is also a **reliability problem**: when the measured marker set provides weak evidence for a target marker (e.g., rare phenotypes or atypical tissue patterns), accurate point prediction is ill-posed and visually plausible reconstructions may still be biologically wrong [17]. In this regime, **uncertainty** must be a part of the model output, highlighting unconfident predictions. Finally, cohort-scale use places a strong emphasis on **inference time**. Transformer-heavy virtual-staining models can become increasingly slow as marker count and spatial resolution grow [18]. To summarize, to the best of our knowledge, no prior virtual-staining model combines **panel flexibility** and **calibrated uncertainty** with **inference speed** for IMC.

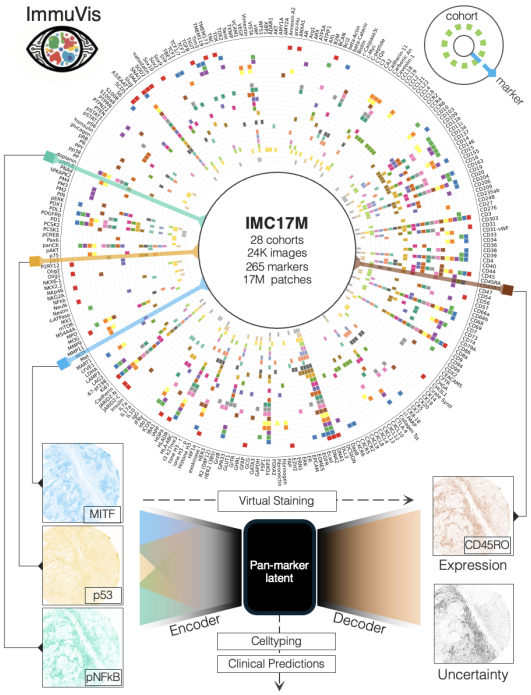



Figure 1: **Motivation for `ImmuVis`: real-world panel heterogeneity.** IMC17M exhibits strong cohort-marker diversity (rings: cohorts; radii: markers; colored ticks: measured markers), motivating a single model that operates on different marker subsets. `ImmuVis` encodes an observed panel (e.g., MITF, p53, pNFkB) into a shared pan-marker latent space for downstream phenotyping (cell typing, clinical prediction) and instantiates a task-specific decoder to virtually stain requested targets (e.g., CD45RO), outputting both expression and predictive uncertainty.

To address these challenges, we introduce `ImmuVis`, a family of foundation models for IMC data that unifies variable-input learning, efficiency, and uncertainty in a single architecture. Our contributions are as follows:

**Hyperconvolutional architecture**: We introduce a channel-adaptive hyperconvolution module that provides operator-level channel adaptivity and generates convolutional kernels conditioned on learned marker embeddings, enabling a single model that uses standard computer vision backbones to process any combination of the pretrained vocabulary without architectural modification or retraining.

**Family of efficient foundation models:** We introduce a family of foundation models based on our hyperconvolutional architecture, that uses different established computer vision backbones spanning from fully convolutional $\texttt{ImmuVis}_{\texttt{Conv}}$ that incorporates ConvNeXt v2 [19], through window-attention-based $\texttt{ImmuVis}_{\texttt{Swin}}$ [20], up to token-based $\texttt{ImmuVis}_{\texttt{ViT}}$ [21]. The entire family offers substantially lower inference cost than token-concatenation-based alternatives [13, 14].

**Uncertainty-aware learning objective:** We model predictive uncertainty via modified Gaussian heteroscedastic regression, producing uncertainty estimates that highly correlate with reconstruction error, making virtual staining reliability-aware rather than purely point-estimated.

We pretrain `ImmuVis` on the largest IMC dataset compiled to our knowledge (IMC17M; 24,405 images, 265 markers, 28 datasets) using masked-channel reconstruction, and show that the learned representations transfer across datasets to state-of-the-art performance in virtual staining and patch-level cell typing, with clinical prediction competitive with state-of-the-art approaches. Taken together, `ImmuVis` provides a practical, panel-flexible IMC foundation models family that is de-

ployable at cohort scale and enables reliability-aware virtual staining. More broadly, it offers a general recipe for variable-input foundation models in multiplex imaging and beyond. The code and pretrained models are available at `https://github.com/szczurek-lab/ImmuVis` and `https://huggingface.co/collections/SzczurekLab/immuvis`, respectively.

## 2 Related Work

**Handling heterogeneous panels and virtual staining for multiplex data.** Virtual staining aims at prediction of unmeasured markers, increasing phenotyping depth and enabling retrospective upgrading of legacy cohorts without additional wet-lab assays [13, 22]. Early approaches to this problem typically assume a fixed input panel and train U-Net-like architectures to predict new markers [12, 23]. However, these methods lack panel flexibility, as they are restricted to datasets whose panels contain markers used during training and cannot exploit additional available markers. To achieve panel flexibility, [14, 24] adopt a channel-extended ViT design from [25], which concatenates marker-specific spatial tokens with marker-identity encodings and processes them using a transformer model. However, this design incurs a heavy combinatorial marker$\times$space computational cost. VirTues [13] mitigates this limitation via marker-space attention factorization. Additionally, VirTues and EVA [14] use pretrained language models embeddings as marker encodings enabling predictions for previously unseen markers. Still, in these transformer tokenization-based designs, marker identity is injected *after* tokenization and therefore may underutilize fine-grained, local cross-marker dependencies that could be captured by conditioning the feature extractor itself.

**Foundation-model trends for multiplex imaging.** Large-scale self-supervised pretraining [26] has enabled transferable visual representations via masked prediction [13, 14, 24] or self-supervised distillation objectives [27]. For multiplex imaging, self-supervision is attractive because heterogeneous collections can be exploited without dense per-marker labels. However, as IMC introduces modality-specific artifacts (e.g., hot pixels and ion-counting noise), a recent self-supervised foundation model for spatial proteomics, KRONOS, explicitly excludes ion-based modalities such as IMC from training and notes that they may require architectural modifications or modality-specific adaptation strategies [28].

## 3 Methodology

### 3.1 `ImmuVis` architecture

Let $\mathcal{M} = \{m_1, \ldots, m_N\}$ denote the global pan-cohort marker vocabulary observed during the training. Given two *ordered* marker sets $\mathcal{I}_e, \mathcal{J}_d \subset \mathcal{M}$, indexed as $\mathcal{I}_e = (i^1, \ldots, i^{C_e})$ and $\mathcal{J}_d = (j^1, \ldots, j^{C_d})$, with $C_e = |\mathcal{I}_e|$ and $C_d = |\mathcal{J}_d|$, `ImmuVis` instantiates a marker-specific encoder-decoder pair:

$$\texttt{ImmuVis}\,(\mathcal{I}_e, \mathcal{J}_d) = \left(\mathrm{Enc}^{\mathcal{I}_e}, \mathrm{Dec}^{\mathcal{J}_d}\right).$$

The encoder $\mathrm{Enc}^{\mathcal{I}_e}$ maps the multiplex image $\mathbf{X} = (\mathbf{X}_{i^1}, \ldots, \mathbf{X}_{i^{C_e}}) \in \mathbb{R}^{C_e \times H_0 \times W_0}$, to a pan-marker latent representation $\mathbf{Z} = \mathrm{Enc}^{\mathcal{I}_e}(\mathbf{X}) \in \mathbb{R}^{d_{lat} \times H_{lat} \times W_{lat}}$ shared across the input markers. The decoder $\mathrm{Dec}^{\mathcal{J}_d}$ maps $\mathbf{Z}$ to image-level point-wise predictions for the markers in $\mathcal{J}_d$, producing the mean $\mathbf{X}^{\mu}$ and the corresponding log-variance $\mathbf{X}^{\log \sigma^2}$:

$$(\mathbf{X}^{\mu},\, \mathbf{X}^{\log \sigma^2}) = \mathrm{Dec}^{\mathcal{J}_d}(\mathbf{Z}) \in \mathbb{R}^{2 \times C_d \times H_0 \times W_0}.$$

`ImmuVis` architecture overview is presented in Figure 2.

#### 3.1.1 $\mathrm{Enc}^{\mathcal{I}_e}$ hypernetwork

Given $\mathcal{I}_e$, the instantiated encoder factors to a composition:

$$\mathbf{Z} = \mathrm{Enc}^{\mathcal{I}_e}(\mathbf{X}) = \mathrm{Enc}_{pm} \circ \mathfrak{H}_{\mathrm{e}}^{\mathcal{I}_e} \circ \mathrm{Enc}_{\mathrm{ma}}(\mathbf{X}),$$

where $\mathrm{Enc}_{ma}$ is a *marker-agnostic* encoder applied independently per input marker channel, $\mathfrak{H}_{e}^{\mathcal{I}_e}$ is a *marker-conditional hyperconvolution operator* instantiated from marker embeddings for the input set

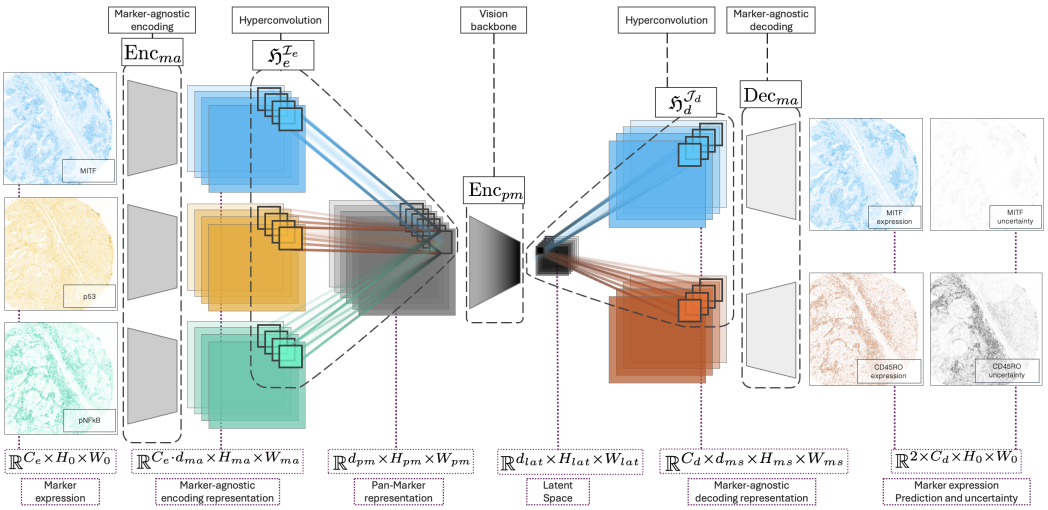



Figure 2: `ImmuVis` **architecture overview.** Marker-agnostic encoder stems embed each input marker channel and a hyperconvolution module, conditioned on learned marker embeddings, fuses them into a shared pan-marker representation processed by a standard vision backbone. A symmetric hyperconvolution and marker-agnostic decoder instantiates the requested output marker set, predicting per-marker reconstructions together with pixel-wise uncertainty (modified heteroscedastic log-variance).

$\mathcal{I}_e$, and $\mathrm{Enc}_{pm}$ maps the resulting features into a shared *pan-marker* latent space. Formal definitions are provided below.

**Marker-agnostic encoding** $\mathrm{Enc}_{ma}$**.** Let $\mathfrak{S} : \mathbb{R}^{1\times H_0\times W_0} \to \mathbb{R}^{d_{ma}\times H_{ma}\times W_{ma}}$ be a convolutional *stem* shared across markers, where $d_{ma}$ is the marker-agnostic feature width. $\mathfrak{S}$ preprocesses each marker-channel independently, downsampling the image while capturing local spatial expression patterns. Applying $\mathfrak{S}$ to each marker channel and concatenating along the channel axis yields $\mathrm{Enc}_{ma} : \mathbb{R}^{C_e\times H_0\times W_0} \to \mathbb{R}^{C_e\cdot d_{ma}\times H_{ma}\times W_{ma}}$ given as:

$$\mathbf{W} = \mathrm{Enc}_{ma}(\mathbf{X}) = \left[\, \mathfrak{S}(\mathbf{X}_{i^1}); \ldots ; \mathfrak{S}(\mathbf{X}_{i^{C_e}}) \,\right]_0$$

providing compressed, marker-agnostic representation $\mathbf{W}$ of $\mathbf{X}$. Here $[\,\cdot;\cdot\,]_{\mathrm{k}}$ denotes the concatenation along the k-th dimension.

**Encoder Hyperconvolution operator** $\mathfrak{H}_e$**.** Let $\phi_e : \mathcal{M} \to \mathbb{R}^{d_{\mathrm{pm}}\times d_{\mathrm{ma}}\times h_e\times w_e}$, be a learnable marker-conditional convolutional kernel generator, where $d_{\mathrm{pm}}$ is the channel width of the pan-marker space, and $(h_e, w_e)$ is the spatial kernel size. For $\mathcal{I}_e$, we formulate a hyperkernel $\mathbf{H}_e^{\mathcal{I}_e} \in \mathbb{R}^{d_{pm}\times C_e\cdot d_{ma}\times h_e\times w_e}$ as:

$$\mathbf{H}_e^{\mathcal{I}_e} = \left[\phi_e(i^1); \ldots ; \phi_e(i^{C_e})\right]_1.$$

Then $\mathfrak{H}_e^{\mathcal{I}_e} : \mathbb{R}^{C_e\cdot d_{ma}\times H_{ma}\times W_{ma}} \to \mathbb{R}^{d_{pm}\times H_{pm}\times W_{pm}}$ is a single dynamic convolution operator [29, 30]:

$$\mathbf{V} = \mathfrak{H}_e^{\mathcal{I}_e}(\mathbf{W}) = \mathbf{W} \otimes_{s,p} \mathbf{H}_e^{\mathcal{I}_e},$$

where $\otimes_{s,p}$ is a standard 2D convolution operation with stride $s$ and padding $p$, and $H_{pm}$ and $W_{pm}$ are the output spatial dimensions. Now $\mathbf{V}$ is an embedding of $\mathbf{X}$ into a unified pan-marker representation. Since $\mathfrak{H}_e^{\mathcal{I}_e}$ depends on $\mathcal{I}_e$, the induced hyperkernel adapts to the observed markers and enables capturing fine-grained local cross-marker dependencies. Equivalently, fusion sums marker-specific linear convolutions,

$$\mathbf{V} = \sum_{m\in\mathcal{I}_e} \mathfrak{S}(\mathbf{X}_m) \otimes_{s,p} \phi_e(m).$$

Jointly permuting marker identities and their channels leaves this sum unchanged. Fusion cost grows linearly with the number of input markers, without a marker-by-marker interaction term.

**Pan-marker backbone** $\mathrm{Enc}_{pm}$**.** Since $\mathbf{V}$ is a universal, fixed-sized embedding, it enables the application of standard computer vision backbones. Specifically, $\mathrm{Enc}_{pm} : \mathbb{R}^{d_{pm} \times H_{pm} \times W_{pm}} \to \mathbb{R}^{d_{lat} \times H_{lat} \times W_{lat}}$ is any fixed-channel vision backbone. This choice lets us leverage advances in general-purpose vision models while remaining flexible to different input marker sets, producing the pan-marker latent representation $\mathbf{Z} = \mathrm{Enc}_{pm}(\mathbf{V})$.

### 3.1.2 $\mathrm{Dec}^{\mathcal{J}_d}$ hypernetwork

Analogously to the encoding procedure, given target marker-set specific $\mathcal{J}_d$, the instantiated decoder $\mathrm{Dec}^{\mathcal{J}_d}$ is a composition: $\mathbf{X}^{\mu}, \mathbf{X}^{\log \sigma^2} = \mathrm{Dec}^{\mathcal{J}_d}(\mathbf{Z}) = \mathrm{Dec}_{\mathrm{ma}} \circ \mathfrak{H}_d^{\mathcal{J}_d}(\mathbf{Z})$, where $\mathfrak{H}_d^{\mathcal{J}_d}$ is a *marker-conditioned* hyperconvolution that maps the latent representation to *marker-specific* decoding features, and $\mathrm{Dec}_{\mathrm{ma}}$ produces per-marker mean predictions and log-variance for markers in $\mathcal{J}_d$. Formal definitions are provided below.

**Decoder Hyperconvolution operator** $\mathfrak{H}_d$**.** Let $\phi_d : \mathcal{M} \to \mathbb{R}^{d_{\mathrm{ms}} \times d_{\mathrm{lat}} \times h_d \times w_d}$, be a learnable marker-conditional convolutional kernel generator, where $d_{\mathrm{ms}}$ is the width of the marker set specific decoder space, and $(h_d, w_d)$ is the spatial kernel size. We define $\mathfrak{H}_d^{\mathcal{J}_d} : \mathbb{R}^{d_{lat} \times H_{lat} \times W_{lat}} \to \mathbb{R}^{C_d \times d_{ms} \times H_{ms} \times W_{ms}}$ as:

$$\mathbf{U} = \mathfrak{H}_d^{\mathcal{J}_d}(\mathbf{Z}) = \left[\mathbf{Z} \otimes_{s,p} \phi_d\left(j^1\right); \ldots; \mathbf{Z} \otimes_{s,p} \phi_d\left(j^{C_d}\right)\right]^0$$

where $\mathbf{U} = (\mathbf{U}_{j^1}, \ldots, \mathbf{U}_{j^{C_d}})$ constitutes an embedding of a latent space $\mathbf{Z}$ into a marker set specific space. Here $[\,\cdot;\cdot\,]^{\mathrm{k}}$ denotes stacking along the k-th dimension.

**Marker-agnostic decoding** $\mathrm{Dec}_{ma}$**.** Let $\mathfrak{R} : \mathbb{R}^{d_{ms} \times H_{ms} \times W_{ms}} \to \mathbb{R}^{2 \times H_0 \times W_0}$, be a convolutional *head* operator, that provides marker predictions and uncertainties from a marker-specific representations. $\mathrm{Dec}_{ma} : \mathbb{R}^{C_d \times d_{ms} \times H_{ms} \times W_{ms}} \to \mathbb{R}^{2 \times C_d \times H_0 \times W_0}$ applies $\mathfrak{R}$ to each marker channel independently and stacks the results:

$$\widehat{\mathbf{X}} = \mathrm{Dec}_{ma}(\mathbf{U}) = \left[\mathfrak{R}\left(\mathbf{U}_{j^1}\right); \ldots; \mathfrak{R}\left(\mathbf{U}_{j^{C_d}}\right)\right]^1.$$

Finally, we set $\mathbf{X}^{\mu} = \widehat{\mathbf{X}^0}$ and $\mathbf{X}^{\log \sigma^2} = \widehat{\mathbf{X}^1}$ to obtain the final point-wise prediction and its uncertainty approximation for markers from $\mathcal{J}_d$, respectively.

## 3.2 Masked Modelling Task

We train `ImmuVis` with a masked modelling objective [31]. Let $\mathbf{X}_{\mathcal{A}}$ denote an multiplexed image $\mathbf{X}$ restricted only to markers from marker set $\mathcal{A}$. During the training, for each image $\mathbf{X}_{\mathcal{I}_{\mathrm{img}}} \in \mathbb{R}^{C_{\mathrm{img}} \times H \times W}$ with a marker set $\mathcal{I}_{\mathrm{img}}$ we: (i) sample a target marker set $\mathcal{I}_{\mathrm{tgt}} \subseteq \mathcal{I}_{\mathrm{img}}$, (ii) sample an input marker set $\mathcal{I}_{\mathrm{in}} \subset \mathcal{I}_{\mathrm{tgt}}$, and (iii) apply patch-wise spatial masking on $\mathbf{X}_{\mathcal{I}_{\mathrm{in}}}$ to obtain a masked input $\tilde{\mathbf{X}}_{\mathcal{I}_{\mathrm{in}}}$. $\tilde{\mathbf{X}}_{\mathcal{I}_{\mathrm{in}}}$ is then fed to a model to obtain unmasked point-wise predictions and uncertainty estimations for all markers from set $\mathcal{I}_{\mathrm{tgt}}$:

$$(\mathbf{X}^{\mu}_{\mathcal{I}_{\mathrm{tgt}}}, \mathbf{X}^{\log \sigma^2}_{\mathcal{I}_{\mathrm{tgt}}}) = \texttt{ImmuVis}\big(\mathcal{I}_{\mathrm{in}}, \mathcal{I}_{\mathrm{tgt}}\big)\left(\tilde{\mathbf{X}}_{\mathcal{I}_{\mathrm{in}}}\right).$$

At test time, we set $\mathcal{I}_{\mathrm{tgt}} = \mathcal{I}_{\mathrm{img}}$. For detailed procedures of sampling and masking see Appendix A.1.

**Loss function** We train `ImmuVis` using a Gaussian heteroscedastic regression scheme [32]. It assumes that each element $\mathbf{X}_{c,h,w}$ of a target image $\mathbf{X}_{\mathcal{I}_{\mathrm{tgt}}}$ follows an independent Gaussian distribution:

$$\mathbf{X}_{c,h,w} \sim \mathcal{N}\Big(\mathbf{X}^{\mu}_{c,h,w}, \exp\Big(\mathbf{X}^{\log \sigma^2}_{c,h,w}\Big)\Big).$$

Then the training objective is, up to an additive constant, the modified Gaussian negative log-likelihood, $\beta$-NLL, following [33]:

$$\hat{\sigma}^2 = \exp\Big(\mathrm{clamp}(\mathbf{X}^{\log \sigma^2}_{\mathcal{I}_{\mathrm{tgt}}})\Big), \quad \epsilon = 10^{-8},$$

$$\mathcal{L}_{\beta\text{-NLL}} = \mathrm{mean}\left[\lfloor \hat{\sigma}^{2\beta} \rfloor \left(\frac{(\mathbf{X}_{\mathcal{I}_{\mathrm{tgt}}} - \mathbf{X}^{\mu}_{\mathcal{I}_{\mathrm{tgt}}})^2}{\hat{\sigma}^2 + \epsilon} + \log \hat{\sigma}^2\right)\right],$$

where the mean is taken over the batch, channels, and spatial dimensions, and $\lfloor \cdot \rfloor$ denotes the stop gradient operation. The variance $\hat{\sigma}^2$ is truncated using a gradient-preserving clamp to improve training stability. For implementation details see Appendix A.2.

# 4 Results

## 4.1 IMC17M dataset

To train the `ImmuVis` family, we curated IMC17M, the largest IMC corpus reported to date, composed of 17M patches from 28 datasets spanning 25 unique marker panels (24,405 images and 265 markers) across 14 histologies (see Appendix A.3 and Table 6). For training we split images into train/test in a 4:1 ratio, stratifying by dataset panel to preserve panel- and histopathology-level representation. All patches derived from the same parent image were assigned to the same split to prevent patch-level leakage.

## 4.2 `ImmuVis` model family

`ImmuVis` is a model family in which the choice of modules' backbone instantiates a specific architecture. We consider three variations: (i) a fully convolutional $\texttt{ImmuVis}_{\texttt{Conv}}$ based on ConvNeXt v2 [19], (ii) a Swin Transformer-based $\texttt{ImmuVis}_{\texttt{Swin}}$ [20], and (iii) a ViT-based $\texttt{ImmuVis}_{\texttt{ViT}}$ [21]. All variants use a pan-marker latent dimension $d_{\text{lat}} = 768$; architectural details are provided in Appendix A.4.

**Implementation of $\phi_e$ and $\phi_d$** We implement the convolutional kernel generators $\phi_e$ and $\phi_d$ as learnable marker-kernel lookup tables trained from scratch. This contrasts with VirTues [13] and EVA [14], which rely on marker encodings derived from protein language models [34, 35]. While this choice restricts inference to markers observed during training, as per marker sequence- or gene-level similarity do not imply similar tissue-level expression patterns it lets the model learn marker-specific convolutional operators directly from IMC measurements (for detailed explanation see Appendix A.5).

**Training procedure** All backbone weights and marker embeddings are randomly initialized and trained end-to-end with a single masked-reconstruction objective. Stem and objective ablations are described in Appendix A.4.1. We trained all models for 200 epochs with AdamW [36] optimizer using weight decay $10^{-4}$ and a cosine annealing learning-rate schedule with linear warm-up (for full details see Appendix A.6). For new panels with unseen markers, one can either skip markers not observed during the pretraining or train only their encoder and decoder embeddings with the remaining network frozen (Appendix A.10).

**Preprocessing Overview** For training `ImmuVis` model family, we preprocess raw IMC images with variance-stabilizing $\text{arcsinh}(x/5)$ transformation, frequency-based Butterworth [37] denoising, and intensity normalization, and then train/evaluate on fixed-size crops with standard spatial augmentations; full procedural details and all preprocessing parameters are given in Appendix A.7.

Because this pipeline differs from previously used IMC preprocessing workflows [13, 14], we include a dedicated signal-stability analysis in Appendix A.8, showing that our preprocessing excels at preserving meaningful marker variation while limiting technology-specific noise.

## 4.3 Experiments

We evaluate `ImmuVis` variants across: (i) virtual staining and uncertainty estimation, (ii) representation learning for cell typing and clinical prediction, and (iii) computational efficiency. We benchmark primarily against VirTues [13], which reports state-of-the-art performance for IMC data, and EVA [14], which further reports improvements over VirTues on selected tasks. We use the publicly released VirTues checkpoint from Hugging Face[1] and the EVA checkpoint released in the authors' GitHub repository. Our attempts to train KRONOS v1 [28] did not converge (Appendix A.9).

To disentangle the effects of training data and preprocessing, we train variants across different datasets and preprocessing pipelines. Model names containing `IMC17M` denote training on IMC17M. The

[1] `https://huggingface.co/bunnelab/virtues`, snapshot as of December 9, 2025.

suffix OP denotes our preprocessing, whereas VP denotes the VirTues preprocessing pipeline. For models without these suffixes, the original dataset and preprocessing of the corresponding method are used. We evaluate EVA trained on IMC17M using OP and the released checkpoint evaluated with our preprocessing (the original IMC preprocessing was not fully specified). We did not train ImmuVis on the VirTues dataset because some of their datasets were not publicly available (see Appendix A.3.1).

### 4.3.1 Virtual Staining

**Setup.** To evaluate virtual staining in an out-of-cohort setting, we use ten held-out test panels. Because ImmuVis and VirTues use different preprocessing pipelines, we train ImmuVis variants for out-of-cohort evaluation and VirTues on IMC17M with all test data excluded, using our preprocessing pipeline (OP). We use OP for the main comparison because VP requires statistics computed from the unseen test cohort, which violates the out-of-cohort setting (Appendix A.8.3). For completeness, we also train ImmuVis variants with VP preprocessing and report HN results in Appendix A.11. For fair comparison with VirTues, we exclude DNA1/DNA2, which are absent from the VirTues training panel as they lack appropriate ESM embeddings .

For each test image, we predict each of the $M$ target markers in a leave-one-marker-out setting. We report pixel-level MSE averaged across all images and markers, and Pearson correlation computed per marker across all images and then averaged over markers (for details see Appendix A.12).

Table 1: Virtual staining on ten test panels: MSE/Pearson. VirTues and EVA denote their IMC17M+OP variants with overlap-only inputs (full-panel comparison: Table 11). Bold marks the lowest MSE and highest Pearson per panel, including ties at the reported precision.

| Panel | EVA | $\texttt{ImmuVis}_{\texttt{ViT}}$ | $\texttt{ImmuVis}_{\texttt{Conv}}$ | $\texttt{ImmuVis}_{\texttt{Swin}}$ | VirTues |
|---|---|---|---|---|---|
| Head&Neck, Immucan [38] | .180/.51 | .081/.72 | .068/.77 | **.061/.78** | .101/.73 |
| Allam et al. [39], Panel 1 | .067/.50 | .045/.64 | .037/**.66** | .044/**.66** | **.035**/.65 |
| Allam et al. [39], Panel 2 | .462/.47 | .287/.56 | .217/.62 | .187/.59 | **.185/.65** |
| Hu et al. [40], Panel 1 | .108/.18 | .113/.32 | .092/.41 | .081/**.45** | **.056**/.34 |
| Hu et al. [40], Panel 2 | .487/.47 | .457/.48 | **.427/.61** | .531/.45 | .455/.56 |
| Hu et al. [40], Panel 3 | .021/.54 | .010/.58 | .017/.53 | **.005**/.55 | .011/**.65** |
| Moldoveanu [41], Panel 1 | .361/.39 | .143/.55 | **.137/.57** | .173/.53 | .201/.54 |
| Moldoveanu [41], Panel 2 | .302/.21 | .192/.35 | **.120/.43** | .180/.41 | .185/.29 |
| Schulz [42], Panel 1 | .470/.48 | .344/.66 | .475/.67 | **.282/.68** | .346/.54 |
| Schulz [42], Panel 2 | .341/.58 | .248/.63 | .267/**.69** | .215/.64 | **.167/.69** |

**Results.** As shown in Table 1, ImmuVis variants achieve the lowest MSE on six of the ten panels, with $\texttt{ImmuVis}_{\texttt{Conv}}$ and $\texttt{ImmuVis}_{\texttt{Swin}}$ leading on three panels each, while VirTues leads on the remaining four. ImmuVis variants also achieve the highest Pearson correlation on eight panels, including a tie with VirTues on Schulz Panel 2 at the reported precision, supporting the preservation of marker-specific expression patterns. These results suggest that convolutional and window-attention pan-marker encoders are well suited to the local spatial structure of IMC images. Paired Wilcoxon signed-rank tests on image-marker MSE with Benjamini-Hochberg FDR correction (significance at $q < 0.05$) yield wins/losses/non-significant differences versus VirTues of 6/1/3 for $\texttt{ImmuVis}_{\texttt{Swin}}$ and 3/3/4 for both $\texttt{ImmuVis}_{\texttt{Conv}}$ and $\texttt{ImmuVis}_{\texttt{ViT}}$ (Appendix A.10). Further per-marker HN results are provided in Appendix A.13.

### 4.3.2 Uncertainty estimation

**Setup.** To evaluate uncertainty fidelity, we use the heteroscedastic Gaussian head of the ImmuVis architecture, which encodes the predictive uncertainty through a log-variance estimate $\log \sigma^2$ for each reconstructed channel. We assess whether the predicted uncertainty is reliably *calibrated* in the sense of tracking reconstruction error.

We perform this analysis on all 40 markers of the IMMUcan Head & Neck (HN) cohort using the leave-one-marker-out setup and evaluate it using Pearson correlation between log variance and reconstruction error, expected calibration error (ECE) [43], and area under sparsification error (AUSE) [44, 45], which measures how closely the predicted $\sigma$ ranks pixels by error compared to an oracle ranking. For details see Appendix A.14.

We also assess whether ImmuVis-predicted uncertainty can identify unreliable cell-level readouts from virtual staining, using PD-1 positivity in CD8+ cells as an example. In a retrospective analysis of 21,058 segmented CD8+ cells from IMMUcan NSCLC2 panel 1, we predict PD-1 leave-one-out with $\texttt{ImmuVis}_{\texttt{Conv}}$ and compare predicted positivity with reference labels derived from measured PD-1 expression. For the details see Appendix A.15.

**Results.** Table 2 shows that all ImmuVis variants produce uncertainty estimates that closely track reconstruction error, with Pearson correlations above .92. $\texttt{ImmuVis}_{\texttt{Conv}}$ leads on Pearson (.941), while $\texttt{ImmuVis}_{\texttt{Swin}}$ achieves the lowest AUSE (.0112), indicating that its predicted $\sigma$ ranks pixels closest to the oracle ranking by true error. Although all variants obtain the same low ECE (.029), their expected calibration curves differ (see Figure 12). These results indicate that the heteroscedastic head provides a meaningful per-pixel confidence signal rather than an auxiliary output disconnected from reconstruction quality. For further uncertainty analyses see Appendix Figures 11-15. Cell-level uncertainty distinguishes incorrect from correct PD-1 positivity calls with an ROC-AUC of .78. Error rates are 8.4% and 77% in the lowest- and highest-uncertainty deciles, respectively, showing that classification errors are concentrated among cells assigned higher uncertainty.

Table 2: Uncertainty calibration evaluation on the HN cohort, measured by Pearson correlation (predicted uncertainty vs reconstruction error), expected calibration error (ECE), and area under sparsification error (AUSE, RMSE-base). **Bold** denotes the best score; ECE values are equal at the reported precision.

| Model | Pearson (↑) | ECE (↓) | AUSE (↓) |
|---|---|---|---|
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | .922 | .029 | .0135 |
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | **.941** | .029 | .0132 |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | .934 | .029 | **.0112** |

### 4.4 Representation Learning: Cell Typing

**Setup** To evaluate the quality of learned single-cell representations, we use three datasets: IMMUcan Phenotyping [46], with expert-curated manual cell-type annotations (kept as a hold-out validation set), and Danenberg et al. [47] and Cords et al. [48], that provide annotations based on unsupervised clustering. We benchmark ImmuVis variants against $\text{EVA}_{\texttt{OP}}$, the original VirTues model, and VirTues variants trained on IMC17M with both VP ($\text{VirTues}_{\texttt{IMC17M}}$) and OP ($\text{VirTues}_{\texttt{IMC17M+OP}}$) preprocessing pipelines. Additionally, we include a mean-marker baseline, representing each segmented cell by its mean expression for each measured marker under raw, OP, or VP preprocessing. This tests whether learned image representations improve over cell-averaged marker expression.

Single-cell patch extraction and embedding construction are described in Appendix A.16. As our extraction procedure differs from that of VirTues, we evaluate all VirTues variants under both protocols and report the better of the two scores. We train a multinomial logistic regression on single-cell embeddings and report per-cell-type F1 scores obtained via 10-fold cross-validation.

Table 3: Cell-typing Macro-F1 from 10-fold cross-validation. All standard deviations were below $10^{-2}$. For VirTues, values report the better score from either the original or ImmuVis cell-typing procedure (see Appendix A.17). Underline denotes the best score within a model family; **bold** denotes the best score per endpoint.

| Model | [46] | [47] | [48] | Avg. |
|---|---|---|---|---|
| $\text{mean-marker}_{\texttt{raw}}$ | .715 | .535 | .684 | .645 |
| $\text{mean-marker}_{\texttt{OP}}$ | .778 | .563 | .705 | .682 |
| $\text{mean-marker}_{\texttt{VP}}$ | .762 | .565 | .706 | .678 |
| $\text{EVA}_{\texttt{OP}}$ | .602 | .510 | .691 | .601 |
| $\text{EVA}_{\texttt{IMC17M+OP}}$ | .694 | .551 | .697 | .647 |
| VirTues | .726 | .569* | .711 | .669 |
| $\text{VirTues}_{\texttt{IMC17M}}$ | .703 | .555* | .712 | .657 |
| $\text{VirTues}_{\texttt{IMC17M+OP}}$ | .706 | .579* | .716 | .667 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | .753 | .592 | **.751** | .699 |
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | **.790** | **.592** | .745 | **.709** |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | .762 | .582 | .738 | .694 |

* Original VirTues cell-typing procedure used.

To assess out-of-cohort transfer, we additionally evaluate the frozen $\texttt{ImmuVis}_{\texttt{Conv}}$ encoder on the two Moldoveanu et al. [41] panels without new-marker adaptation, comparing it with released VirTues (Appendix A.10).

**Results** Table 3 shows that $\texttt{ImmuVis}_{\texttt{Conv}}$ achieves the best average Macro-F1 (.709), outperforming the strongest mean-marker baseline (.682), the strongest VirTues variant (.669), and retrained EVA (.647). For the mean-marker baseline, OP and VP increase average Macro-F1 from .645 on raw

inputs to .682 and .678, respectively. This shows that preprocessing improves the predictive value of cell-averaged marker expression even without a learned image encoder, providing a stronger reference for evaluating the pretrained representations. On Eling, mean-marker OP (.778) exceeds $\texttt{ImmuVis}_{\texttt{ViT}}$ (.753) and $\texttt{ImmuVis}_{\texttt{Swin}}$ (.762), while $\texttt{ImmuVis}_{\texttt{Conv}}$ reaches .790. All three ImmuVis variants exceed the mean-marker baselines on Danenberg, Cords, and on average. On Moldoveanu, $\texttt{ImmuVis}_{\texttt{Conv}}$ without new-marker adaptation achieves Macro-F1 scores of .765 on panel P1 and .643 on panel P2, compared with .734 and .617, respectively, for released VirTues, whose pretraining included this cohort (Table 10). These results support transfer of $\texttt{ImmuVis}_{\texttt{Conv}}$ representations to a cohort absent from pretraining, without adapting the encoder.

### 4.5 Representation Learning: Clinical predictions

**Setup** We further evaluate whether ImmuVis variants learn representations that transfer to patient-level clinical prediction. We consider two published IMC cohorts with clinical labels [47, 48]. For each cohort, we extract patch embeddings from the encoder, aggregate them to the patient level by mean pooling, and train logistic regression and ABMIL [49] for each clinical endpoint. We report Macro-F1 across 10-fold cross-validation splits as mean±standard deviation, and compare against VirTues variants, $\text{EVA}_{\texttt{OP}}$, and $\text{EVA}_{\texttt{IMC17M+OP}}$ under the same protocol. We also include mean-marker baselines using raw, OP, and VP preprocessing. Appendix A.18 details their construction and statistical testing, and reports ABMIL results.

**Results** Table 4 shows that ImmuVis variants provide competitive patient-level representations for clinical prediction. $\texttt{ImmuVis}_{\texttt{Swin}}$ achieves the best average Macro-F1 (.641), followed by $\texttt{ImmuVis}_{\texttt{Conv}}$ (.638), compared with .627 for the strongest VirTues variant and .608 for the strongest mean-marker baseline. Released and retrained EVA achieve .585 and .543, respectively. The VP mean-marker baseline achieves the highest displayed PAM50 mean (.43). The differences between ImmuVis and VirTues are not statistically significant on any of the seven endpoints (all $q \geq .34$). These results indicate that marker-adaptive pretraining yields representations that remain informative for clinical predictions under minimal downstream supervision. Appendix A.18 provides extended results and discusses the information limits of these endpoints.

Table 4: Representation learning for clinical endpoints (Macro-F1; higher is better). Values report mean±std Macro-F1 across 10-fold cross-validation; bold marks the highest displayed mean, including rounded ties.

| Model | Danenberg et al. [47] | | | | Cords et al. [48] | | | |
|---|---|---|---|---|---|---|---|---|
| | PAM50 | Grade | ER | ERBB2 | Subtype | Relapse | Grade | Avg. |
| $\text{mean-marker}_{\texttt{raw}}$ | .41±.06 | .47±.05 | .77±.08 | .75±.11 | .77±.03 | .56±.02 | .49±.05 | .602 |
| $\text{mean-marker}_{\texttt{OP}}$ | .42±.06 | .49±.06 | .76±.09 | .75±.11 | .78±.03 | .56±.02 | .49±.04 | .606 |
| $\text{mean-marker}_{\texttt{VP}}$ | **.43±.05** | .47±.05 | .75±.09 | .77±.10 | .79±.03 | .55±.03 | .49±.05 | .608 |
| $\text{EVA}_{\texttt{OP}}$ | .38±.05 | .50±.08 | .75±.07 | .61±.08 | .79±.03 | .57±.05 | .50±.05 | .585 |
| $\text{EVA}_{\texttt{IMC17M+OP}}$ | .34±.06 | .43±.05 | .68±.05 | .47±.002 | .79±.03 | .58±.04 | .51±.08 | .543 |
| VirTues | .37±.05 | **.53±.05** | .78±.08 | .76±.11 | .82±.03 | .57±.03 | .51±.07 | .620 |
| $\text{VirTues}_{\texttt{IMC17M}}$ | .36±.06 | .49±.05 | .77±.06 | .78±.13 | **.83±.02** | .58±.04 | .54±.06 | .621 |
| $\text{VirTues}_{\texttt{IMC17M+OP}}$ | .40±.08 | .49±.08 | .80±.06 | .77±.13 | **.83±.03** | .57±.05 | .53±.09 | .627 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | .40±.06 | .48±.05 | **.81±.04** | .77±.12 | .82±.03 | .58±.04 | .54±.06 | .628 |
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | .40±.09 | .52±.08 | **.81±.07** | **.80±.13** | .81±.03 | .59±.03 | .54±.05 | .638 |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | .42±.09 | .49±.06 | .79±.05 | **.80±.13** | **.83±.03** | **.61±.04** | **.55±.06** | **.641** |

### 4.6 Computational efficiency

**Setup.** We compare the inference-time computational cost of ImmuVis variants, VirTues, and $\text{EVA}_{\texttt{OP}}$ on a single forward pass for an IMC image of shape $[40, 128, 128]$. We report active parameter count, GFLOPs, mean±std inference time in milliseconds, and peak activation memory, all obtained on a single NVIDIA RTX 5000 Ada Generation GPU aggregated over 100 random input runs.

**Results.** Table 5 shows that ImmuVis variants operate in a substantially lower-compute time regime than prior foundation models for IMC. $\texttt{ImmuVis}_{\texttt{Conv}}$ requires 89.9 GFLOPs and 13.4±0.2 ms per forward pass, compared with 1049.3 GFLOPs and 28.52±0.5 ms for VirTues, and 3590 GFLOPs and 45.82±0.3 ms for $\text{EVA}_{\texttt{OP}}$. Thus, $\texttt{ImmuVis}_{\texttt{Conv}}$ reduces compute by more than an order of magnitude relative to VirTues and nearly 40-fold relative to EVA. These results transfer to being faster in wall-clock inference. Among ImmuVis variants, $\texttt{ImmuVis}_{\texttt{ViT}}$ is the fastest in runtime (11.1±0.3 ms) but uses the largest parameter count and activation memory within the family (105.2M parameters, 1.39 GB). $\texttt{ImmuVis}_{\texttt{Conv}}$ provides the most efficient overall trade-off, with the fewest parameters among ImmuVis variants (44.6M), the lowest GFLOPs, and lower activation memory than $\texttt{ImmuVis}_{\texttt{ViT}}$. $\texttt{ImmuVis}_{\texttt{Swin}}$ remains substantially more efficient than VirTues and EVA, while achieving strong virtual staining accuracy. Overall, these results combine strong cell typing and virtual staining with lower computational cost, while clinical prediction remains competitive with VirTues [13, 14].

Table 5: Computational efficiency. Inference cost measured on 100 random inputs of shape $[40, 128, 128]$ on a single NVIDIA RTX 5000 Ada Generation GPU.

| Model | Active param. | GFLOPs | Time [ms] | Mem. [GB] |
|---|---|---|---|---|
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | 44.6M | **89.9** | 13.4±0.2 | 1.01 |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | 57.0M | 138.8 | 18.2±0.9 | 1.03 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | 105.2M | 141.5 | **11.1±0.3** | 1.39 |
| VirTues | **42.5M** | 1049.3 | 28.5±0.5 | **0.45** |
| EVA | 122.4M | 3590.0 | 45.8±0.3 | 0.95 |

## 5 Discussion

ImmuVis addresses a central practical constraint of IMC, namely the absence of a fixed channel marker space, by introducing marker-adaptive hyperconvolutions that generate convolutional kernels from learned marker embeddings. Such a design enables a single foundation model to operate on any subset of markers seen during pretraining without re-training and architectural changes. Virtual staining results across ten held-out test panels demonstrate out-of-cohort transfer. Across the cell-typing benchmarks, all ImmuVis variants exceed the evaluated baselines on average, while clinical prediction remains competitive with VirTues. The locality analysis suggests that pan-marker representations enabled by the hyperconvolutional architecture preserve local spatial contexts and capture short-range dependencies relevant to tissue architecture and cellular interaction (see Appendix A.19). Beyond that, the heteroscedastic objective provides per-pixel uncertainty estimates that closely track reconstruction error, offering a practical reliability signal for downstream applications (e.g., prioritizing regions or markers requiring caution).

**Model choice.** $\texttt{ImmuVis}_{\texttt{Conv}}$ offers the strongest average cell-typing performance and lowest computational cost within the family. $\texttt{ImmuVis}_{\texttt{Swin}}$ is preferable when reconstruction fidelity and uncertainty ranking are priorities, while $\texttt{ImmuVis}_{\texttt{ViT}}$ offers the lowest inference latency at the cost of higher parameter count and activation memory.

**Limitations.** While the proposed design is computationally efficient and cohort-deployable, several limitations remain. First, markers absent from pretraining require learning new marker embeddings before they can be encoded or virtually stained. Second, the training distribution is constrained by marker availability and dataset bias; rare biomarkers and under-represented tissue contexts may yield less reliable reconstructions, especially under out-of-cohort transfer. Third, our likelihood model assumes conditional independence across pixels and channels, which is a pragmatic approximation but may under-represent structured noise or cross-channel dependencies in IMC. Finally, the clinical cohorts are small, and the observed differences from Virtues are not statistically significant.

**Future overlook.** ImmuVis could be extended to additional multiplex modalities (e.g., MIBI, CODEX), incorporate richer priors or structured uncertainty (e.g., spatially correlated noise), and develop more explainable representations characterizing cross-marker dependencies. These extensions may require modality-specific preprocessing, as the current pipeline is designed and challenged with IMC. If IMC adoption reaches clinical practice, ImmuVis could support diagnostic pipelines through virtual staining of missing markers and calibrated uncertainty flagging low-confidence predictions. Overall, our results suggest that operator-level channel adaptivity is an effective recipe for building practical foundation models in multiplex imaging, combining panel flexibility, efficiency, and reliability-aware predictions within a single architecture.

## Acknowledgments and Disclosure of Funding

**Funding.** This project was supported by the IMMUcan project that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 821558. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA (https://IMI.europa.eu). This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101136552 (SPACETIME). This work was supported by the Polish National Science Centre SONATA BIS grant No. 2020/38/E/NZ2/00305 and SONATA grant No. 2023/51/D/NZ2/02892.

**Computational resources.** We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grants no. PLG/2025/018195 and PLG/2026/019387.

**Conflicts of Interest.** Nothing to declare.

**Disclaimer.** Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

## References

[1] Charlotte Giesen, Hao A O Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, March 2014.

[2] Qing Chang, Olga I. Ornatsky, Iram Siddiqui, Alexander Loboda, Vladimir I. Baranov, and David W. Hedley. Imaging mass cytometry. *Cytometry Part A*, 91(2):160–169, February 2017.

[3] Vladan Milosevic. Different approaches to imaging mass cytometry data analysis. *Bioinformatics Advances*, 3(1):vbad046, 04 2023.

[4] Nat. Methods Editorial. Method of the year 2024: spatial proteomics. *Nature Methods*, 21(12):2195–2196, December 2024.

[5] Alina Bollhagen and Bernd Bodenmiller. Highly multiplexed tissue imaging in precision oncology and translational cancer research. *Cancer Discovery*, 14(11):2071–2088, November 2024.

[6] Natalie de Souza, Shan Zhao, and Bernd Bodenmiller. Multiplex protein imaging in tumour biology. *Nature Reviews Cancer*, 24(3):171–191, February 2024.

[7] Sean C. Bendall, Garry P. Nolan, Mario Roederer, and Pratip K. Chattopadhyay. A deep profiler's guide to cytometry. *Trends in Immunology*, 33(7):323–332, 2012.

[8] Stéphane Chevrier, Helena L. Crowell, Vito R. T. Zanotelli, Stefanie Engler, Mark D. Robinson, and Bernd Bodenmiller. Compensation of signal spillover in suspension and imaging mass cytometry. *Cell Systems*, 6(5):612–620.e5, 2018.

[9] Yuval Bussi and Leeat Keren. Multiplexed image analysis: what have we achieved and where are we headed? *Nature Methods*, 21(12):2212–2215, December 2024.

[10] Leena Latonen, Sonja Koivukoski, Umair Khan, and Pekka Ruusuvuori. Virtual staining for histology by deep learning. *Trends in Biotechnology*, 42(9):1177–1191, September 2024.

[11] Yu-Chen Lo, Timothy J. Keyes, Astraea Jager, Jolanda Sarno, Pablo Domizi, Ravindra Majeti, Kathleen M. Sakamoto, Norman Lacayo, Charles G. Mullighan, Jeffrey Waters, Bita Sahaf, Sean C. Bendall, and Kara L. Davis. Cytofin enables integrated analysis of public mass cytometry datasets using generalized anchors. *Nature Communications*, 13(1), February 2022.

[12] Muhammad Shaban, Wiem Lassoued, Kenneth Canubas, Shania Bailey, Yanling Liu, Clint Allen, Julius Strauss, James L Gulley, Sizun Jiang, Faisal Mahmood, George Zaki, and Houssein A Sater. Deep learning model imputes missing stains in multiplex images. *bioRxiv*, November 2023.

[13] Johann Wenckstern, Eeshaan Jain, Kiril Vasilev, Matteo Pariset, Andreas Wicki, Gabriele Gut, and Charlotte Bunne. Ai-powered virtual tissues from spatial proteomics for clinical diagnostics and biomedical discovery. *arXiv*, 2025.

[14] Yufan Liu, Rishabh Sharma, Matthew Bieniosek, Amy Kang, Eric Wu, Peter Chou, Irene Li, Maha Rahim, Erica Bauer, Ran Ji, Wei Duan, Li Qian, Ruibang Luo, Padmanee Sharma, Renu Dhanasekaran, Christian M. Schürch, Gregory Charville, Aaron T. Mayer, James Zou, Alexandro E. Trevino, and Zhenqin Wu. Modeling patient tissues at molecular resolution with eva. *openRxiv*, December 2025.

[15] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017.

[16] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. *arXiv*, 2019.

[17] James M. Dolezal, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Brittany Cody, Aaron S. Mansfield, Sagar Rakshit, Radhika Bansal, Melanie C. Bois, Aaron O. Bungum, Jefree J. Schulte, Everett E. Vokes, Marina Chiara Garassino, Aliya N. Husain, and Alexander T. Pearson. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature Communications*, 13(1), November 2022.

[18] Lorenzo Papa, Paolo Russo, Irene Amerini, and Luping Zhou. A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7682–7700, December 2024.

[19] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv*, 2021.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[22] Mohammed Zidane, Ahmad Makky, Matthias Bruhns, Alexander Rochwarger, Sepideh Babaei, Manfred Claassen, and Christian M. Schürch. A review on deep learning applications in highly multiplexed tissue imaging data analysis. *Frontiers in Bioinformatics*, 3, July 2023.

[23] Luke Ternes, Jia-Ren Lin, Yu-An Chen, Joe W. Gray, and Young Hwan Chang. Computational multiplex panel reduction to maximize information retention in breast cancer tissue microarrays. *PLOS Computational Biology*, 18(9):e1010505, September 2022.

[24] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11757–11768. IEEE, June 2024.

[25] Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie Chen, Kathleen Creel, Jared Q. Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang M. Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv*, 2021.

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[28] Muhammad Shaban, Yuzhou Chang, Huaying Qiu, Yao Yu Yeo, Andrew H. Song, Guillaume Jaume, Yuchen Wang, Luca L. Weishaupt, Tong Ding, Anurag Vaidya, Abdallah Lamane, Daniel Shao, Mohammed Zidane, Yunhao Bai, Paige McCallum, Shuli Luo, Wenrui Wu, Yang Wang, Precious Cramer, Chi Ngai Chan, Pierre Stephan, Johanna Schaffenrath, Jia Le Lee, Hendrik A. Michel, Caiwei Tian, Cristina Almagro-Perez, Sophia J. Wagner, Sharifa Sahai, Ming Y. Lu, Richard J. Chen, Andrew Zhang, Mark Edward M. Gonzales, Ahmad Makky, Jia-Ying Joey Lee, Hao Cheng, Nourhan El Ahmar, Sayed Matar, Maximilian Haist, Darci Phillips, Yuqi Tan, Garry P. Nolan, W. Richard Burack, Jacob D. Estes, Jonathan T. C. Liu, Toni K Choueiri, Neeraj Agarwal, Marc Barry, Scott J. Rodig, Long Phi Le, Georg Gerber, Christian M. Schürch, Fabian J. Theis, Youn H Kim, Joe Yeong, Sabina Signoretti, Brooke E. Howitt, Lit-Hsin Loo, Qin Ma, Sizun Jiang, and Faisal Mahmood. A foundation model for spatial proteomics. *arXiv*, 2025.

[29] Chao Li and Anbang Yao. KernelWarehouse: Rethinking the design of dynamic convolution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29201–29221. PMLR, 21–27 Jul 2024.

[30] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15979–15988. IEEE, June 2022.

[32] Andrew Stirn, Hans-Hermann Wessels, Megan D. Schertzer, Laura Peralta Pereira, Neville E. Sanjana, and David A. Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[33] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022.

[34] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

[35] Yiqun Chen and James Zou. Genept: A simple but effective foundation model for genes and cells built from chatgpt. *openRxiv*, October 2023.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[37] S. Butterworth. On the Theory of Filter Amplifiers. *Experimental Wireless & the Wireless Engineer*, 7:536–541, October 1930.

[38] IMMUcan Consortium. Immucan – an integrated european immuno-oncology profiling platform. `https://immucan.eu/`, 2026. Accessed: 2026-05-01.

[39] Mayar Allam, Thomas Hu, Jeongjin Lee, Jeffrey Aldrich, Sunil S. Badve, Yesim Gökmen-Polar, Manali Bhave, Suresh S. Ramalingam, Frank Schneider, and Ahmet F. Coskun. Spatially variant immune infiltration scoring in human cancer tissues. *npj Precision Oncology*, 6(1), 2022.

[40] Thomas Hu, Mayar Allam, Shuangyi Cai, Walter Henderson, Brian Yueh, Aybuke Garipcan, Anton V. Ievlev, Maryam Afkarian, Semir Beyaz, and Ahmet F. Coskun. Single-cell spatial metabolomics with cell-type specific protein profiling for tissue systems biology. *Nature Communications*, 14(1), December 2023.

[41] Dan Moldoveanu, LeeAnn Ramsay, Mathieu Lajoie, Luke Anderson-Trocme, Marine Lingrand, Diana Berry, Lucas J.M. Perus, Yuhong Wei, Cleber Moraes, Rached Alkallas, Shivshankari Rajkumar, Dongmei Zuo, Matthew Dankner, Eric Hongbo Xu, Nicholas R. Bertos, Hamed S. Najafabadi, Simon Gravel, Santiago Costantino, Martin J. Richer, Amanda W. Lund, Sonia V. Del Rincon, Alan Spatz, Wilson H. Miller, Rahima Jamal, Réjean Lapointe, Anne-Marie Mes-Masson, Simon Turcotte, Kevin Petrecca, Sinziana Dumitra, Ari-Nareg Meguerditchian, Keith Richardson, Francine Tremblay, Beatrice Wang, May Chergui, Marie-Christine Guiot, Kevin Watters, John Stagg, Daniela F. Quail, Catalin Mihalcioiu, Sarkis Meterissian, and Ian R. Watson. Spatially mapping the immune landscape of melanoma using imaging mass cytometry. *Science Immunology*, 7(70), April 2022.

[42] Daniel Schulz, Vito Riccardo Tomaso Zanotelli, Jana Raja Fischer, Denis Schapiro, Stefanie Engler, Xiao-Kang Lun, Hartland Warren Jackson, and Bernd Bodenmiller. Simultaneous multiplexed imaging of mrna and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Systems*, 6(4):531, April 2018.

[43] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR, 10–15 Jul 2018.

[44] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[45] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[46] Nils Eling, Julien Dorier, Sylvie Rusakiewicz, Robin Liechti, Preethi Devanand, Michelle Daniel, Jonas Windhager, Bruno Palau Fernandez, Sophie Déglise, Lucie Despland, Abdelkader Benyagoub, Marcin Możejko, Dawid Uchal, Ewa Szczurek, Alexander Loboda, Daaf Sandkuijl, Nikesh Parsotam, Henoch S. Hong, Marie Morfouace, Nicolas Guex, George Coukos, Bernd

Bodenmiller, Stephanie Tissot, and Daniel Schulz. Multi-modal image analysis for large-scale cancer tissue studies within immucan. *Cell Reports Methods*, 5(9):101170, September 2025.

[47] Esther Danenberg, Helen Bardwell, Vito R. T. Zanotelli, Elena Provenzano, Suet-Feung Chin, Oscar M. Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O. Ellis, Bernd Bodenmiller, Carlos Caldas, and H. Raza Ali. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature Genetics*, 54(5):660–669, April 2022.

[48] Lena Cords, Stefanie Engler, Martina Haberecker, Jan Hendrik Rüschoff, Holger Moch, Natalie de Souza, and Bernd Bodenmiller. Cancer-associated fibroblast phenotypes are associated with patient outcome in non-small cell lung cancer. *Cancer Cell*, 42(3):396–412.e5, March 2024.

[49] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018.

[50] Shoaib Ajaib, Joshua Winter-Luke, Richard J Digby, Steven Pollock, Gemma Hemmings, Arief Gusnanto, Aruna Chakrabarty, Azzam Ismail, Erica Wilson, Bethany Hunter, Andrew Filby, David McDonald, Asa A Brockman, Rebecca A Ihrie, and Lucy F Stead. Spatial profiling of longitudinal glioblastoma reveals consistent changes in cellular architecture, post-treatment. *Neuro-Oncology*, September 2025.

[51] Henrike Salié, Lara Wischer, Antonio D'Alessio, Ira Godbole, Yuan Suo, Patricia Otto-Mora, Juergen Beck, Olaf Neumann, Albrecht Stenzinger, Peter Schirmacher, Claudia A M Fulgenzi, Andreas Blaumeiser, Melanie Boerries, Natascha Roehlen, Michael Schultheiß, Maike Hofmann, Robert Thimme, David J Pinato, Thomas Longerich, and Bertram Bengsch. Spatial single-cell profiling and neighbourhood analysis reveal the determinants of immune architecture connected to checkpoint inhibitor therapy outcome in hepatocellular carcinoma. *Gut*, 74(3):451–466, September 2024.

[52] Charles M. Bowen, Nan Deng, Laura Reyes-Uribe, Edwin Roger Parra, Pedro Rocha, Luisa M. Solis, Ignacio I. Wistuba, Valerie O. Sepeda, Lana Vornik, Marjorie Perloff, Eva Szabo, Asad Umar, Krishna M. Sinha, Powel H. Brown, and Eduardo Vilar. Naproxen chemoprevention induces proliferation of cytotoxic lymphocytes in lynch syndrome colorectal mucosa. *Frontiers in Immunology*, 14, May 2023.

[53] Yeonju Cho, Jae W. Lee, Sarah M. Shin, Alexei G. Hernandez, Xuan Yuan, Jowaly Schneider, Jody E. Hooper, Laura D. Wood, Elizabeth M. Jaffee, Atul Deshpande, and Won Jin Ho. Modeling cellular influence delineates functionally relevant cellular neighborhoods in primary and metastatic pancreatic ductal adenocarcinoma. *BiorXiv*, June 2025.

[54] Lena Cords, Sandra Tietscher, Tobias Anzeneder, Claus Langwieder, Martin Rees, Natalie de Souza, and Bernd Bodenmiller. Cancer-associated fibroblast classification in single-cell and spatial proteomics data. *Nature Communications*, 14(1), July 2023.

[55] Nicolas Damond, Stefanie Engler, Vito R.T. Zanotelli, Denis Schapiro, Clive H. Wasserfall, Irina Kusmartseva, Harry S. Nick, Fabrizio Thorel, Pedro L. Herrera, Mark A. Atkinson, and Bernd Bodenmiller. A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metabolism*, 29(3):755–768.e5, March 2019.

[56] Benjamin Ehret. Imc data associated with the publication 'shaking organoid co-cultures: human-relevant models of epithelial, stromal and immune interactions for cancer research and immunotherapy development', 2025.

[57] Jakob Einhaus, Dyani K. Gaudilliere, Julien Hedou, Dorien Feyaerts, Michael G. Ozawa, Masaki Sato, Edward A. Ganio, Amy S. Tsai, Ina A. Stelzer, Karl C. Bruckman, Jonas N. Amar, Maximilian Sabayev, Thomas A. Bonham, Joshua Gillard, Maïgane Diop, Amelie Cambriel, Zala N. Mihalic, Tulio Valdez, Stanley Y. Liu, Leticia Feirrera, David K. Lam, John B. Sunwoo, Christian M. Schürch, Brice Gaudilliere, and Xiaoyuan Han. Spatial subsetting enables integrative modeling of oral squamous cell carcinoma multiplex imaging data. *iScience*, 26(12):108486, December 2023.

[58] Michael J. Haley, Leoma Bere, James Minshull, Sokratia Georgaka, Natalia Garcia-Martin, Gareth Howell, David J. Coope, Federico Roncaroli, Andrew King, David C. Wedge, Stuart M. Allan, Omar N. Pathmanaban, David Brough, and Kevin N. Couper. Hypoxia coordinates the spatial landscape of myeloid cells within glioblastoma to affect survival. *Science Advances*, 10(20), May 2024.

[59] Tobias Hoch, Daniel Schulz, Nils Eling, Julia Martínez Gómez, Mitchell P. Levesque, and Bernd Bodenmiller. Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy. *Science Immunology*, 7(70), April 2022.

[60] Hartland W. Jackson, Jana R. Fischer, Vito R. T. Zanotelli, H. Raza Ali, Robert Mechera, Savas D. Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P. Weber, and Bernd Bodenmiller. The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620, January 2020.

[61] Emre Küçükköse, Matthijs J. D. Baars, Mojtaba Amini, Suzanna J. Schraa, Evelien Floor, Guus M. Bol, Inne H. M. Borel Rinkes, Jeanine M. L. Roodhart, Miriam Koopman, Jamila Laoukili, Onno Kranenburg, and Yvonne Vercoulen. Stromal localization of inactive cd8+ t cells in metastatic mismatch repair deficient colorectal cancer. *British Journal of Cancer*, 130(2):213–223, December 2023.

[62] Lasse Meyer, Hartland W. Jackson, Nils Eling, Shan Zhao, Genki Usui, Haithem Dakhli, Peter Schraml, Susanne Dettwiler, Constanze Elfgen, Zsuzsanna Varga, Holger Moch, Natalie de Souza, and Bernd Bodenmiller. A stratification system for breast cancer based on basoluminal tumor cells and spatial tumor architecture. *Cancer Cell*, 43(9):1637–1655.e9, September 2025.

[63] Kentaro Ohara, André Figueiredo Rendeiro, Bhavneet Bhinder, Kenneth Wha Eng, Hiranmayi Ravichandran, Duy Nguyen, David Pisapia, Aram Vosoughi, Evan Fernandez, Kyrillus S. Shohdy, Jyothi Manohar, Shaham Beg, David Wilkes, Brian D. Robinson, Francesca Khani, Rohan Bareja, Scott T. Tagawa, Madhu M. Ouseph, Andrea Sboner, Olivier Elemento, Bishoy M. Faltas, and Juan Miguel Mosquera. The evolution of metastatic upper tract urothelial carcinoma through genomic-transcriptomic and single-cell protein markers analysis. *Nature Communications*, 15(1), March 2024.

[64] Alessandra Rigamonti, Marika Viatore, Rebecca Polidori, Daoud Rahal, Marco Erreni, Maria Rita Fumagalli, Damiano Zanini, Andrea Doni, Anna Rita Putignano, Paola Bossi, Emanuele Voulaz, Marco Alloisio, Sabrina Rossi, Paolo Andrea Zucali, Armando Santoro, Vittoria Balzano, Paola Nisticò, Friedrich Feuerhake, Alberto Mantovani, Massimo Locati, and Federica Marchesi. Integrating ai-powered digital pathology and imaging mass cytometry identifies key classifiers of tumor cells, stroma, and immune cells in non–small cell lung cancer. *Cancer Research*, 84(7):1165–1177, February 2024.

[65] Nathan Steenbuck, Nicolas Damond, Stefanie Engler, Irina Kusmartseva, Amanda L. Posgai, Denise M. Drotar, MacKenzie D. Williams, Natalie de Souza, Todd M. Brusko, Maigan A. Brusko, Clive H. Wasserfall, Mark A. Atkinson, and Bernd Bodenmiller. Imaging mass cytometry reveals early $\beta$-cell dysfunction and changes in immune signatures during type 1 diabetes progression in human pancreata. *BiorXiv*, March 2025.

[66] Jonathan H. Sussman, Nathalia Kim, Samantha B. Kemp, Daniel Traum, Takeshi Katsuda, Benjamin M. Kahn, Jason Xu, Il-Kyu Kim, Cody Eskandarian, Devora Delman, Gregory L. Beatty, Klaus H. Kaestner, Amber L. Simpson, and Ben Z. Stanger. Multiplexed imaging mass cytometry analysis characterizes the vascular niche in pancreatic cancer. *Cancer Research*, 84(14):2364–2376, May 2024.

[67] Xu Xiao, Qian Guo, Chuanliang Cui, Yating Lin, Lei Zhang, Xin Ding, Qiyuan Li, Minshu Wang, Wenxian Yang, Yan Kong, and Rongshan Yu. Multiplexed imaging mass cytometry reveals distinct tumor-immune microenvironments linked to immunotherapy responses in melanoma. *Communications Medicine*, 2(1), October 2022.

[68] Xiao Qian Wang, Esther Danenberg, Chiun-Sheng Huang, Daniel Egle, Maurizio Callari, Begoña Bermejo, Matteo Dugo, Claudio Zamagni, Marc Thill, Anton Anton, Stefania Zambelli,

Stefania Russo, Eva Maria Ciruelos, Richard Greil, Balázs Győrffy, Vladimir Semiglazov, Marco Colleoni, Catherine M. Kelly, Gabriella Mariani, Lucia Del Mastro, Olivia Biasi, Robert S. Seitz, Pinuccia Valagussa, Giuseppe Viale, Luca Gianni, Giampaolo Bianchini, and H. Raza Ali. Spatial predictors of immunotherapy response in triple-negative breast cancer. *Nature*, 621(7980):868–876, 2023.

[69] Bo Zhu, Pingjun Chen, Muhammad Aminu, Jian-Rong Li, Junya Fujimoto, Yanhua Tian, Lingzhi Hong, Hong Chen, Xin Hu, Chenyang Li, Natalie Vokes, Andre L. Moreira, Don L. Gibbons, Luisa M. Solis Soto, Edwin Roger Parra Cuentas, Ou Shi, Songhui Diao, Jie Ye, Frank R. Rojas, Eduardo Vilar, Anirban Maitra, Ken Chen, Nicolas Navin, Monique Nilsson, Beibei Huang, Simon Heeke, Jianhua Zhang, Cara L. Haymaker, Vamsidhar Velcheti, Daniel H. Sterman, Veena Kochat, William I. Padron, Ludmil B. Alexandrov, Zhubo Wei, Xiuning Le, Linghua Wang, Junya Fukuoka, J. Jack Lee, Ignacio I. Wistuba, Harvey I. Pass, Mark Davis, Samir Hanash, Chao Cheng, Steven Dubinett, Avrum Spira, Kunal Rai, Scott M. Lippman, P. Andrew Futreal, John V. Heymach, Alexandre Reuben, Jia Wu, and Jianjun Zhang. Spatial and multiomics analysis of human and mouse lung adenocarcinoma precursors reveals tim-3 as a putative target for precancer interception. *Cancer Cell*, 43(6):1125–1140.e10, 2025.

[70] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014.

[71] Alina Bollhagen, James Whipman, Ricardo Coelho, Viola Heinzelmann-Schwarz, Francis Jacob, and Bernd Bodenmiller. High-resolution imaging mass cytometry to map subcellular structures. *Nature Methods*, 22(12):2601–2608, October 2025.

[72] Jonas Windhager, Vito Riccardo Tomaso Zanotelli, Daniel Schulz, Lasse Meyer, Michelle Daniel, Bernd Bodenmiller, and Nils Eling. An end-to-end workflow for multiplexed image processing and analysis. *Nature Protocols*, 18(11):3565–3613, October 2023.

[73] Tsuyoshi Hosogane, Ruben Casanova, and Bernd Bodenmiller. Dna-barcoded signal amplification for imaging mass cytometry enables sensitive and highly multiplexed tissue imaging. *Nature Methods*, 20(9):1304–1309, August 2023.

[74] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El-ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Pe'er, Scott D. Tanner, and Garry P. Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, May 2011.

[75] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, March 2007.

# A Appendix

## A.1 Sampling and masking procedures for Masked Modelling task

Each training sample consists of a full measured panel $(\mathbf{X}_{\mathcal{I}_{\text{img}}}, \mathcal{I}_{\text{img}})$ with $\mathbf{X}_{\mathcal{I}_{\text{img}}} \in \mathbb{R}^{C_{\text{im}} \times H \times W}$ where $C_{\text{img}} = |\mathcal{I}_{\text{img}}|$. We construct minibatches using a panel-grouped sampler, so all samples within a minibatch share $C_{\text{img}}$ (the same panel size) using the following steps: (i) target-set subsampling, (ii) full-channel dropout within the targets, and (iii) patch-wise spatial masking of the remaining inputs.

**(i) Target-set subsampling.** For each minibatch we first sample a target size

$$K \sim \text{Unif}\Big(\{\lceil \alpha C_{\text{img}} \rceil, \dots, C_{\text{img}}\}\Big),$$

where $\alpha$ by default is set to 0.75 and, for each sample, we choose $K$ markers uniformly without replacement from $\mathcal{I}_{\text{img}}$ to form $\mathcal{I}_{\text{tgt}}$ and the corresponding tensor $\mathbf{X}_{\mathcal{I}_{\text{tgt}}} \in \mathbb{R}^{K \times H \times W}$. Each sample in minibatch has its $K$ markers drawn independently.

**(ii) Full-marker dropout within targets.** Given $K$, we drop a non-empty subset of target markers by sampling

$$M \sim \text{Unif}\Big(\{1, \dots, \lceil \beta K \rceil\}\Big),$$

where $\beta$ by default is set to 0.5, and selecting $K - M$ markers uniformly without replacement from $\mathcal{I}_{\text{tgt}}$ to form the encoder input set $\mathcal{I}_{\text{in}} \subset \mathcal{I}_{\text{tgt}}$. The dropped channels $\mathcal{I}_{\text{tgt}} \setminus \mathcal{I}_{\text{in}}$ are completely unobserved by the encoder, but are included in the reconstruction targets. Each sample in minibatch has its $K - M$ channels drawn independently. This setting simulates a virtual staining scenario, where models aims at prediction of the markers missing from panel.

**(iii) Patch-wise spatial masking.** On the remaining input channels $\mathcal{I}_{\text{in}}$, we follow a masking protocol from [19] and additionally mask spatial patches by setting them to zero. Let $p = 8$ and assume $p$ divides $H$ and $W$. Define $(h, w) = (H/p, W/p)$ and sample a Bernoulli mask on the patch grid

$$\mathbf{m} \in \{0,1\}^{|\mathcal{I}_{\text{in}}| \times h \times w}, \qquad \mathbb{P}[\mathbf{m}_{c,u,v} = 1] = \rho,$$

where $\rho$ by default is set to 0.6. We then expand each entry of $\mathbf{m}$ to a $p \times p$ block to obtain $\mathbf{M} \in \{0,1\}^{|\mathcal{I}_{\text{in}}| \times H \times W}$ and define the masked encoder input

$$\tilde{\mathbf{X}}_{\mathcal{I}_{\text{in}}} = \mathbf{X}_{\mathcal{I}_{\text{in}}} \odot (1 - \mathbf{M}),$$

where $\odot$ denotes element wise multiplication.

## A.2 Stabilizing log-variance training using a gradient clamping

Let $\ell = \mathbf{X}_{\mathcal{I}_{\text{tgt}}}^{\log \sigma^2}$ denote the predicted log-variance. For numerical stability, we use the clamped log-variance

$$\hat{\ell} = \text{clamp}(\ell; a, b), \qquad (a, b) = (-15, 15).$$

In the *forward pass*, clamp returns the hard clip $\text{clip}(\ell; a, b) = \min(\max(\ell, a), b)$. In the *backward pass*, the gradient is propagated unchanged inside the interval and is smoothly downweighted outside it:

$$\frac{\partial \,\text{clamp}(\ell; a, b)}{\partial \ell} = \begin{cases} 1, & a < \ell < b, \\ 1 - \tanh^2(\ell), & \ell \le a \text{ or } \ell \ge b. \end{cases}$$

For the $\beta$-NLL loss, we set $\beta = 0.5$, following its original authors suggestion [33].

## A.3 Dataset details

IMC17M is a curated aggregation of 28 IMC datasets comprising 25 distinct marker panels, totalling 24,405 images and 265 unique markers across 14 histologies. A detailed composition of used datasets is presented in Table 6. Throughout, we use *dataset* to denote a source study/cohort, and *panel* to

Table 6: **Cohort-level composition of IMC17M** . The dataset spans 28 cohorts, 24,405 images, 265 markers, and over 17M patches. IMMUcan BC2/NSCLC2 (P1/2)/SCCHN1 datasets are IMMUcan [38] proprietary data, awaiting publication. The *VirTues* column denotes with ✓cohorts included in the VirTues pretraining corpus.

| Dataset | Subjects | Images | Markers | Histopathology | VirTues |
|---|---|---|---|---|---|
| ajaib [50] | 5 | 34 | 34 | Glioblastoma | |
| bengsch [51] | 54 | 108 | 42 | Hepatocellular carcinoma | |
| bowen-lynch [52] | 18 | 85 | 38 | Lynch Syndrome colorectal mucosa | |
| cho [53] | 12 | 140 | 41 | Pancreatic adenocarcinoma | |
| cords [48] | 1070 | 2070 | 42 | Non-small cell lung cancer | ✓ |
| cords-fibro [54] | 12 | 110 | 42 | Breast cancer | ✓ |
| damond [55] | 12 | 845 | 36 | Type I diabetes pancreas | ✓ |
| danenberg [47] | 693 | 794 | 39 | Breast cancer | ✓ |
| ehret-p1 [56] | 52 | 56 | 42 | Organoid | |
| ehret-p2 [56] | 52 | 52 | 44 | Organoid | |
| einhaus [57] | 24 | 94 | 41 | Oral squamous cell carcinoma | |
| haley-glio [58] | 8 | 24 | 40 | Glioblastoma | |
| hoch-protein [59] | 69 | 167 | 46 | Melanoma | ✓ |
| hoch-rna [59] | 69 | 166 | 41 | Melanoma | ✓ |
| IMMUcan BC2 [38]† | 223 | 813 | 40 | Breast cancer | |
| IMMUcan NSCLC2 (P1) [38]† | 192 | 745 | 40 | Non-small-cell lung cancer | |
| IMMUcan NSCLC2 (P2) [38]† | 190 | 566 | 44 | Non-small-cell lung cancer | |
| IMMUcan SCCHN1 [38]‡,† | 218 | 743 | 40 | Head and neck cancer | |
| IMMUcan Phenotyping [46]‡ | | | 40 | pan cancer | ✓ |
| jackson-basel [60] | 285 | 381 | 39 | Breast cancer | ✓ |
| jackson-zurich [60] | 72 | 365 | 39 | Breast cancer | ✓ |
| kucukkose [61] | 9 | 27 | 34 | Colorectal cancer | |
| meyer [62] | 215 | 495 | 39 | Breast cancer | |
| ohara [63] | 7 | 58 | 28 | Urothelial carcinoma | |
| rigamonti [64] | 108 | 158 | 29 | Non-small-cell lung cancer | ✓ |
| steenbuck-immune [65] | 88 | 7557 | 47 | Type I diabetes pancreas | |
| steenbuck-islets [65] | 88 | 7558 | 47 | Type I diabetes pancreas | |
| sussman [66] | 9 | 35 | 35 | Pancreatic adenocarcinoma | |
| xu [67] | 26 | 158 | 37 | Melanoma | |
| **TOTAL** | **3880** | **24405** | | | |

† IMMUcan BC2, NSCLC2 P1, NSCLC2 P2, and SCCHN1 are IMMUcan consortium proprietary datasets awaiting publication.
‡ IMMUcan SCCHN1 and Phenotyping were held out from training for all `ImmuVis` variants and VirTues and used exclusively for out-of-cohort evaluation.

denote its measured marker configuration; multiple datasets may share a panel, and a single study may provide multiple panels (e.g., protocol variants).

The corpus spans diverse biological contexts, including tumour microenvironments (e.g., breast, lung, glioblastoma, colorectal, melanoma, oral squamous cell carcinoma, urothelial carcinoma), immune system characterisation, and tissue-specific studies (e.g., pancreas in Type I diabetes and organoids). Panel sizes range from 28 to 47 markers, with most panels in the 35-45 range, reflecting the strong marker-set heterogeneity typical of IMC collections.

For all experiments, we partition images into train/test with a 4:1 ratio, stratified by panel to ensure that each panel contributes proportionally to both splits and to reduce confounding by panel-specific acquisition characteristics. Table 6 reports per-dataset counts and histopathologies; datasets labelled IMMUcan (except IMMUcan Phenotyping [46]) are consortium data and are currently proprietary (awaiting publication).

### A.3.1 Notes on VirTues overlap.

VirTues was pretrained on 14 datasets [13]. Eight of these overlap with IMC17M , corresponding to ten IMC17M cohorts after cohort-level splitting: cords, cords-fibro, damond, danenberg, hoch-protein, hoch-rna, IMMUcan Phenotyping, jackson-basel, jackson-zurich, and rigamonti. The remaining six VirTues pretraining datasets (Allam et al. [39], Hu et al. [40], Moldoveanu et al. [41], Schultz et al. [42], Wang et al. [68], and Zhu et al. [69]) are not part of IMC17M . In particular, Zhu et al. [69] dataset is behind DTA, while Wang et al. [68] under protected zenodo access. Meyer et al. [62] was used by VirTues only for downstream transfer evaluation, not for pretraining.

### A.4 ImmuVis Models configurations

In this section we introduce the detailed architecture of $\texttt{ImmuVis}_{\texttt{Conv}}$, $\texttt{ImmuVis}_{\texttt{Swin}}$, and $\texttt{ImmuVis}_{\texttt{ViT}}$ models. We considered the following configurations for the masked modelling task:

- Marker-Agnostic stages $L_{ma} \in \{\texttt{None}, 1, 2\}$,
- Hyperkernel size $(h_e \times w_e) \in \{(1 \times 1), (3 \times 3)\}$
- Stage dimensionalities for CONV and SWIN variants $\in \{(192, 384, 768), (128, 256, 512)\}$,
- Hidden size for VIT variant $\in \{512, 768\}$,

The details of all selected architectures configurations are shown in Table 7.

#### A.4.1 Stem and objective ablations

Table 8 varies the marker-agnostic stem depth of $\texttt{ImmuVis}_{\texttt{Conv}}$ at $d_{lat} = 768$. With no stem ($L_{ma} = 0$), the hyperkernel acts directly on pixels and is equivalent to a fixed-channel convolution over the vocabulary with missing channels zero-filled. The one-stage stem improves both downstream scores over this control; a second stage lowers validation NLL without improving those scores. These results do not establish validation NLL alone as a model-selection rule. The matched objective comparison in Table 9 changes only the training objective. Plain MSE slightly improves Conv staining, whereas $\beta$-NLL improves its cell typing and clinical scores and provides the uncertainty output.

#### A.4.2 Marker-agnostic decoder architecture

Let $\mathbf{U}_{j^c}$ be the output of the decoder Hyperconvolution operator $\mathfrak{H}_d$ for target marker $j^c$. Then $\mathfrak{R}$ of the marker-agnostic decoder $\mathrm{Dec}_{\mathrm{ma}}$ performs the following steps:

$$\mathbf{D}_{j^c} = \text{Blocks}^{(B_{dec})}(\mathbf{U}_{j^c})$$
$$\mathbf{P}_{j^c} = \text{Conv}_{1\times1}(\mathbf{D}_{j^c}, 2\lambda^2)$$
$$\hat{\mathbf{X}}_{j^c} = \text{PixelShuffle}(\mathbf{P}_{j^c})$$

where $\lambda = 8$ is the total upsampling factor. Since in the preprocessing stage of IMC17M we normalize marker intensities to a $[0, 1]$ interval, we map

$$\hat{\mathbf{X}}^0 = \mathrm{sigmoid}(\hat{\mathbf{X}}^0).$$

#### A.4.3 Decoder hyperkernel size

All three ImmuVis variants use $1 \times 1$ hyperkernels in the decoder ($h_d \times w_d = 1 \times 1$). We evaluated larger spatial decoder kernels (e.g., $3 \times 3$) but these configurations led to consistent training instability and divergence (larger decoder kernels substantially increased parameter count). No formal ablation is reported as no stable larger-kernel decoder was obtained.

#### A.4.4 Comment on the $\texttt{ImmuVis}_{\texttt{ViT}}$ architecture

For $\texttt{ImmuVis}_{\texttt{ViT}}$ we initially evaluated a configuration analogous to ImmuVis that employed a one-stage Marker-agnostic encoder, however training of this configuration diverged. Subsequently, we investigated a pixel-wise hyperkernel embedding strategy equivalent to the $L_{ma} = 0$ configuration, followed by a standard patch tokenizer, but it also suffered from the convergence difficulties. Consequently, we adopted the minimal strategy with the hyperkernel acting directly as a multiplex tokenizer ($\mathrm{Enc}_{\mathrm{ma}}$ was set as an identity), creating pan-marker patch embeddings from the original input.

### A.5 Marker-conditional kernel generators

For both encoder and decoder generators $\phi_e, \phi_d$ we first map markers to their vocabulary indices $i \in \mathcal{N}$. Then, each marker index is assigned to a standard learnable embedding of a flattened intended kernel shape (either $d_{\mathrm{pm}} \cdot d_{\mathrm{ma}} \cdot h_e \cdot w_e$ for encoder and $d_{\mathrm{ms}} \cdot d_{\mathrm{lat}} \cdot h_d \cdot w_d$ for decoder), subsequently reshaped to the final kernel dimensions.

Table 7: `ImmuVis` models architecture and parameter specifications

| Component | Parameter | Value / Specification |
|---|---|---|
| **`ImmuVis`$_{\texttt{Conv}}$** | | |
| Encoder (Marker-Agnostic) | Stages ($L_{ma}$) | 1 |
| | Stage composition | 4-fold Downsampling + 4 ConvNeXt v2 blocks |
| | Stage dimension ($d_{ma}$) | 16 |
| Encoder (Pan-Marker) | Hyperkernel size ($h_e \times w_e$) | $1 \times 1$ |
| | Pan-Marker dimension ($d_{pm}$) | 192 |
| | Stages ($L_{pm}$) | 3 |
| | Stage composition | 4-fold Downsampling (skipped in first stage) + 4 ConvNeXt v2 blocks |
| | Stage dimensionalities | 192, 384, 768 |
| Latent feature map | Dimension ($d_{lat}$) | 768 |
| | Shape ($H_{lat} \times W_{lat}$) | $16 \times 16$ |
| Decoder | Hyperkernel size ($h_d \times w_d$) | $1 \times 1$ |
| | Marker-Specific dimension ($d_{ms}$) | 512 |
| | Marker-Agnostic decoder | 1 ConvNeXt v2 Block ($B_{dec} = 1$) |
| **`ImmuVis`$_{\texttt{Swin}}$** | | |
| Encoder (Marker-Agnostic) | Stages ($L_{ma}$) | 1 |
| | Stage composition | 4-fold Downsampling + 4 Swin Transformer blocks |
| | Stage dimension ($d_{ma}$) | 16 |
| Encoder (Pan-Marker) | Hyperkernel size ($h_e \times w_e$) | $1 \times 1$ |
| | Pan-Marker dimension ($d_{pm}$) | 192 |
| | Stages ($L_{pm}$) | 3 |
| | Stage composition | 4-fold Downsampling (skipped in first stage) + 4 Swin Transformer blocks |
| | Stage dimensionalities | 192, 384, 768 |
| Latent feature map | Dimension ($d_{lat}$) | 768 |
| | Shape ($H_{lat} \times W_{lat}$) | $16 \times 16$ |
| Decoder | Hyperkernel size ($h_d \times w_d$) | $1 \times 1$ |
| | Marker-Specific dimension ($d_{ms}$) | 512 |
| | Marker-Agnostic decoder | 1 ViT Block ($B_{dec} = 1$) |
| **`ImmuVis`$_{\texttt{ViT}}$** | | |
| Encoder (Marker-Agnostic) | Stages ($L_{ma}$) | 0 (Identity) |
| | Dimension ($d_{ma}$) | 1 |
| Encoder (Pan-Marker) | Hyperkernel / Patch size ($h_e \times w_e$) | $8 \times 8$ (Stride 8) |
| | Pan-Marker dimension ($d_{pm}$) | 768 |
| | Layers | 16 |
| | Attention heads | 8 |
| Latent feature map | Dimension ($d_{lat}$) | 768 |
| | Shape ($H_{lat} \times W_{lat}$) | $16 \times 16$ |
| Decoder | Hyperkernel size ($h_d \times w_d$) | $1 \times 1$ |
| | Marker-Specific dimension ($d_{ms}$) | 512 |
| | Marker-Agnostic decoder | 1 ViT Block ($B_{dec} = 1$) |

### A.5.1 Caveats of protein language model-based embeddings

Contrary to VirTues [13] and EVA [14], `ImmuVis` uses randomly initialized, learnable marker embeddings trained from scratch instead of PLM-based embeddings [34, 35]. This design choice is motivated by two main reasons:

Table 8: Conv stem ablation. Cell typing is Macro-F1; lower validation NLL and staining MSE are better.

| $L_{ma}$ | Validation NLL | Cell typing | Staining MSE |
|---|---|---|---|
| 0 (fixed-channel control) | $-6.601$ | .758 | .0672 |
| 1 (used) | $-7.410$ | .790 | .0573 |
| 2 | $-7.546$ | .788 | .0833 |

Table 9: Conv Objective ablation. Cell typing and clinical prediction report Macro-F1.

| Objective | Staining MSE | Pearson | Cell typing | Clinical |
|---|---|---|---|---|
| $\beta$-NLL | .0573 | .800 | .790 | .637 |
| MSE | .0501 | .806 | .768 | .632 |

- PLM embeddings are poorly defined for a substantial fraction of IMC markers; for example, DNA1/2, CD45RA+RO, and panCK lack unique UniProt entries or aggregate multiple proteins.
- Sequence similarity does not imply spatial or functional similarity in IMC, with notable examples including CD45RA/CD45RO (which have nearly identical ESM / GenePT embeddings but an expression correlation of only 0.35) and CK subtypes.

To illustrate the latter, we tested VirTues zero-shot marker prediction on AMY1A, which is present in the Steenbuck dataset [65] included in IMC17M, but absent from VirTues' training data. We selected this particular marker as AMY2A, a closely related isoform with a similar ESM2 embedding ($L_2 = 0.154$) and high sequence similarity, was present in VirTues' training data. We further compared two conditions:

- **Zero-shot:** VirTues predicts AMY1A based solely on the ESM embedding (AMY1A was absent from the training data, but with similar AMY2A present).
- **Trained:** VirTues$_{\texttt{IMC17M}}$, trained on IMC17M, with both AMY1A and AMY2A using VirTues preprocessing.

Despite the high sequence similarity between AMY1A and AMY2A, reconstruction quality degrades dramatically in the zero-shot setting (mean MSE 0.96 vs. 0.24; see Figure 3, for example reconstructions see Figure 4). This demonstrates that ESM sequence similarity does not reliably transfer to spatial co-expression similarity in IMC even when the accurate prediction is possible (as confirmed by a **Trained** setup).

### A.6 Optimization and training protocol.

All models are trained with AdamW [36] using weight decay $10^{-4}$ and a cosine annealing learning-rate schedule with linear warmup. The learning rate increases linearly during the first 5 epochs to a peak of $5 \times 10^{-4}$, and is then annealed to a final value of $10^{-6}$ by the end of training. We use mixed-precision training with `bfloat16` for efficiency. We used batch size of 8, as we observed lower batches tend to improve the convergence of models. We additionally apply $\ell_2$-norm gradient clipping with maximum norm 1.0. We train all `ImmuVis` variants for 200 epochs.

**Training compute resources.** All main models were trained on NVIDIA A100 SXM 40 GB GPUs, using one GPU per run. Pretraining each `ImmuVis` variant on 200 epochs required approximately 4 days. Downstream linear-probe experiments were run on CPU and required less than 24 hours per dataset.

### A.7 IMC Preprocessing Pipeline

IMC raw image data requires additional preprocessing prior to model training. We apply the following steps in order.

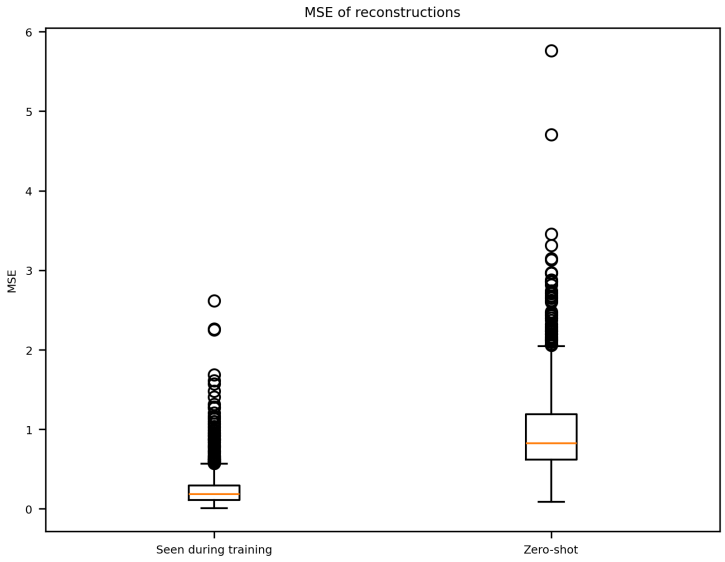



Figure 3: MSE of VirTues reconstructions in trained vs zero-shot settings.

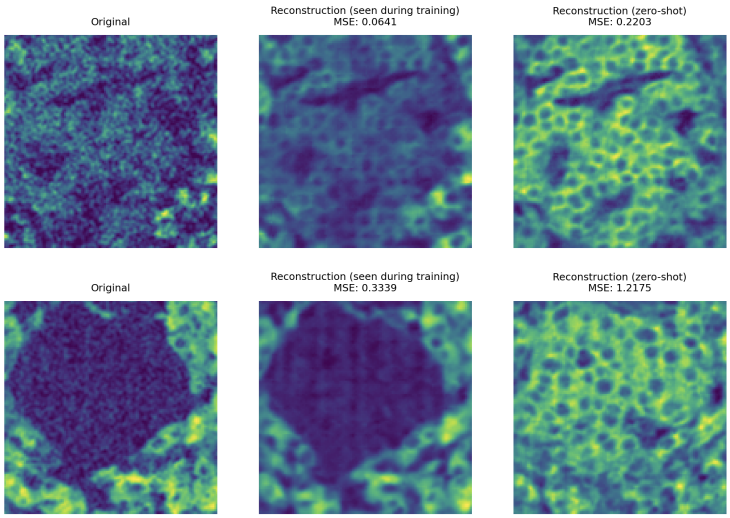



Figure 4: Examples of VirTues reconstructions for trained and zero-shot settings.

**Arcsinh transformation.** We apply a pixel-wise `arcsinh` transform with cofactor $5$ to compress the high dynamic range of IMC intensities while amplifying foreground signal and accommodating zero/negative values.

**Denoising via low-pass filtering.** We observed that IMC-specific background noise and hot pixels can destabilise optimisation, particularly when training on smaller image crops. To address this, we applied a Butterworth low-pass filter [37] implementation from `skimage.filters.butterworth` (v 0.25.2) [70], with cutoff frequency ratio $0.2$ relative to the Nyquist frequency and filter order $2$. These settings provide a maximally flat passband response up to $0.2 f_{\mathrm{N}}$, attenuating high-frequency spatial noise while preserving low-frequency tissue structure.

**Panel-wise intensity standardization.** We further normalize marker intensities to a $[0, 1]$ interval using a modified min-max scaling applied independently per dataset panel. We fix the lower bound at

0 and set the upper bound to the 99th percentile of the pooled intensity distribution across all markers within the panel, rounded up to one decimal place. This improves cross-dataset comparability while limiting the influence of extreme outliers.

**Subimage extraction and augmentation.** The model input is a $128 \times 128$ crops, where each pixel corresponds to approximately $1\ \mu\mathrm{m}^2$ tissue area. To mitigate I/O bound during training, we extracted non-overlapping subimages of shape $256 \times 256$ from the training split images. During training, we apply random rotation, reflection and cropping to generate crops from subimages of intended shape. For one training epoch, all subimages from all images are used once. During evaluation, we extract image crops via a deterministic central crop from each source image to ensure consistency and reproducibility.

### A.8 Comparison with VirTues preprocessing

Because our IMC Preprocessing pipeline differs from preprocessing of previously published methods [13, 14], below we will provide a justification for our choices. As EVA [14] does not provide preprocessing details for IMC, we will concentrate on VirTues [13] preprocessing pipeline.

#### A.8.1 `ImmuVis` arcsinh vs VirTues log1p transformation

Following established IMC/CyTOF analyses practice [71, 72, 73, 74], we apply $\texttt{arcsinh}(x/5)$ to raw per-pixel ion counts rather than the `log1p` transformation used by VirTues [13]. The distinction matters because IMC pixel values are discrete, non-negative ion counts, and the two transformations treat low-count measurements very differently.

`log1p` is asymptotically logarithmic, so equal multiplicative intervals receive equal compression regardless of the count magnitude. Concretely, the mapped differences for three count ranges of identical multiplicative width ($\times 4$) are:

$$\begin{aligned}\texttt{log1p}(3) - \texttt{log1p}(0) &= \ln 4 \approx 1.39,\\ \texttt{log1p}(255) - \texttt{log1p}(63) &= \ln 4 \approx 1.39,\\ \texttt{log1p}(1023) - \texttt{log1p}(255) &= \ln 4 \approx 1.39.\end{aligned}$$

The first interval (0–3 counts) falls squarely within the range attributable to spatial or channel spillover noise; the latter two correspond to biologically meaningful low-to-medium and medium-to-high expression transitions. Under `log1p`, all three differences receive identical weight, inflating the contribution of noise-level counts in any subsequent distance- or covariance-based computation.

$\texttt{arcsinh}(x/5)$ alleviates this by behaving approximately linearly near zero and logarithmically for large $x$, with the cofactor $c = 5$ controlling the transition point. The same three intervals now map to:

$$\begin{aligned}\texttt{arcsinh}(3/5) - \texttt{arcsinh}(0/5) &\approx 0.57,\\ \texttt{arcsinh}(255/5) - \texttt{arcsinh}(63/5) &\approx 1.39,\\ \texttt{arcsinh}(1023/5) - \texttt{arcsinh}(255/5) &\approx 1.39.\end{aligned}$$

The noise-level interval is compressed by a factor of roughly 2.5 comparing to the weight assigned to the biologically informative ranges, while the latter two remain comparably scaled, preserving the dynamic range that matters for biological analysis.

#### A.8.2 `ImmuVis` Denoising via low-pass filtering vs VirTues Gaussian blur

Instead of 3x3-truncated Gaussian blur used by VirTues we used a Butterworth [37] low-pass filtering (as described in Appendix A.7). The main reason behind this choice is a high intrinsic noise present in the IMC data (see Figure 5). To quantitatively measure this noise we introduce two pixel-level metrics: *gradient* and *curvature*. Let $x_0$ denote the intensity of a pixel with eight-connected neighbours $\{x_k\}_{k=1}^{8}$ (N, NE, E, SE, S, SW, W, NW). The *curvature*

$$m_1 = x_0 - \frac{1}{8}\sum_{k=1}^{8} x_k \tag{1}$$

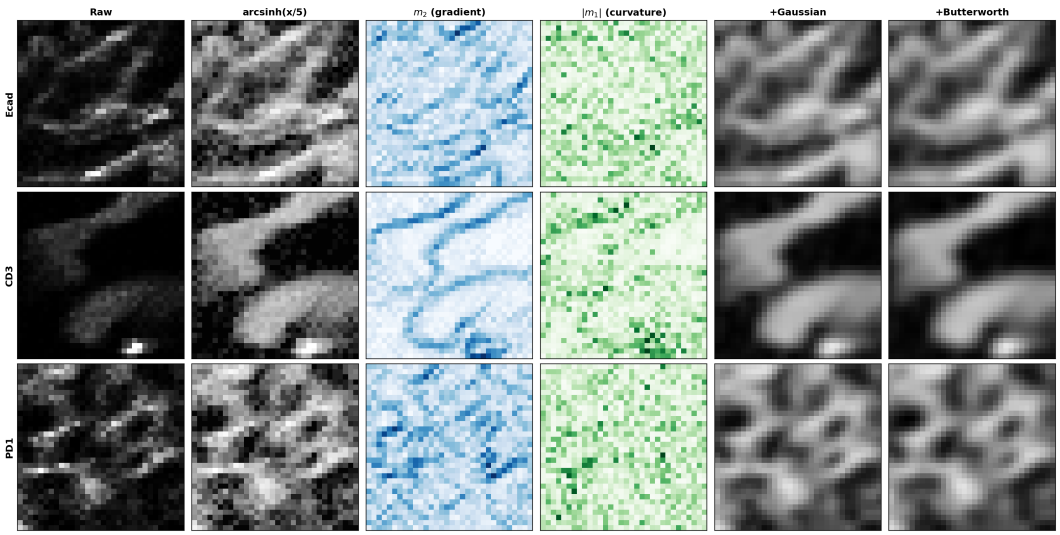



Figure 5: **Exemplary qualitative comparison of spatial preprocessing on $32 \times 32$ IMC patches from [46].** Each row shows a single antibody channel: E-cadherin (Ecad, epithelial marker), CD3 (T-cell marker), and PD-1. The six columns are: raw counts; arcsinh$(x/5)$ normalisation; the corresponding omnidirectional gradient $m_2$ and absolute curvature $|m_1|$ maps (computed on the arcsinh-normalised full image and coloured with sequential blue and green scales, respectively); arcsinh+Gaussian blur ($\sigma$=1 px, 3×3 kernel, analogous to VirTues preprocessing); and arcsinh+Butterworth low-pass filter (order 2, cutoff 0.2 Nyquist, used by `ImmuVis` ). Intensity columns share a common colour scale across arcsinh and filtered variants within each row; the raw column and the two metric columns are each scaled independently per row.

measures how much a pixel deviates from its local neighbourhood mean; it is equivalent to the discrete isotropic Laplacian and is sensitive to high-frequency structure such as shot noise and ringing artefacts. The *gradient*

$$m_2 = \sqrt{\left(\frac{x_E - x_W}{2}\right)^2 + \left(\frac{x_S - x_N}{2}\right)^2} \tag{2}$$

is the Euclidean magnitude of the central-difference gradient along the two principal axes, measuring the strength of local intensity transitions irrespective of orientation. Both metrics are computed after applying the full preprocessing pipeline; $|m_1|$ is used to obtain a non-negative curvature magnitude. Periodic boundary conditions are applied via circular array shifts.

The analysis of these pixel-level metrics computed on IMC images from [46] reveal interesting noise dependencies (see Figure 6). For both raw data (without filtering) and Gaussian blur, there is a significant curvature divergence for a flat image areas ($|m2| < 10^{-2}$). Low-pass filtering removes this divergence providing almost linear dependency between $|m1|$ and $|m2|$ within this range (as visualized by $\mathbb{E}[|m_1| \mid m_2]$ curve).

### A.8.3 `ImmuVis` min–99th-percentile panel-wise scaling vs. VirTues per-marker z-score intensity standardization

Compared with our min–99th-percentile scaling computed over the full dataset, VirTues applies per-marker z-score normalization based on the statistics computed on cell masks. MSE values obtained under these two preprocessing protocols are therefore not directly comparable. Our choice is motivated by three main reasons:

- Even after preprocessing with `log1p` or `arcsinh`, IMC data remain zero-inflated and hence highly non-Gaussian. Therefore, we use min–99th-percentile scaling, which is better suited to this type of data.

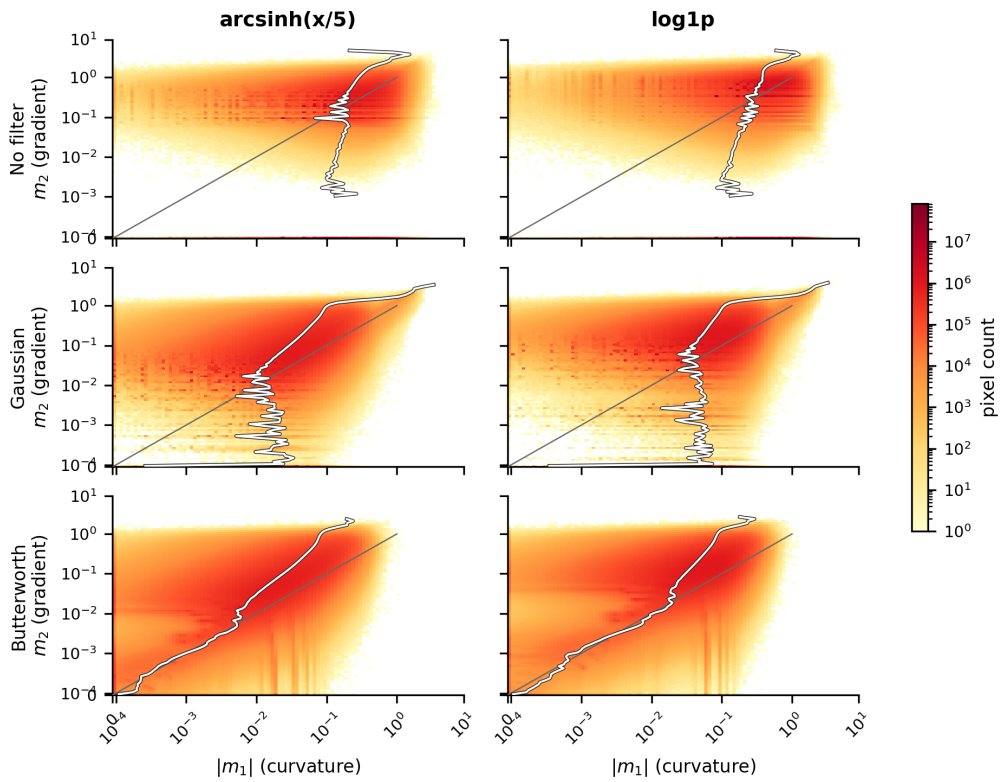



Figure 6: **Joint distribution of curvature $|m_1|$ and gradient $m_2$ under six preprocessing pipelines.** Rows: no spatial filter / Butterworth / Gaussian; columns: arcsinh$(x/5)$ / $\log(1+x)$ normalisation. Each panel accumulates a 2D histogram over all pixels, images, and antibody channels, using a mixed bin scale (one zero-anchor bin $[0, 10^{-4}]$ followed by 99 log-spaced bins up to 10). All six panels share a logarithmic colour scale. The white curve shows $\mathbb{E}[|m_1| \mid m_2]$; the diagonal $|m_1| = m_2$ is for a reference. Both filters reduce pixel curvature relative to the unfiltered signal (top row); Gaussian smoothing introduces discrete curvature artefacts at fixed $|m_1|$ values, whereas Butterworth low-pass filtering yields a smoother, more uniform $(|m_1|, m_2)$ distribution.

- The per-marker, per-dataset statistics required for VirTues z-score normalization are unavailable for markers absent from a panel, what makes results of virtual staining difficult to interpret in such cases, as they cannot be converted back to an absolute value scale. Additionally such transformation invalidates out-of-panel and out-of-cohort task assumptions. In contrast, `ImmuVis` quantile scaling is zero-preserving, and because the scaling quantile is computed over the full dataset, it can be easily approximated using only the markers present in the panel.
- As z-score statistics used by VirTues require cell segmentation, this approach is dependent on a segmentation model and quality of the segmentation masks.

### A.9 KRONOS reproduction

The initial attempts to pre-train a self-supervised foundation model KRONOS [28] on our IMC dataset used the native KRONOS implementation. This architecture relied on a standard Vision Transformer stem and incorporated per-channel patch embeddings together with sinusoidal marker encodings. However, these initial experiments exhibited severe optimisation instabilities, resulting in consistent loss divergence.

This empirical challenge is consistent with observations reported by the original KRONOS authors, who explicitly excluded ion-based modalities, such as IMC, from their training corpus, noting that such data might require modality-specific architectural adaptations. In addition, unlike `ImmuVis` and

VirTues, KRONOS employs ambiguous marker encodings, where markers are arranged in an ordered list such that biologically or semantically similar markers are positioned close to one another. This introduces implicit assumptions about marker relationships and ordering that may not generalise well across heterogeneous IMC panels. Despite evaluating multiple alternative marker orderings, none of these configurations resulted in stable convergence during training.

To mitigate the exploding loss, we first experimented with expanding the marker embedding vocabulary using custom sinusoidal encodings and explicitly injecting spatial coordinate information via a meshgrid representation. However, these modifications proved insufficient to stabilise the training dynamics.

### A.10 Transfer to external panels

The ten test panels comprise held-out IMMUcan Head & Neck [38] and nine external panels from four studies absent from IMC17M: Allam et al. [39], Hu et al. [40], Moldoveanu et al. [41], and Schulz et al. [42]. These four studies were included in the original VirTues pretraining corpus. The retrained VirTues and EVA baselines below use IMC17M and OP. All models receive the same in-vocabulary inputs and are scored on the same overlapping targets; no new-marker adaptation is used in this comparison. $\texttt{ImmuVis}_{\texttt{Conv}}$ and $\texttt{ImmuVis}_{\texttt{Swin}}$ each achieve the lowest panel-average MSE on three of the ten panels; VirTues leads on four. Paired Wilcoxon signed-rank tests on image–marker MSE with Benjamini–Hochberg FDR correction yield wins/losses/non-significant differences versus VirTues of 6/1/3 for $\texttt{ImmuVis}_{\texttt{Swin}}$, and 3/3/4 for both $\texttt{ImmuVis}_{\texttt{Conv}}$ and $\texttt{ImmuVis}_{\texttt{ViT}}$. Thus, the multi-panel staining advantage is carried by $\texttt{ImmuVis}_{\texttt{Swin}}$.

**Embedding-only adaptation.** The external Moldoveanu cohort contains 48 markers, of which 43 are in the pretrained vocabulary. For $\texttt{ImmuVis}_{\texttt{Conv}}$, we add embeddings for CCR7, CD29, MEK1-2, pMEK1-2, and S100, training only these embeddings with the remaining network frozen. Table 10 compares adaptation with no adaptation and with pretraining that includes the cohort. The no-adaptation staining score is omitted because it averages only 43 markers and is not directly comparable to the 48-marker evaluations. Embedding-only adaptation used the pretraining objective and hyperparameters and took approximately 2 hours on one A100 (about 2% of pretraining compute). Downstream tasks on the other three external studies were not run.

Table 10: Moldoveanu transfer with $\texttt{ImmuVis}_{\texttt{Conv}}$: cell-typing Macro-F1 for panels P1/P2 and staining on all 48 markers where available. The cohort was included in original VirTues pretraining.

| Setting | P1 | P2 | MSE | Pearson |
|---|---|---|---|---|
| No adaptation | .765 | .643 | – | – |
| New-marker embeddings only | .765 | .641 | .081 | .531 |
| Cohort included in pretraining | .776 | .649 | .040 | .687 |
| Original VirTues | .734 | .617 | – | – |

**Full-panel versus overlap-only inputs.** For the five panels with markers absent from IMC17M, we compare full-panel and overlap-only inputs for the retrained IMC17M+OP baselines. Both conditions are scored on the same overlapping targets (Table 11). Extra marker inputs worsen VirTues MSE on three of five panels and Pearson on four; EVA MSE improves on four and worsens on one. Thus, unseen-marker inputs are not uniformly beneficial, and the VirTues pattern does not generalize to EVA.

Table 11: MSE/Pearson on shared targets with full-panel (F) and overlap-only (O) inputs.

| Panel | VirTues F | VirTues O | EVA F | EVA O |
|---|---|---|---|---|
| A1 | .0348/.62 | .0350/.65 | .0659/.50 | .0672/.50 |
| A2 | .1853/.65 | .1853/.65 | .3498/.48 | .4618/.47 |
| H1 | .0661/.29 | .0559/.34 | .1060/.18 | .1080/.18 |
| M1 | .2634/.52 | .2011/.54 | .3577/.37 | .3612/.39 |
| M2 | .2299/.24 | .1849/.29 | .3112/.20 | .3019/.21 |

Table 12: Virtual staining accuracy measured by MSE and Pearson correlation under the VirTues preprocessing protocol. **Bold** denotes the best score.

| Model | MSE (↓) | Pearson (↑) |
|---|---|---|
| VirTues | .797 | .700 |
| $\text{VirTues}_{\texttt{IMC17M}}$ | .410 | .764 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | .502 | .731 |
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | **.400** | **.783** |

### A.11 Virtual staining results for VirTues preprocessing (VP)

Quantitative results for virtual staining on HN with VirTues preprocessing (VP) are presented in Table 12. A marker-wise MSE breakdown, together with statistical significance evaluation, are presented in Figure 7. Figure 8 presents Pearson correlation distributions together with statistical significance evaluation.

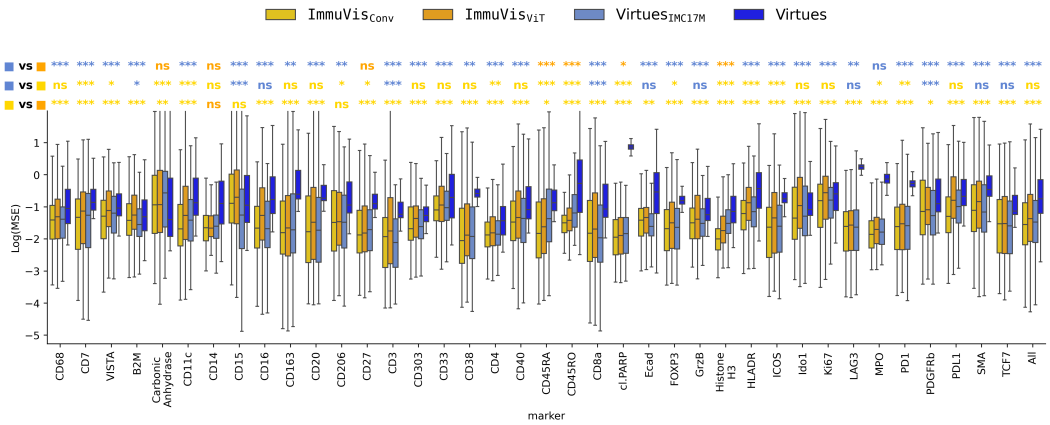



Figure 7: **Quantitative virtual staining evaluation for MSE on the Head & Neck cohort for models train using VirTues preprocessing (VP).** Per-marker reconstruction accuracy measured by image-level $\log(\text{MSE})$ for $\texttt{ImmuVis}_{\texttt{Conv}}$, $\texttt{ImmuVis}_{\texttt{ViT}}$, $\text{VirTues}_{\texttt{IMC17M}}$ and VirTues (original checkpoint from [13]). Each boxplot summarizes the distribution of image-level scores, where each image score is obtained by averaging patch-level errors over all patches from that image. The three rows above the plot report paired Wilcoxon signed-rank tests results across images for the corresponding model pairs (as indicated in the left top corner), with significance after FDR correction (ns - not significant; $(^{*}) < 0.05$; $(^{**}) < 0.01$; $(^{***}) < 0.001$).

### A.12 On computing Pearson correlation across the whole dataset rather than on a per-image basis

Contrary to VirTues [13], we compute the Pearson correlation between the ground truth and the model predictions on a per-marker basis across the entire dataset, rather than independently for each image. The main motivation for this choice is that Pearson correlation computed on a per-image basis is linearly equivalent to the mean squared error (MSE) computed on data that have been z-scored independently for each image. This method completely loses cross-image marker dependencies and may additionally suffer from variance explosion for frequent images and markers with near-constant zero expression values.

To demonstrate this linear equivalence, assume that two variables $x$ and $y$ are z-scored, i.e.,

$$\mathbb{E}[x] = \mathbb{E}[y] = 0, \qquad \text{Var}(x) = \text{Var}(y) = 1.$$

The mean squared error between $x$ and $y$ is

$$\text{MSE}(x, y) = \mathbb{E}\left[(x - y)^2\right].$$

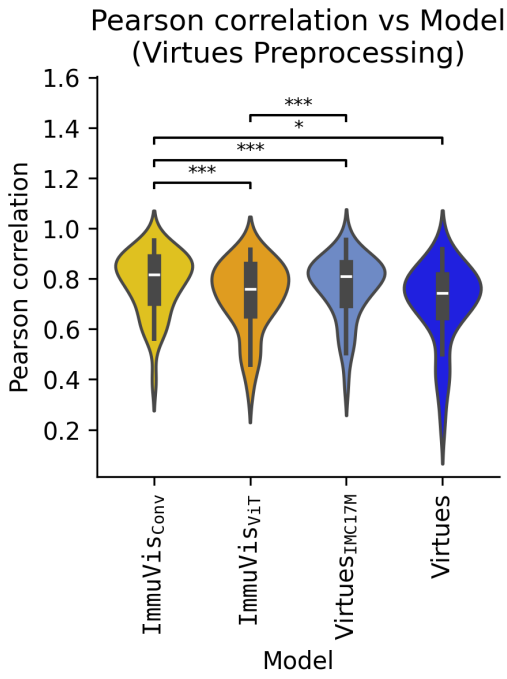



Figure 8: **Quantitative virtual staining evaluation for Pearson correlation on the Head & Neck cohort for models trained using VirTues preprocessing (VP).** Per-model Pearson correlation measured on a marker-level for $\texttt{ImmuVis}_{\texttt{Conv}}$, $\texttt{ImmuVis}_{\texttt{ViT}}$, $\text{VirTues}_{\texttt{IMC17M}}$ and VirTues (original checkpoint from [13]). Each boxplot summarizes the distribution of marker-level Pearson correlation computed across all images. P-values are obtained via Wilcoxon signed-rank tests results across markers for the corresponding model pairs (ns - not significant; $(^{*}) < 0.05$; $(^{**}) < 0.01$; $(^{***}) < 0.001$).

Expanding the square gives

$$\mathrm{MSE}(x, y) = \mathbb{E}[x^2] + \mathbb{E}[y^2] - 2\mathbb{E}[xy].$$

Because $x$ and $y$ are z-scored,

$$\mathbb{E}[x^2] = \mathbb{E}[y^2] = 1,$$

and since their means are zero,

$$\mathrm{Corr}(x, y) = \mathbb{E}[xy].$$

Therefore,

$$\mathrm{MSE}(x, y) = 1 + 1 - 2\,\mathrm{Corr}(x, y) = 2 - 2\,\mathrm{Corr}(x, y).$$

Rearranging yields

$$\mathrm{Corr}(x, y) = 1 - \frac{1}{2}\mathrm{MSE}(x, y).$$

Thus, for z-scored variables, Pearson correlation and mean squared error are linearly related.

### A.13 Extended virtual staining results for `ImmuVis` preprocessing (OP)

A marker-wise MSE breakdown on HN, together with statistical comparisons, is presented in Figure 9. Figure 10 presents Pearson correlation distributions together with statistical comparisons.

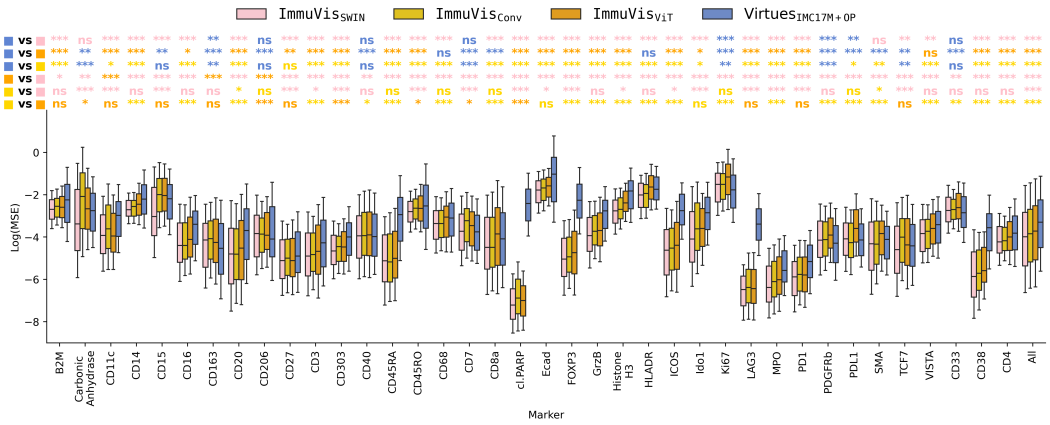



Figure 9: **Quantitative virtual staining evaluation for MSE on the Head & Neck cohort for models train using `ImmuVis` preprocessing (`OP`).** Per-marker reconstruction accuracy measured by image-level $\log(\mathrm{MSE})$ for $\texttt{ImmuVis}_{\texttt{Conv}}$, $\texttt{ImmuVis}_{\texttt{Swin}}$, $\texttt{ImmuVis}_{\texttt{ViT}}$, and $\text{VirTues}_{\texttt{IMC17M+OP}}$. Each boxplot summarizes the distribution of image-level scores, where each image score is obtained by averaging patch-level errors over all patches from that image. The three rows above the plot report paired Wilcoxon signed-rank tests results across images for the corresponding model pairs (as indicated in the left top corner), with significance after FDR correction (ns - not significant; $(^{*}) < 0.05$; $(^{**}) < 0.01$; $(^{***}) < 0.001$).

### A.14 Extended uncertainty calibration analysis

This appendix expands the headline results in Section 4.3.2 with the full set of calibration diagnostics. All figures are computed on out-of-cohort leave-one-out (LOO) virtual staining over the IMMUcan SCCHN1 test split [38] (the cohort was excluded at training), across all 40 markers of the cohort. Per model this comprises 149 patches $\times$ 40 markers $=$ 5 960 patch–channel reconstructions.

**Notation.** For each pixel let $\mu$ denote the predicted mean, $\sigma^2$ the predicted variance from the heteroscedastic head, $y$ the arcsinh-normalised target, and $r = y - \mu$ the residual. Throughout, $|r|$ denotes absolute residual at the pixel level and MAE the per-(image, channel) mean of $|r|$.

#### A.14.1 Per-(image, channel) error–uncertainty correlation

**Method.** For every (image, channel) pair we summarise the predicted uncertainty by $\log \overline{\sigma^2} = \log\big(\frac{1}{N}\sum_i \sigma_i^2\big)$ and the reconstruction error by $\log \mathrm{MAE}$. Each LOO reconstruction contributes one such (uncertainty, error) point in log–log space. Pearson correlation is computed on the 5 960 patch–channel rows per model; we overlay an ordinary-least-squares fit per model.

#### A.14.2 Reliability of the predictive Gaussian

**Method.** A reliability diagram tests whether the predictive distribution $\mathcal{N}(\mu, \sigma^2)$ assigns calibrated coverage to its central intervals. For each nominal level $\alpha \in \{.01, .02, \ldots, .99\}$ (99-point grid) we compute the half-width $z_\alpha = \Phi^{-1}\big((1+\alpha)/2\big)$ and the empirical coverage

$$\widehat{\mathrm{cov}}(\alpha) \;=\; \frac{1}{N_{\mathrm{px}}} \sum_{i=1}^{N_{\mathrm{px}}} \mathbf{1}\big[|r_i| \le z_\alpha\, \sigma_i\big],$$

on all pixels. Implementation note: the per-$\alpha$ indicator is counted once per (patch, channel) and summed at the group level, so the coverage at every $\alpha$ is computed losslessly without weighted-average approximation. A perfectly calibrated heteroscedastic head produces $\widehat{\mathrm{cov}}(\alpha) = \alpha$ for every $\alpha$ - the diagonal $y = \alpha$ in the main panel of Figure 12. The single-number summary is the regression

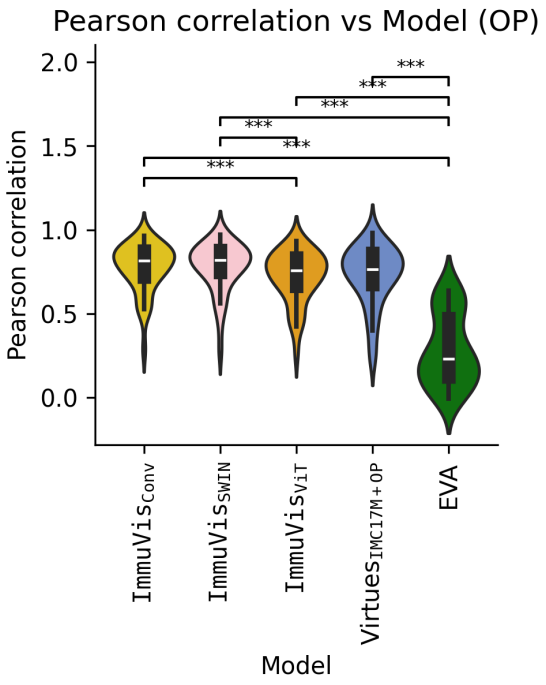



Figure 10: **Quantitative virtual staining evaluation for Pearson correlation on the Head & Neck cohort for models train using `ImmuVis` preprocessing (`OP`).** Per-model Pearson correlation measured on a marker-level for $\texttt{ImmuVis}_{\texttt{Conv}}$, $\texttt{ImmuVis}_{\texttt{Swin}}$, $\texttt{ImmuVis}_{\texttt{ViT}}$, $\text{VirTues}_{\texttt{IMC17M+OP}}$ and $\text{EVA}_{\texttt{OP}}$ (released checkpoint evaluated outside its native preprocessing). Each boxplot summarizes the distribution of marker-level Pearson correlation computed across all images. P-values are obtained via Wilcoxon signed-rank tests results across markers for the corresponding model pairs (ns - not significant; $(^{*}) < 0.05$; $(^{**}) < 0.01$; $(^{***}) < 0.001$).

Expected Calibration Error

$$\mathrm{ECE} \;=\; \mathbb{E}_{\alpha}\left|\widehat{\mathrm{cov}}(\alpha) - \alpha\right| \;\approx\; \frac{1}{99}\sum_{\alpha}\left|\widehat{\mathrm{cov}}(\alpha) - \alpha\right|,$$

following [43] which is the metric reported in Table 2.

### A.14.3 Sharpness vs. calibration

**Method.** Low ECE confirms that the predictive intervals contain the right fraction of pixels but not that those intervals are tight. Sharpness captures this complementary axis [75]; for a marker $m$ we report the pixel-weighted mean of the predicted standard deviation,

$$\mathrm{sharpness}(m) \;=\; \frac{\sum_{\mathrm{pc}\,:\,\mathrm{marker}=m} N_{\mathrm{pc}}\,\overline{\sigma}_{\mathrm{pc}}}{\sum_{\mathrm{pc}\,:\,\mathrm{marker}=m} N_{\mathrm{pc}}}.$$

The global sharpness, averaged across the 40 markers, is .046, .048, and .041 for $\texttt{ImmuVis}_{\texttt{Conv}}$, $\texttt{ImmuVis}_{\texttt{ViT}}$, and $\texttt{ImmuVis}_{\texttt{Swin}}$ respectively. Plotted jointly with per-marker ECE in Figure 13, the desirable region is the lower-left (small $\sigma$ *and* well-calibrated). The upper-right corner (large $\sigma$, large ECE) flags markers where the model is both unsure and miscalibrated; the upper-left flags markers where the model is precise but its uncertainty does not match the residuals.

### A.14.4 Sparsification curves and AUSE

**Method.** The sparsification analysis evaluates how well $\sigma$ *ranks* pixels by error, integrating over thresholds rather than picking one. We sort all pixels by their predicted $\sigma$ and progressively remove

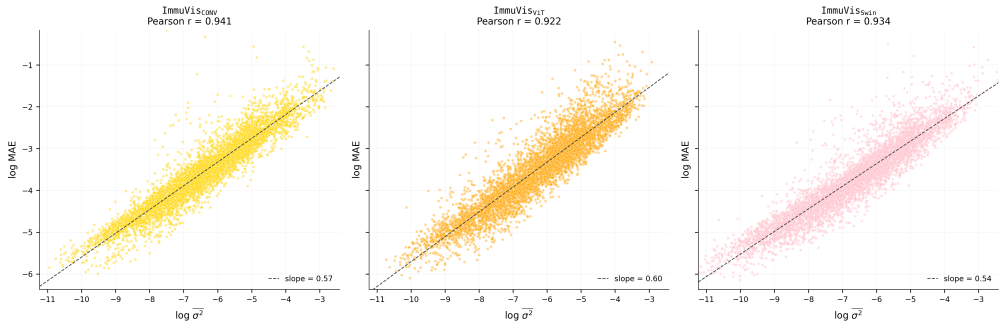



Figure 11: **Predicted uncertainty tracks reconstruction error across (image, channel) pairs under out-of-cohort LOO virtual staining.** Each point is one LOO reconstruction (5 960 per model). Strong log–log scaling between mean predicted variance and MAE holds for all three `ImmuVis` variants: Pearson $r_{\mathrm{CONV}} = .941$, $r_{\mathrm{ViT}} = .922$, $r_{\mathrm{Swin}} = .934$. The OLS fit slopes (`ImmuVis`$_{\texttt{Conv}}$, `ImmuVis`$_{\texttt{ViT}}$, `ImmuVis`$_{\texttt{Swin}}$) are .57, .60, .54 respectively. The relationship confirms that the heteroscedastic head captures local ambiguity in the reconstruction.

the most-uncertain fraction $f \in [0, 1)$; the residual error of the kept set, with RMSE as the base metric (matching Table 2), is

$$\mathrm{err}_\sigma(f) \;=\; \mathrm{RMSE}\Big(\big\{r_i \;:\; \sigma_i \le Q_{1-f}(\sigma)\big\}\Big).$$

The same construction with MAE in place of RMSE gives an alternative AUSE that is more sensitive to the bulk of the residual distribution and less to outliers; we report it in Figure 15 as a complementary view. The *oracle* curve $\mathrm{err}_{\mathrm{oracle}}(f)$ replaces $\sigma$ with $|r|$ in the sort key; this is the lowest possible sparsification curve given the same residual distribution. The *sparsification error* is

$$\mathrm{SE}(f) \;=\; \mathrm{err}_\sigma(f) - \mathrm{err}_{\mathrm{oracle}}(f) \;\ge\; 0,$$

and its area integrates to AUSE,

$$\mathrm{AUSE} \;=\; \int_0^1 \big[\mathrm{err}_\sigma(f) - \mathrm{err}_{\mathrm{oracle}}(f)\big]\, df,$$

the standard scalar uncertainty-quality metric [44, 45]. Lower is better; $\mathrm{AUSE} = 0$ iff $\sigma$ ranks pixels identically to $|r|$. The integral is evaluated by the trapezoidal rule over 100 equally spaced fractions. The metric is computed on the full pixel population.

### A.15 Retrospective PD-1 uncertainty analysis

We use `ImmuVis`$_{\texttt{Conv}}$ to predict PD-1 leave-one-out on 21,058 segmented CD8+ cells from IMMUcan NSCLC2 panel 1. We extract cell-centred $32 \times 32$ crops, exclude PD-1 from the inputs, and average predicted expression over the cell mask. Reference PD-1 labels are assignments from a two-component Gaussian mixture fitted to measured, preprocessed cell-mean expression on the arcsinh scale; the component with the larger mean is labelled positive. A prediction threshold is chosen by maximizing Youden's $J$ against these reference labels on the evaluated cells.

Cell uncertainty is summarized by the mean predicted pixel standard deviation. The error-ranking ROC-AUC is .78, with error rates of 8.4% and 77% in the lowest- and highest-uncertainty deciles. Mean pixel variance and maximum pixel variance give ROC-AUCs of .78 and .74, respectively. These are descriptive uncertainty summaries, rather than estimates of the variance of the cell mean under a validated pixel-covariance model. Since reference-label fitting and threshold selection use the evaluated cells, this is a retrospective error-ranking analysis, not an independent assessment of clinical diagnostic performance.

### A.16 Single-cell pre-processing

For each segmented cell from the [46] dataset, we extract a $32 \times 32$ pixel patch centred at the cell mask centroid. All pixels outside the target cell mask are zeroed to remove neighborhood information while

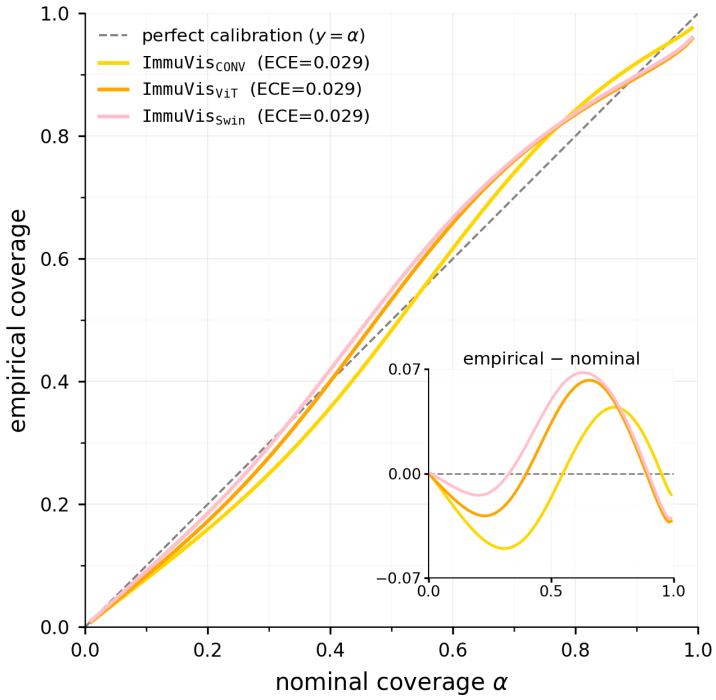



Figure 12: **Reliability diagram - global, out-of-cohort LOO.** Empirical coverage of the predictive Gaussian's central $\alpha$-interval as a function of nominal $\alpha$. All three `ImmuVis` variants closely track the perfect-calibration line and their ECE values agree to within .0006. The inset, plotted on a magnified $y$-axis, reveals that the three architectures miscalibrate in qualitatively different directions: `ImmuVis`$_{\texttt{Conv}}$ predominantly under-covers (peaking near $\alpha \approx 0.3$), while `ImmuVis`$_{\texttt{ViT}}$ and `ImmuVis`$_{\texttt{Swin}}$ predominantly over-cover (peaking near $\alpha \approx 0.7$, with `ImmuVis`$_{\texttt{Swin}}$ deviating most). Integrating the absolute deviations gives nearly identical ECE values ($.029 \pm .0003$), indicating that overall calibration quality is comparable across architectures despite distinct miscalibration shapes.

preserving within-cell marker intensities. Each crop is encoded by the frozen encoder (no fine-tuning) of all considered models from `ImmuVis` family, and VirTues to obtain a spatial embedding map, which we reduce to a fixed-length vector ($d_{\text{lat}}$) via spatial average pooling.

### A.17 Cell phenotyping results for different cell phenotyping methods

Quantitative results for cell phenotyping across model families and preprocessing configurations are presented in Table 13, extending the main results reported in Table 3. In addition to the `ImmuVis` cell-typing approach used in the main paper (denoted as `patch`), we also evaluate the cell-typing strategy introduced by the VirTues authors (denoted as `context`). Subscripts denote training and preprocessing configurations: `OP` refers to ImmuVis preprocessing, `VP` to VirTues preprocessing, and `IMC17M` to VirTues trained on our dataset. The mean-marker baseline requires segmentation masks and produces feature vectors whose dimensions correspond to the measured markers. Transferring a classifier across different panels therefore requires explicit alignment of marker features. This baseline does not provide virtual staining.

In the `patch` method, for each cell identified from the segmentation mask, we extract a $32 \times 32$ pixel patch centered at the cell centroid. All pixels outside the target cell mask are zeroed to remove neighborhood information. Each patch is then processed by the frozen encoder of all considered models from the `ImmuVis` family, as well as VirTues, to obtain a spatial embedding map, which is subsequently reduced to a fixed-length vector ($d_{\text{lat}}$) via spatial average pooling.

The `context` method follows the design proposed by the VirTues authors and incorporates both cellular and local microenvironment information. Each image, together with its corresponding cell mask, is partitioned into overlapping $128 \times 128$ crops using a stride of 42 pixels. Each crop

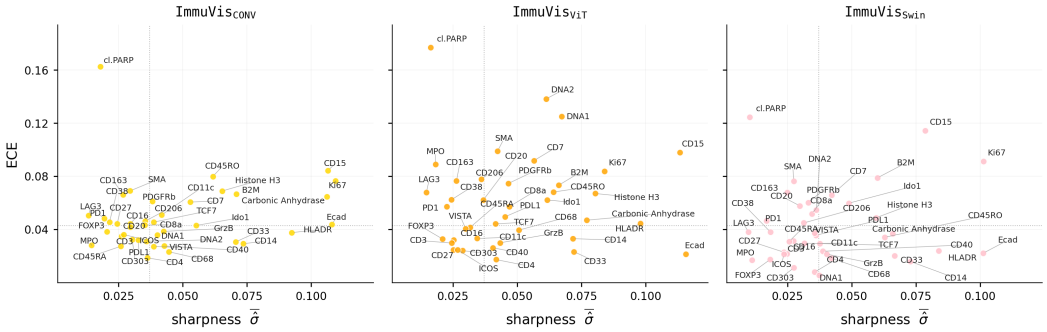



Figure 13: **Sharpness vs. ECE per marker.** Each point is one of the 40 markers in the Head & Neck cohort. The bulk of markers cluster in the lower-left quadrant (mean $\sigma \lesssim .05$, ECE $< .05$), confirming that the calibrated coverage in Figure 12 is achieved at narrow predictive intervals.

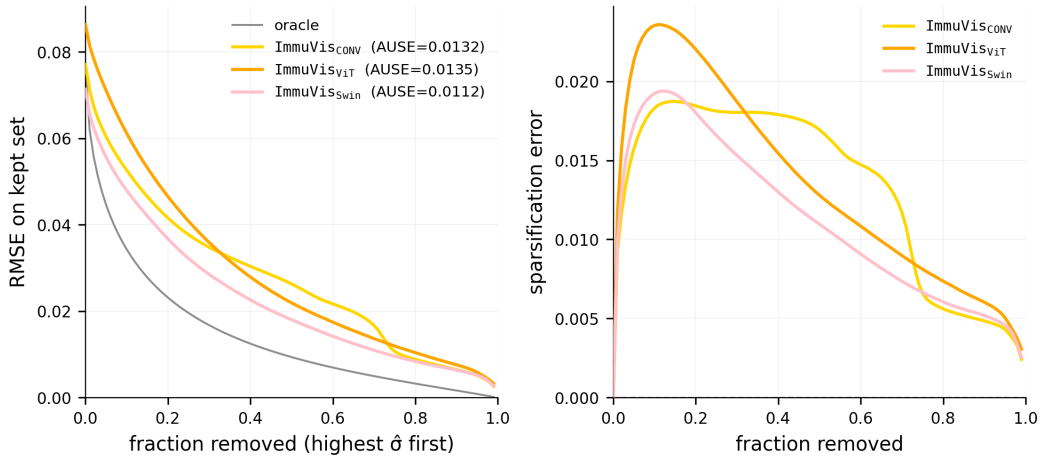



Figure 14: **Sparsification analysis - global, out-of-cohort LOO (RMSE base, matches Table 2).** *Left:* residual RMSE on the kept $(1 - f)$ fraction of pixels after removing the most-uncertain $f$ by predicted $\sigma$ (coloured lines per model); the grey oracle curve is obtained by ranking with $|r|$ instead and represents the lowest achievable curve at this residual distribution. AUSE in the legend is the area between each model's curve and the oracle. *Right:* the sparsification-error integrand $\mathrm{SE}(f) = \mathrm{err}_{\sigma}(f) - \mathrm{err}_{\mathrm{oracle}}(f)$. All three `ImmuVis` variants track the oracle closely across the full range; the residual gap is concentrated in the highest-$\sigma$ tail ($f \lesssim 0.2$), where outlier pixels are hardest to rank exactly. `ImmuVis`$_{\texttt{Swin}}$ achieves the lowest AUSE, while `ImmuVis`$_{\texttt{ViT}}$ exhibits the largest peak sparsification error.

is further decomposed into $8 \times 8$ patches. For VirTues, which is based on a Vision Transformer architecture, embeddings are produced natively at the patch level. For ImmuVis, each $8 \times 8$ patch is independently embedded as an image region. To obtain cell-level representations, patch embeddings are aggregated using a mask-informed weighted average: for each cell, the contribution of a given patch is proportional to the fraction of pixels within that patch belonging to the cell, ensuring that patches overlapping multiple cells contribute proportionally to each cell's embedding.

The reported `ImmuVis`$_{\texttt{Conv}}$ context scores use the original protocol of independently encoding $8 \times 8$ patches, although the network was trained on $128 \times 128$ inputs and downsamples fourfold per stage. Encoding the full crop once and pooling patch-aligned latent regions with the same mask weights improves the Eling context Macro-F1 from .533 to .752 (patch protocol: .790; `ImmuVis`$_{\texttt{ViT}}$ context: .759). Danenberg and Cords context scores (.488 and .666) were not recomputed. Tables 13 and 15 retain the original context protocol for `ImmuVis`$_{\texttt{Conv}}$; the corrected Eling score is reported separately here.

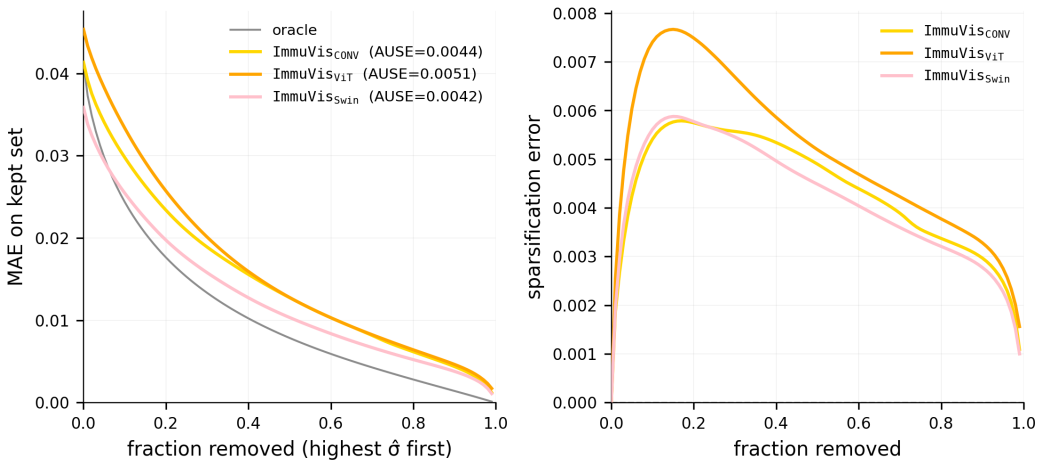



Figure 15: **Same construction as Figure 14** with MAE in place of RMSE. The model ranking is preserved ($\texttt{ImmuVis}_{\texttt{Swin}} < \texttt{ImmuVis}_{\texttt{Conv}} < \texttt{ImmuVis}_{\texttt{ViT}}$), confirming the $\sigma$-ordering quality differences are robust to the error norm choice.

Complementary results in terms of Average Precision are reported in Tables 14 and 15, while Figures 16, 17, and 18 provide dataset-specific macro-F1 comparisons with statistical significance testing across all evaluated datasets.

Table 13: **Representation learning performance for cell typing.** Macro-F1 scores are reported for 10-fold cross-validation; higher values indicate better performance. Standard deviations across folds were below $10^{-2}$ for all entries and are therefore omitted. The best score for each endpoint is shown in **bold**. The best method per endpoint / method is underlined.

| Model | [46] | | [47] | | [48] | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| Method | patch | context | patch | context | patch | context | patch | context |
| $\text{EVA}_{\texttt{OP}}$ | .602 | .450 | .510 | .487 | .691 | .571 | .601 | .503 |
| VirTues | .726 | .675 | .555 | .569 | .711 | .699 | .664 | .648 |
| $\text{VirTues}_{\texttt{IMC17M}}$ | .703 | .672 | .517 | .555 | .712 | .704 | .644 | .644 |
| $\text{VirTues}_{\texttt{IMC17M+OP}}$ | .706 | .686 | .557 | .579 | .716 | .702 | .660 | .656 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | .753 | **.759** | **.592** | **.618** | **.751** | **.740** | .699 | **.706** |
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | **.790** | .533 | **.592** | .488 | .745 | .666 | **.709** | .562 |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | .762 | .752 | .582 | .611 | .738 | .732 | .694 | .698 |
| $\texttt{ImmuVis}_{\texttt{VIT+VP}}$ | .757 | .750 | .581 | .610 | .723 | .722 | .687 | .694 |
| $\texttt{ImmuVis}_{\texttt{Conv+VP}}$ | .671 | .638 | .551 | .540 | .691 | .652 | .638 | .610 |

### A.18 Extended clinical endpoint prediction results across model families and preprocessing protocols

Quantitative results for clinical endpoint prediction across model families and preprocessing configurations are reported in Table 16, extending the main results of Table 4. In addition to the linear probe (LogReg) used in the main paper, this extended evaluation also includes the abMIL probe. Subscripts denote training and preprocessing configurations: `OP` refers to ImmuVis preprocessing, `VP` to VirTues preprocessing, and `IMC17M` to VirTues trained on our dataset.

**Mean-marker baseline.** We compute mean marker intensities within non-overlapping $128 \times 128$ patches and average the resulting vectors per image. We evaluate raw intensities, OP, and VP preprocessing. The VP configuration applies per-marker 99th-percentile clipping based on non-zero pixels, log1p transformation, Gaussian smoothing ($\sigma = 1$), and intensity clipping and rescaling. It

Table 14: Representation learning performance for cell typing (Average Precision; higher is better). Scores are reported from 10-fold cross-validation. Standard deviations across folds were below $10^{-2}$ for all entries and are therefore omitted. For VirTues evaluations, the reported value is the higher score obtained using either the authors' original or the `ImmuVis` cell-typing approach (see A.17). (best score within models family, **best score** for each endpoint)

| Model | [46] | [47] | [48] | Avg. |
|---|---|---|---|---|
| $\text{EVA}_{\text{OP}}$ | .664 | .559 | .748 | .657 |
| VirTues | .812 | .635* | .781 | .743 |
| $\text{VirTues}_{\text{IMC17M}}$ | .786 | .624* | .783 | .731 |
| $\text{VirTues}_{\text{IMC17M+OP}}$ | .796 | .651* | .783 | .743 |
| $\texttt{ImmuVis}_{\text{ViT}}$ | .845 | .662 | **.827** | .778 |
| $\texttt{ImmuVis}_{\text{Conv}}$ | **.882** | **.668** | .822 | **.791** |
| $\texttt{ImmuVis}_{\text{Swin}}$ | .850 | .654 | .807 | .770 |

* VirTues cell-typing method gave better results.

Table 15: **Representation learning performance for cell typing.** $\texttt{ImmuVis}_{\text{Conv}}$ context entries use the original $8 \times 8$ protocol (see text for corrected Eling Macro-F1). Average Precision scores are reported for 10-fold cross-validation; higher values indicate better performance. Standard deviations across folds were below $10^{-2}$ for all entries and are therefore omitted. The best score for each endpoint is shown in **bold**. The best method per endpoint / method is underlined.

| Model | [46] | | [47] | | [48] | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| Method | `patch` | `context` | `patch` | `context` | `patch` | `context` | `patch` | `context` |
| $\text{EVA}_{\text{OP}}$ | .664 | .472 | .559 | .534 | .748 | .606 | .657 | .537 |
| VirTues | .812 | .760 | .625 | .635 | .781 | .776 | .739 | .724 |
| $\text{VirTues}_{\text{IMC17M}}$ | .786 | .756 | .574 | .624 | .783 | .778 | .714 | .719 |
| $\text{VirTues}_{\text{IMC17M+OP}}$ | .796 | .771 | .624 | .651 | .783 | .771 | .734 | .731 |
| $\texttt{ImmuVis}_{\text{ViT}}$ | .845 | **.848** | .662 | **.695** | **.827** | **.821** | .778 | **.788** |
| $\texttt{ImmuVis}_{\text{Conv}}$ | **.882** | .585 | **.668** | .532 | .822 | .718 | **.791** | .612 |
| $\texttt{ImmuVis}_{\text{Swin}}$ | .850 | .841 | .654 | .691 | .807 | .811 | .770 | .781 |
| $\texttt{ImmuVis}_{\text{VIT+VP}}$ | .847 | .840 | .649 | .686 | .793 | .799 | .763 | .775 |
| $\texttt{ImmuVis}_{\text{Conv+VP}}$ | .752 | .714 | .617 | .599 | .752 | .702 | .707 | .672 |

does not use the per-marker z-score normalization described for VP. Features are standardized within each training fold and classified using logistic regression with stratified 10-fold cross-validation.

**Statistical testing.** We assess differences between `ImmuVis` and VirTues using paired Wilcoxon signed-rank tests on per-fold Macro-F1 scores with Benjamini–Hochberg FDR correction.

**Extended results.** Table 16 extends Table 4 with additional preprocessing configurations and ABMIL results. For `ImmuVis` variants, logistic regression achieves higher average Macro-F1 than ABMIL. Among the models in the main table, $\texttt{ImmuVis}_{\text{Swin}}$ achieves the highest displayed means for relapse and Cords grade, and shares the highest ERBB2 mean with $\texttt{ImmuVis}_{\text{Conv}}$. $\texttt{ImmuVis}_{\text{Conv}}$ and $\texttt{ImmuVis}_{\text{ViT}}$ share the highest ER mean. VirTues achieves the highest Danenberg grade mean and shares the highest Cords subtype mean with $\texttt{ImmuVis}_{\text{Swin}}$.

These endpoints are constrained by the information in a small IMC tissue region: relapse depends on follow-up and systemic disease, while grade and subtype are assessed from wider tissue material. Small clinical cohorts also limit the precision of between-model comparisons.

### A.19 Locality analysis

In order to assess the locality of different models, we introduced a saliency-based *locality index*. This index was computed for a leave-one-out virtual staining task (performed on the IMMUcan HN, Cords and Danenberg cohorts). For each image and each marker, we computed a saliency map as the

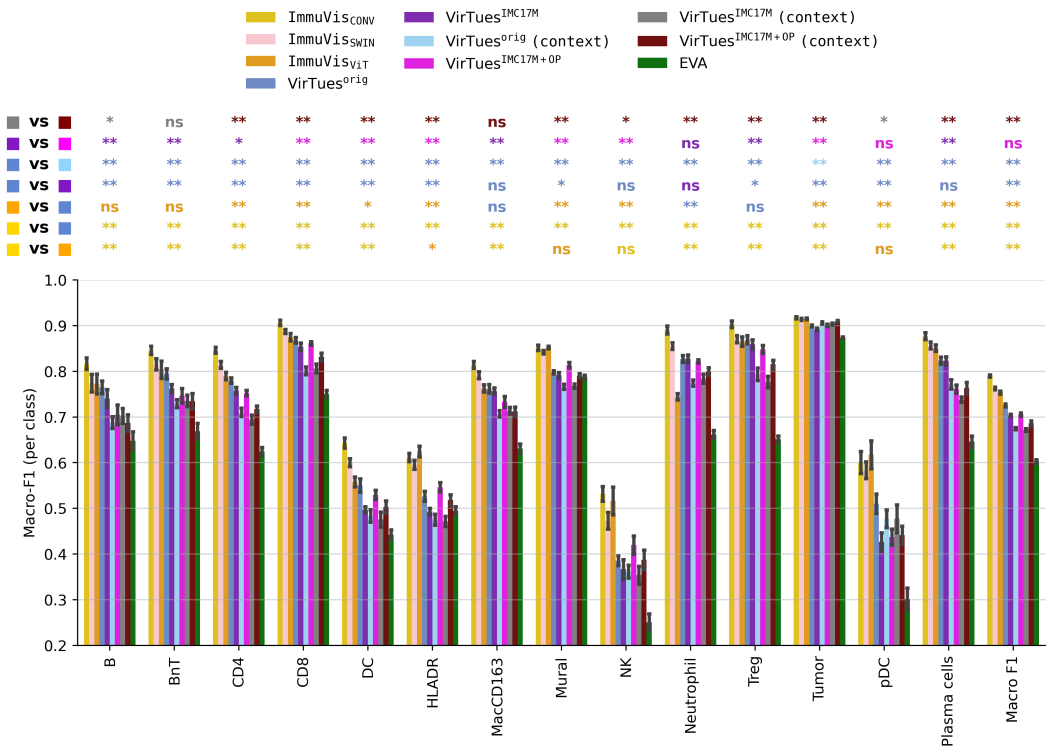



Figure 16: **Cell-typing performance from learned representations on the [46] dataset.** Bars show mean marco-F1 across cross-validation folds with confidence intervals (whiskers); higher is better. The seven rows above the plot report paired Wilcoxon signed-rank tests on per-fold scores for the corresponding model pairs (indicated by the colored squares on the left). Significance markers are color-coded to match the model with the higher mean score. Significance is shown after FDR correction (ns - not significant; $(*) < 0.05$; $(**) < 0.01$; $(***) < 0.001$).

element-wise absolute value of the gradient of the predicted marker expression at the central pixel. We then normalized this gradient to sum to 1 across all input marker expression images, yielding a probability distribution. For this distribution, we defined the locality index as the average $L_\infty$ distance from the central pixel.

Figure 19 shows the locality-index distributions on HN, while Table 17 reports mean locality radii across HN, Cords, and Danenberg. $\texttt{ImmuVis}_{\texttt{Conv}}$ achieved the lowest average locality index across images and markers, outperforming $\texttt{ImmuVis}_{\texttt{Swin}}$ and $\texttt{ImmuVis}_{\texttt{ViT}}$. VirTues exhibited a substantially higher locality index.

$\texttt{ImmuVis}_{\texttt{Conv}}$ and $\texttt{ImmuVis}_{\texttt{Swin}}$ fuse markers after the stem, whereas the $\texttt{ImmuVis}_{\texttt{ViT}}$ hyperkernel acts as the $8 \times 8$ tokenizer; token-based alternatives add marker identity after tokenization. The replication supports the separation from VirTues, but differences within the ImmuVis family are small and the index does not explain their task-performance ordering.

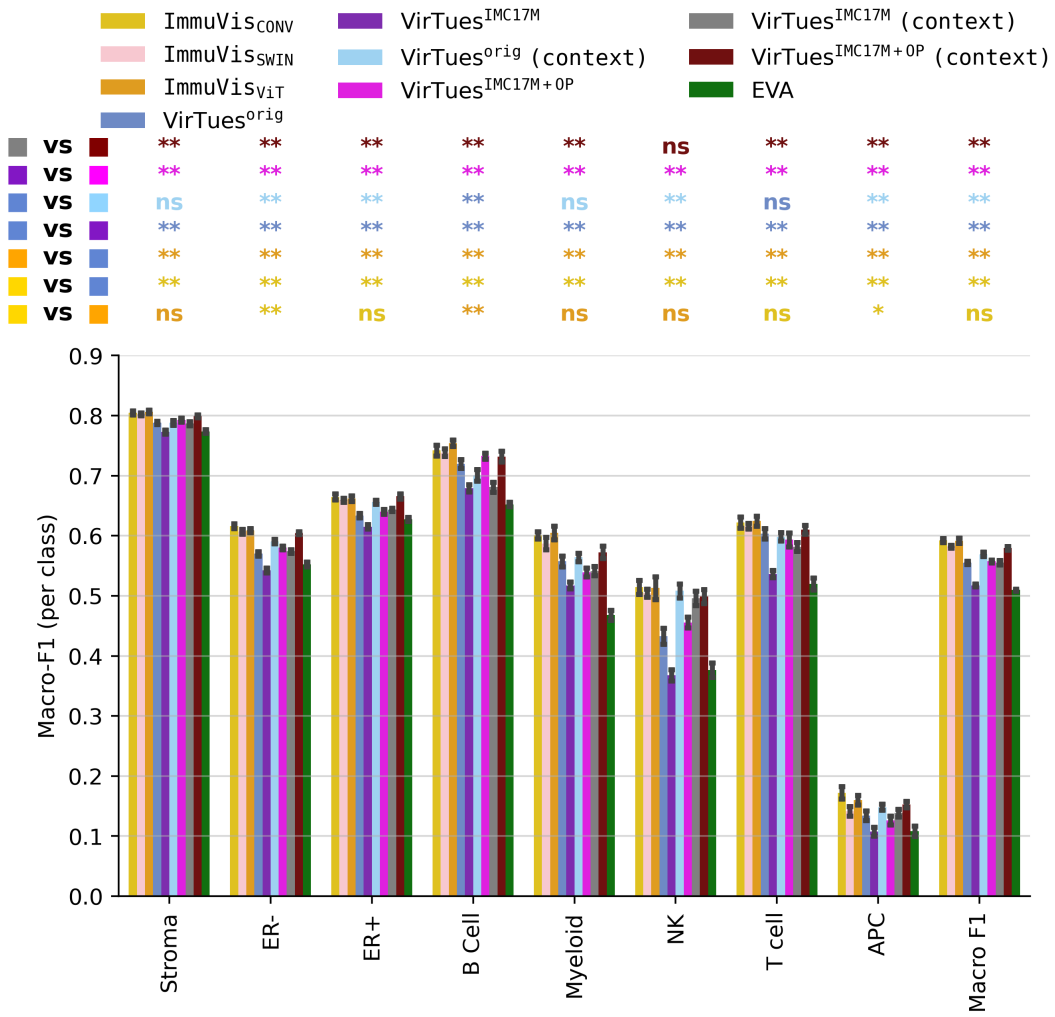



Figure 17: **Cell-typing performance from learned representations on the [47] dataset.** Bars show mean marco-F1 across cross-validation folds with confidence intervals (whiskers); higher is better. The seven rows above the plot report paired Wilcoxon signed-rank tests on per-fold scores for the corresponding model pairs (indicated by the colored squares on the left). Significance markers are color-coded to match the model with the higher mean score. Significance is shown after FDR correction (ns - not significant; $(^{*}) < 0.05$; $(^{**}) < 0.01$; $(^{***}) < 0.001$).

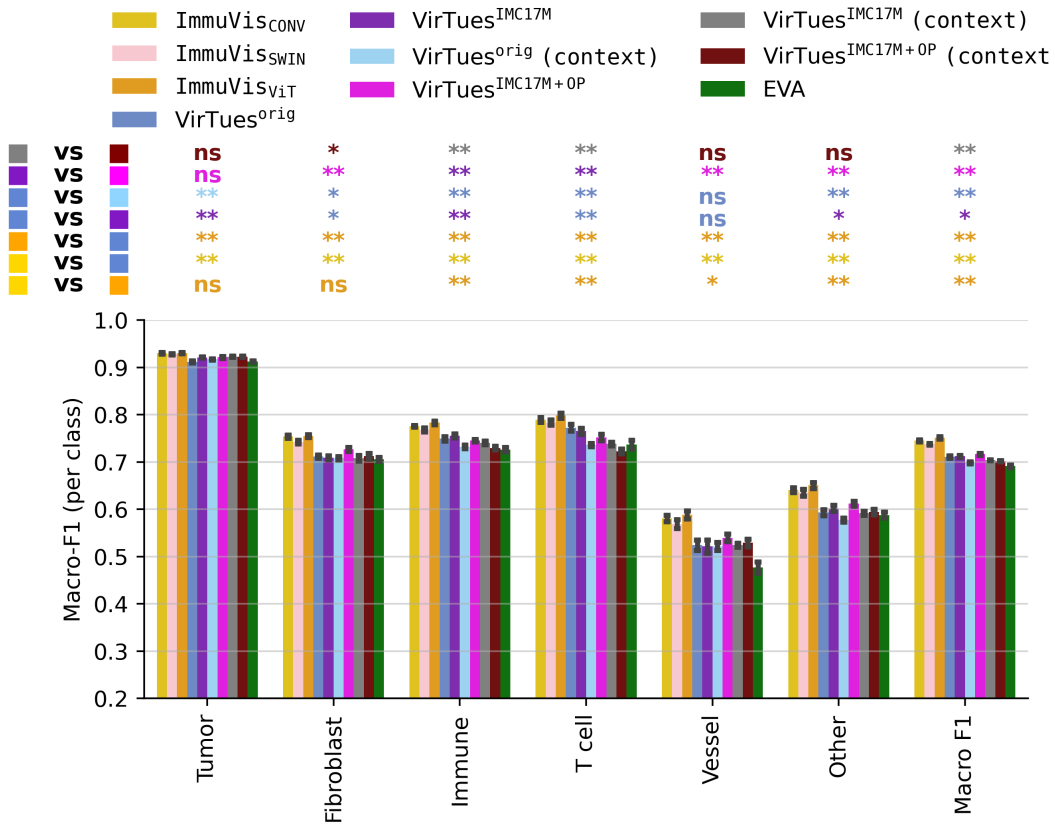



Figure 18: **Cell-typing performance from learned representations on the [48] dataset.** Bars show mean marco-F1 across cross-validation folds with confidence intervals (whiskers); higher is better. The seven rows above the plot report paired Wilcoxon signed-rank tests on per-fold scores for the corresponding model pairs (indicated by the colored squares on the left). Significance markers are color-coded to match the model with the higher mean score. Significance is shown after FDR correction (ns - not significant; $(*) < 0.05$; $(**) < 0.01$; $(***) < 0.001$).

Table 16: Representation learning for clinical endpoints (Macro-F1; higher is better). Values report mean±std Macro-F1 across 10-fold cross-validation; best performance per endpoint is shown in bold.

| Model | Probe | Danenberg et al. [47] | | | | Cords et al. [48] | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PAM50 | Grade | ER | ERBB2 | Subtype | Relapse | Grade | Avg. |
| $\text{EVA}_{\texttt{OP}}$ | LogReg | .38±.05 | .50±.08 | .75±.07 | .61±.08 | .79±.03 | .57±.05 | .50±.05 | .585 |
| | abMIL | .40±.03 | .45±.03 | .80±.07 | .81±.15 | .81±.02 | .55±.05 | .48±.04 | .614 |
| VirTues | LogReg | .37±.05 | **.53±.05** | .78±.08 | .76±.11 | .82±.03 | .57±.03 | .51±.07 | .620 |
| | abMIL | .42±.06 | .48±.04 | .77±.09 | .66±.11 | .83±.02 | .58±0.03 | .48±.03 | .603 |
| $\text{VirTues}_{\texttt{IMC17M}}$ | LogReg | .36±.06 | .49±.05 | .77±.06 | .78±.13 | **.83±.02** | .58±.04 | .54±.06 | .621 |
| | abMIL | **.43±.04** | .52±.06 | .76±.09 | .74±.10 | **.83±.02** | .56±.03 | .52±.05 | .623 |
| $\text{VirTues}_{\texttt{IMC17M+OP}}$ | LogReg | .40±.08 | .49±.08 | .80±.06 | .77±.13 | .83±.03 | .57±.05 | .53±.09 | .627 |
| | abMIL | .41±.04 | .46±.05 | .78±.08 | .79±.15 | .80±.03 | .54±.04 | .47±0.06 | .607 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | LogReg | .40±.06 | .48±.05 | **.81±.04** | .77±.12 | .82±.03 | .58±.04 | .54±.06 | .628 |
| | abMIL | **.43±.04** | .46±.02 | .76±.08 | .76±.15 | .81±.02 | .56±.03 | .47±.04 | .608 |
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | LogReg | .40±.09 | .52±.08 | .81±.07 | .80±.13 | .81±.03 | .59±.03 | .54±.05 | .638 |
| | abMIL | .42±.05 | .45±.01 | .78±.08 | .79±.16 | .82±.02 | .58±.04 | .48±.04 | .618 |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | LogReg | .42±.09 | .49±.06 | .79±.05 | .80±.13 | .83±.03 | **.61±.04** | **.55±.06** | **.641** |
| | abMIL | **.43±.04** | .47±.04 | .78±.07 | .80±.16 | **.83±.02** | .57±.04 | .50±.04 | .626 |
| $\texttt{ImmuVis}_{\texttt{ViT+VP}}$ | LogReg | .39±.11 | .49±.10 | .78±.06 | .79±.10 | .83±.02 | .59±.02 | .51±.04 | .626 |
| | abMIL | .43±.07 | .46±.04 | .74±.06 | .81±.10 | .80±.03 | .56±.04 | .46±.06 | .609 |
| $\texttt{ImmuVis}_{\texttt{Conv+VP}}$ | LogReg | .41±.09 | .49±.06 | .80±.08 | .78±.09 | .81±.03 | .57±.03 | .51±.06 | .624 |
| | abMIL | .41±.08 | .44±.06 | .73±.05 | **.82±.07** | .79±.03 | .55±.06 | .45±.04 | .600 |

Table 17: Mean saliency locality radius on three cohorts; smaller values indicate more local influence.

| Model | Immucan H&N | Cords | Danenberg |
|---|---|---|---|
| $\texttt{ImmuVis}_{\texttt{Conv}}$ | 26.0 | 26.1 | 24.1 |
| $\texttt{ImmuVis}_{\texttt{Swin}}$ | 28.5 | 26.6 | 25.8 |
| $\texttt{ImmuVis}_{\texttt{ViT}}$ | 29.9 | 29.8 | 29.1 |
| VirTues | 36.5 | 37.9 | 37.8 |

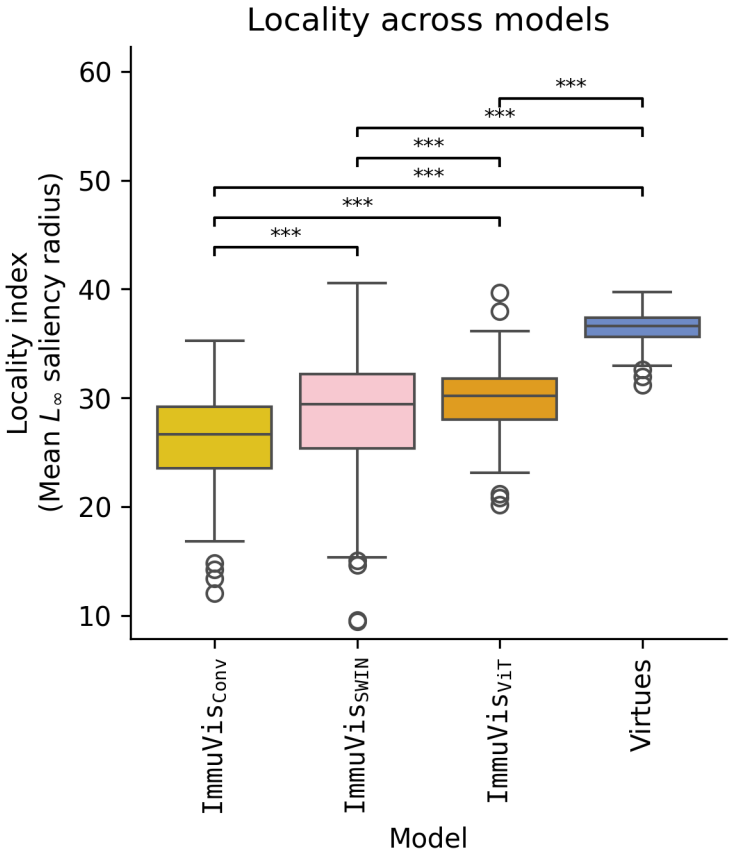



Figure 19: **Locality index analysis on HN.** Distributions of the mean $L_\infty$ saliency radius for `ImmuVis` variants and VirTues. Smaller values indicate more spatially local influence. Pairwise comparisons use paired Wilcoxon signed-rank tests (***: $p < 0.001$).

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's contributions, including the marker-adaptive `ImmuVis` model family, the IMC17Mdataset, uncertainty-aware virtual staining, and downstream evaluations. The claims are supported by the reported experiments and are framed within the intended IMC/spatial proteomics setting without overstating generalization beyond the evaluated data.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discussion section includes dedicated paragraph that mentions current limitations of the `ImmuVis` family as well as the future work paragraph that addresses what is not yet covered by the proposed work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Across the entire paper all mathematical/theoretical claims are justified in-place or the reader is redirected to a dedicated Appendix section covering the proofs supporting any presented claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the architecture, preprocessing, training setup, dataset provenance, splitting strategy, baseline checkpoints, and evaluation protocols required to reproduce the main results. Public datasets are cited to their original sources, and restricted consortium-controlled data are documented with their access constraints. The codebase is provided in anonymized form.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides an anonymized repository with code and instructions for reproducing the main experiments, including environment setup, preprocessing, training, and evaluation scripts. Public datasets are cited with access instructions, while consortium-controlled IMMUcan data cannot be redistributed and are used under the applicable data-governance terms.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main experimental settings in the core text, including dataset splits, held-out cohorts, preprocessing protocols, evaluation tasks, and baseline comparisons. Detailed training configurations, model architectures, hyperparameters, optimizer settings, downstream evaluation procedures, and additional reproducibility details are provided in the appendices and accompanying repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Main quantitative results include appropriate uncertainty or significance reporting, including cross-validation mean±standard deviation and paired statistical tests with FDR correction where applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Inference-time resource usage is reported in Table 5; training and evaluation compute are detailed in the Training compute resources paragraph of Appendix A.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Discussion section (Future overlook paragraph) acknowledges that if IMC adoption extends from research to clinical settings, `ImmuVis` could contribute to diagnostic pipelines, and notes the role of calibrated uncertainty as a reliability safeguard. No direct negative societal impacts are foreseen for this biomedical imaging method.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with high misuse risk in the sense of the checklist. The model does not generate natural images, text, or other content categories typically associated with dual-use or unsafe release concerns.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, model checkpoints, and code assets used in the paper are credited through citations and documented in the appendix. Public datasets are referenced to their source publications, while consortium-controlled IMMUcan data are used under the applicable data-governance terms and are not redistributed. Baseline models such as VirTues and EVA are cited and used from the authors' released resources where available. Released derived assets will document provenance, access conditions, and usage terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new model assets and documents the curated IMC17M resource. We provide documentation of the model architecture, preprocessing, training, evaluation protocols, and limitations in the main paper and appendix. IMC17M includes public datasets and consortium-controlled data; therefore, protected data are not redistributed (until data embargo holds), and access is governed by the corresponding original licenses, approvals, and data-use agreements. The code is available at `https://anonymous.4open.science/r/ImmuVis-9173`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing, participant-facing experiments, or collection of new data from human subjects. The study uses existing IMC datasets and consortium-controlled data; no instructions were given to participants and no compensation was provided by the authors.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper does not collect new human-subject data or interact with participants. It analyzes existing IMC datasets with clinical annotations: public datasets are used under their original data-use terms and approvals, and IMMUcan consortium data are used under the consortium's controlled data-governance framework. No additional participant-facing procedures were performed by the authors.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.