AI-POWERED VIRTUAL TISSUES FROM SPATIAL PROTEOMICS FOR CLINICAL DIAGNOSTICS AND BIOMEDICAL DISCOVERY

Johann Wenckstern[*,1] **Eeshaan Jain**[*,1] **Kiril Vasilev**[2] **Matteo Pariset**[3]
Andreas Wicki[4] **Gabriele Gut**[4] **Charlotte Bunne**[1]
[1] EPFL, [2] ETH Zurich, [3] Straintest, [4] USZ
{johann.wenckstern,eeshaan.jain,charlotte.bunne}@epfl.ch
kvasilev@student.ethz.ch
matteopariset@gmail.com
{andreas.wicki,gabriele.gut}@usz.ch

ABSTRACT

Spatial proteomics technologies have transformed our understanding of complex tissue architectures by enabling simultaneous analysis of multiple molecular markers and their spatial organization. The high dimensionality of these data, varying marker combinations across experiments and heterogeneous study designs pose unique challenges for computational analysis. Here, we present Virtual Tissues (VirTues), a foundation model framework for biological tissues that operates across the molecular, cellular and tissue scale. VirTues introduces innovations in transformer architecture design, including a novel tokenization scheme that captures both spatial and marker dimensions, and attention mechanisms that scale to high-dimensional multiplex data while maintaining interpretability. Trained on diverse cancer and non-cancer tissue datasets, VirTues demonstrates strong generalization capabilities without task-specific fine-tuning, enabling cross-study analysis and novel marker integration. As a generalist model, VirTues outperforms existing approaches across clinical diagnostics, biological discovery and patient case retrieval tasks, while providing insights into tissue function and disease mechanisms.

1 INTRODUCTION

Tissues, particularly in cancer, display pronounced heterogeneity across patients, disease stages and even within individual tumors—evident in diverse cell phenotypes, states and spatial organization (de Souza et al., 2024). Accounting for this heterogeneity is critical: tumor development and response to therapy depend not just on cancer cells, but on their complex interactions with their surrounding environment. Different cell phenotypes may act as promoters or suppressors of tumor development and progression, depending on the biological context (Hanahan & Weinberg, 2011; Hanahan, 2022), with their spatial co-occurrence patterns predicting immunotherapy response (Wang et al., 2023; Phillips et al., 2021), disease relapse (Radtke et al., 2024) and survival (Sorin et al., 2023). Understanding the spatial organization, composition and function of the tumor microenvironment (TME) has thus emerged as an important element for advancing cancer treatment.

Capturing complex tissue structure requires advanced molecular imaging techniques that go beyond traditional methods (Lin et al., 2023a). The emergence of multiplexed imaging technologies—including co-detection by indexing (CODEX) (Black et al., 2021), imaging mass cytometry (IMC) (Giesen et al., 2014) and multiplex immunohistochemistry (IHC) (Lewis et al., 2021)—has revolutionized our ability to study the TME by enabling *in situ* detection of multiple markers on a single slide (Nordmann et al., 2024). IMC, in particular, can measure up to 150 target proteins at cellular and sub-cellular scales, serving as a crucial tool for precision oncology (Danenberg et al., 2022).

Advancement in computational tools of multiplexed tissue imaging data from patient tumors may provide three critical functions in clinical oncology: **clinical diagnostics**, e.g., the identification

and characterization of tumors, **biological discovery**, e.g., understanding disease mechanisms and therapy responses, and clinical decision support through **retrieval** of comparable patient cases for molecular tumor boards (Tsimberidou et al., 2023). The complexity of multiplexed data, characterized by immense scale and highly non-linear, context-dependent relationships between molecular markers, necessitates artificial intelligence (AI) methods for meaningful pattern interpretation and prediction (Rigamonti et al., 2024).

To achieve this, we present Virtual Tissues, an AI-driven foundation model framework for representing biological tissues from spatial proteomics data. Trained on diverse datasets covering breast (Danenberg et al., 2022; Jackson et al., 2020), lung (Cords et al., 2023) and melanoma (Hoch et al., 2022) tissues and evaluated via zero-shot inference on diabetic pancreatic tissue (Damond et al., 2019), VirTues achieves strong performance across various clinical and biological tasks without task-specific training. Importantly, it can generate virtual tissue representations of new cancer types or diseases without fine-tuning, while maintaining robustness to dataset-specific artifacts.

Recent foundation models in digital pathology (Chen et al., 2024; Xu et al., 2024; Wang et al., 2024) and cellular microscopy (Kraus et al., 2024; Kenyon-Dean et al., 2024; Gupta et al., 2024; Bao et al., 2023) (such as CellPainting data (Bray et al., 2016)) have established vision transformers (ViT) (Dosovitskiy et al., 2021) as the predominant architecture for analyzing spatial biomedical data. However, neither these transformer-based approaches nor CNN-based methods (Sorin et al., 2023) adequately address the unique challenges inherent in multiplex imaging data: First, multiplex images are high-dimensional—instead of the three-dimensional RGB format of H&E slices, these images often comprise more than 40 channels of intensity values indicating the presence of different protein or RNA molecules. Second, every experiment may record a *different* set of markers, requiring architectures that scale and deal with flexible inputs of potentially never-measured markers and their distribution within cells and tissues. Lastly, multiplex tissue imaging remains costly, resulting in fewer, often smaller datasets. The relative data scarcity necessitates models that can robustly generalize across diverse study contexts and incorporate future patient cohorts while maintaining reliable performance.

Following the vision of constructing multi-scale foundation models for biology (Bunne et al., 2024), VirTues provides new innovations in AI architecture development: Its success stems from its transformer architecture that respects biological hierarchy across multiple scales, providing representations and insights into tissue function. At the molecular level, it is powered by protein language models (PLM), enabling it to capture complex relationships between protein markers, distinguish the semantic meanings of different markers, and generalize to previously unseen markers.

VirTues is a *generalist* model: Contrary to existing tools, VirTues is the *first* model able to tackle a wide range of prediction and retrieval tasks across cell, niche and tissue levels, while consistently outperforming all baselines and demonstrating *emergent capabilities* for embedding unseen markers, patients and diseases.

2 METHOD

2.1 VIRTUES ARCHITECTURE

To address these unique challenges posed by multiplex imaging data, we propose VirTues, an encoder-decoder model based on the ViT architecture. VirTues is designed for the efficient processing of highly-multiplexed image data accommodating varying numbers and combinations of measured markers. Furthermore, VirTues incorporates the attribution of distinct biological meaning to each measured marker. VirTues operates on tokenized image crops of size $d_c \times d_c = 128 \times 128$, capturing tissue niches. Restricting VirTues' input to such crops increases the number and diversity of pretraining samples while decreasing the dimensionality per sample.

Tokenization. To preserve the biologically distinct meaning of each channel and allow for a flexible number of channels per image, we employ a multi-channel tokenization procedure (Kraus et al., 2024; Bao et al., 2023). Each channel is spatially divided into patches of size $d_p \times d_p = 8 \times 8$, since this approximately captures one cell per patch. Flattening each patch results in a three-dimensional grid of image tokens $\boldsymbol{x} \in \mathbb{R}^{M \times H \times W \times d_p^2}$, where M is the number of measured markers and $H = W = d_c/d_p$ the grid height resp. width. For all M markers, we retrieve from a precom-

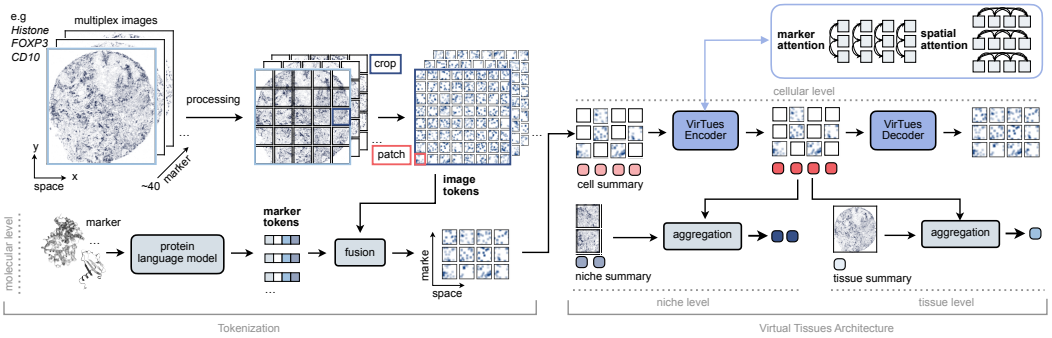

Figure 1: Overview of the VirTues tokenization procedure and architecture.

puted lookup table the corresponding protein embeddings $\pi \in \mathbb{R}^{M \times d_{\text{PLM}}}$ given by the protein language model ESM-2 (Lin et al., 2023b) with $d_{\text{PLM}} = 640$. We refer to these embeddings as marker tokens. For each channel m and each grid position (i, j), we project the image token $\boldsymbol{x}_{mij}$ and the corresponding marker token π_m to the same dimension d_{model} using learnable linear projections, to get $\boldsymbol{x}'_{mij} \in \mathbb{R}^{d_{\text{model}}}$ and $\pi'_m \in \mathbb{R}^{d_{\text{model}}}$ respectively. The image and marker tokens are fused through summation, resulting in the image tokens $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{M \times H \times W \times d_{\text{model}}}$, where $\widetilde{\boldsymbol{x}}_{mij} = \boldsymbol{x}'_{mij} + \pi'_m$. The fusion of the marker token with the image tokens serves two main purposes: (1) enabling VirTues to differentiate the channel origins of input tokens, and (2) introducing an inductive bias regarding the marker function, which cannot be added through other marker tokenization schemes (such as one-hot or learnable marker embeddings). We note that this is the first of many building blocks enabling VirTues to generalize across unseen markers. Further, to allow VirTues to capture an aggregated representation for each patch, we introduce an additional layer of learnable cell summary tokens $\boldsymbol{c} \in \mathbb{R}^{H \times W \times d_{\text{model}}}$, one for each spatial position. Each cell summary token $\boldsymbol{c}_{ij} \in \mathbb{R}^{d_{\text{model}}}$ is initialized using the same weights.

Masking. During training, a portion of the image tokens $\{\widetilde{\boldsymbol{x}}_{mij}\}$ is masked by replacing them with a special masking token $\square \in \mathbb{R}^{d_{\text{model}}}$ initialized with learnable weights. Masking is applied channel-wise by sampling a masking ratio r_{masking} between 60% and 100% and uniformly selecting the corresponding $\lceil r_{\text{masking}} HW \rceil$ tokens to mask within the channel. We denote the resulting 3D binary mask by $\boldsymbol{M} \in \{0, 1\}^{M \times H \times W}$, where the value 1 marks masking. Masked tokens remain linked to their specific markers, which is indicated by adding the marker tokens to the masked tokens.

VirTues Encoder. The set of all non-masked image tokens $\{\widetilde{\boldsymbol{x}}_{mij} \mid \boldsymbol{M}_{mij} = 0\}$ and the set of cell summary tokens $\{\boldsymbol{c}_{ij}\}$ is passed as an input to the VirTues Encoder. This encoder is constructed by modifying the vision transformer's architecture (Dosovitskiy et al., 2021), to adapt it to work with varying input channels efficiently, and capture marker correlations and spatial patterns separately. In contrast to standard Vision Transformers, which use full multi-head self-attention where all tokens attend pairwise to each other (Fig. 1e), we use two specialized sparse multi-head self-attention mechanisms, marker attention and channel attention, (Fig. 1d) akin to space and time attention employed in video transformers (Bertasius et al., 2021). In marker attention, only tokens which are placed at the same spatial grid position attend to each other, thereby capturing inter-marker dependencies and correlations. We denote the set of input tokens to the ℓ-th transformer block as $\{t^{\ell}_{mij}\}$, where the token t^{ℓ}_{mij} is associated to the m-th channel and position (i, j). In this notation, we treat the layer of cell summary tokens simply as a further channel. Then, a marker attention transformer block computes

$$\forall i^*, j^* : \left\{t^{\ell+1}_{mij} \;\middle|\; \substack{i=i^* \\ j=j^*}\right\} = \text{MHSA}\left(\left\{t^{\ell}_{mij} \;\middle|\; \substack{i=i^* \\ j=j^*}\right\}\right).$$

where MHSA denotes a transformer block with standard multi-head self-attention. In contrast, in channel attention, only tokens present in the same channel attend to each other, hence capturing spatial patterns across tissue. Following the notation for marker attention, a channel attention transformer block computes as

$$\forall m^* : \left\{t^{\ell+1}_{mij} \mid m = m^*\right\} = \text{MHSA}\left(\left\{t^{\ell}_{mij} \mid m = m^*\right\}\right).$$

VirTues Encoder outputs a set of encoded image tokens $\left\{\widetilde{\boldsymbol{x}}^{\text{enc}}_{mij} \mid \boldsymbol{M}_{mij} = 0\right\}$ and a set of encoded cell summary tokens $\left\{\boldsymbol{c}^{\text{enc}}_{ij}\right\}$.

VirTues Decoder. The VirTues Decoder is used during training and inference to reconstruct the original image. It is comprised of a Vision Transformer (Dosovitskiy et al., 2021) followed by a single linear projection. To reconstruct the original image, the encoded tokens $\left\{\widetilde{\boldsymbol{x}}^{\text{enc}}_{mij} \mid \boldsymbol{M}_{mij} = 0\right\}$, the encoded cell summary tokens $\left\{\boldsymbol{c}^{\text{enc}}_{ij}\right\}$ and the masked tokens $\{\widetilde{\boldsymbol{x}}_{mij} \mid \boldsymbol{M}_{mij} = 1\}$ are regrouped as follows: For each channel m^*, we group the encoded and the masked tokens of that channel with a copy of the encoded cell summary tokens:

$$\left\{\widetilde{\boldsymbol{x}}^{\text{enc}}_{m^*ij} \mid \boldsymbol{M}_{m^*ij} = 0\right\} \cup \{\widetilde{\boldsymbol{x}}_{m^*ij} \mid \boldsymbol{M}_{m^*ij} = 1\} \cup \left\{\boldsymbol{c}^{\text{enc}}_{ij}\right\}.$$

These groups of tokens are passed individually to the decoder one by one. Hence, in the decoder tokens of different channels do not interact with each other. This design is intended to force the decoder to extract the majority of the information to reconstruct each channel from the cell summary tokens rather than relying on other channels and thus incentivize the encoder to store a meaningful representation in these. This regrouping further limits the individual token set sizes to $2HW$, thus allowing us to use full multi-head self-attention instead of marker and channel attention. Processing all tokens by the decoder's transformer followed by the linear projection yields the final grid of reconstructed image tokens $\boldsymbol{x}^{\text{rec}} \in \mathbb{R}^{M \times H \times W \times d_p^2}$. For details on hyperparameters and the implementation, see the appendix.

Aggregation into niche and tissue level representations. During inference, VirTues Encoder represents each image crop as a grid of cell summary tokens $\boldsymbol{c}^{\text{enc}} \in \mathbb{R}^{H \times W \times d_{\text{model}}}$. To process an entire tissue, the multiplexed image is divided into non-overlapping crops, each representing a niche, which are then embedded individually. Niche or tissue level representations are obtained by aggregating all encoded cell summary tokens from the crop or the full image, respectively. For unsupervised tasks, such as the retrieval experiments in Suppl. Fig. 4, a simple average $z = \frac{1}{HW}\sum_{i,j} \boldsymbol{c}^{\text{enc}}_{ij}$ is used for aggregation, generating task-agnostic embeddings. In supervised settings, a dynamically weighted average is employed, achieved by training an attention-based multiple instance learning classifier (Ilse et al., 2018) on the given task (while the parameters of VirTues' are kept frozen), generating task-specific embeddings. This attention weighted average computes as

$$z = \sum_{i,j} a_{ij} \boldsymbol{c}^{\text{enc}}_{ij} \quad \text{with} \quad a_{ij} = \frac{\exp w^T(\tanh\left(V\boldsymbol{c}^{\text{enc}}_{ij}\right) \odot \sigma(U\boldsymbol{c}^{\text{enc}}_{ij}))}{\sum_{i'j'} \exp w^T(\tanh\left(V\boldsymbol{c}^{\text{enc}}_{i'j'}\right) \odot \sigma(U\boldsymbol{c}^{\text{enc}}_{i'j'}))},$$

where $U, V \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{model}}}$ and $w \in \mathbb{R}^{d_{\text{hidden}}}$ are learnable weights, σ is the sigmoid activation function and $\odot$ indicates element-wise multiplication. In a multi-head setting, this computation is repeated for each head with a different weight vector $w^h \in \mathbb{R}^{d_{\text{hidden}}}$ and the resulting representations are concatenated. Per default, we use 4 heads.

2.2 VIRTUES PRETRAINING

Loss function. VirTues is trained end-to-end to reconstruct image crops in a masked autoencoding framework (He et al., 2022; Vincent et al., 2008). Our reconstruction loss is the mean squared error between the reconstructed pixels' intensity values and the original pixels' intensity values, i.e., $\mathcal{L}_{\text{MAE}} = \frac{1}{d_c^2 M}\left\|\boldsymbol{x}^{\text{rec}} - \boldsymbol{x}\right\|_2^2$. Note that this loss is computed over all pixels of both masked and non-masked tokens.

Datasets. For training VirTues, we curated four publicly available image mass cytometry datasets, each mapping the spatial organization of tumor microenvironments across various cancer types and tissue sites. The Cords et al. (2023) dataset contains samples of non-small cell lung cancer, Jackson et al. (2020) and Danenberg et al. (2022) focus on breast cancer tissues, and Hoch et al. (2022) examines primary and metastatic melanoma tissues. Images smaller than 256 $\times$ 256 pixels were excluded, resulting in a total of 3,473 distinct images. An overview of the measured marker panels and the individual dataset sizes is shown in Suppl. Fig. S1. Each dataset was split into an 80%-20% train-test partition, grouped by patient identities. For each dataset, we computed the ESM-2 (Lin et al., 2023b) embeddings of the measured markers. For mRNA markers, we used the sequences of the encoded proteins. For additional details on data preprocessing and augmentation, see Appendix.

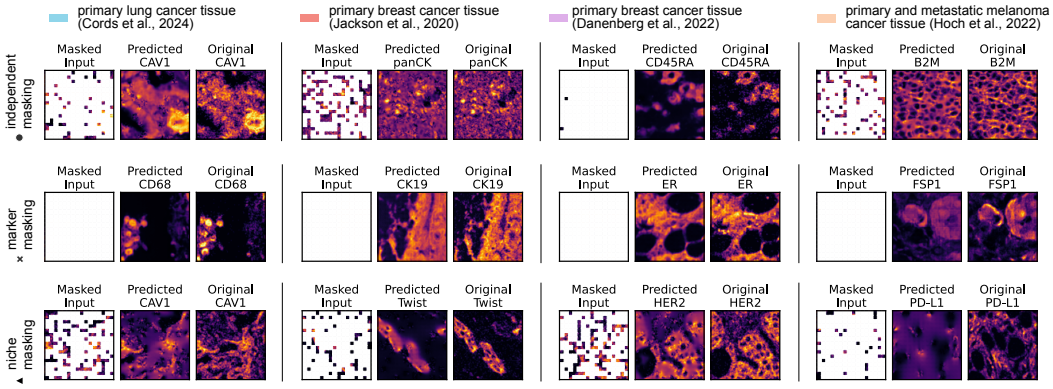

Figure 2: Examples of images reconstructed by VirTues for independent, marker and niche masking.

2.3 VirTues Database and Retrieval

We construct the Virtual Tissue database by extracting from each image the central 4×4 grid of crops with each crop sized 128×128, excluding those smaller than this grid. The remaining crops are embedded using VirTues or one of the baselines and the resulting niche-level representations are stored in the database. For VirTues and CA-MAE (Kraus et al., 2024), we use self-supervised niche-level representations, computed as the average of cell-level embeddings.

The database is used to retrieve tissues similar to a given reference image, measured using the 2-Wasserstein distance between sets of niche-level representations. Given the niche-level representations $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{N \times d_{\text{model}}}$ for two tissues, the Wasserstein distance computes as

$$W_2(\boldsymbol{a}, \boldsymbol{b}) = \left(\min_{\pi \in \Gamma} \sum_{i,j=1}^{N} \pi_{ij} \left\| \boldsymbol{a}_i - \boldsymbol{b}_j \right\|_2^2 \right)^{\frac{1}{2}},$$

where $\Gamma = \left\{ \pi \in \mathbb{R}^{N \times N} \mid \pi \vec{1} = \vec{1}/N \text{ and } \pi^T \vec{1} = \vec{1}/N \right\}$. For a given reference image, we identify the closest matches based on this distance metric.

3 Results

3.1 Masked Autoencoder Reconstructions

We evaluate VirTues' understanding of molecular tissue structure and biological relationships between markers by assessing the reconstruction ability of VirTues for three different masking strategies: independent masking, marker masking and niche masking. For independent masking, akin to the training, we sample a masking ratio $r_m \sim \mathcal{U}([0.6, 1.0])$ independently for each marker m in the input image and mask the corresponding number of patches. For marker masking, we select a single marker from the input image and mask all corresponding patches. In contrast, in niche masking, we sample for each image a single masking ratio $r \sim \mathcal{U}([0.6, 1.0])$ and use this ratio to select the corresponding number of grid positions, where we mask all patches across all channels. Marker masking and niche masking are intended to assess VirTues understanding of spatial patterns and marker correlations in isolation of each other.

In the independent masking experiments, VirTues successfully reconstructs masked regions across markers, e.g., CAV1, panCK, CD45RA and B2M in Fig. 2, preserving both spatial distribution and intensity patterns. This demonstrates the model's ability to leverage contextual information across both spatial and marker dimensions. The marker masking experiments reveal the model learns the correlations between different markers and is thus able to fully inpaint a masked channel, e.g., CD68, CK19, ER or FSP1 in Fig. 2. Since with marker masking the VirTues Decoder receives only the encoded cell summary tokens as input, its ability to decode a channel absent from the encoder's input

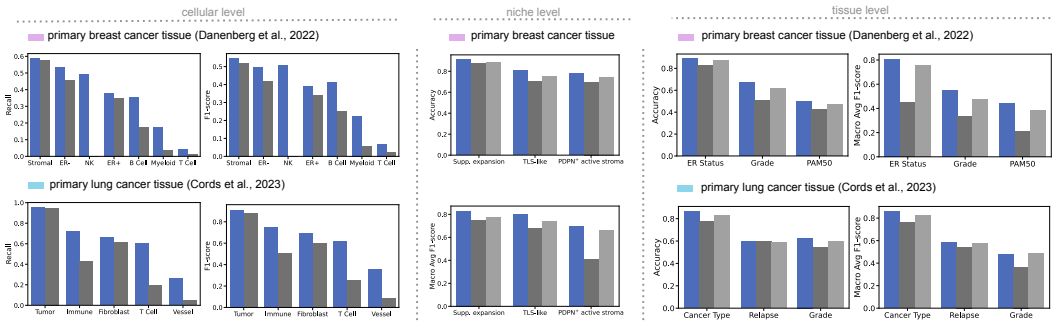

Figure 3: Comparison of prediction performance of VirTues, CA-MAE (Kraus et al., 2024) and ResNet (Sorin et al., 2023) across scales.

further demonstrates that the cell summary tokens generated by VirTues effectively capture a meaningful and robust representation of the molecular tissue architecture. The niche masking experiments (e.g. CAV1, Twist, HER2 and PD-L1 in Fig. 2) show that the model is able recover complex spatial tissue architectures. Quantitative evaluation in terms of the average mean squared error (MSE) computed over all masked token's pixels across different markers, datasets and masking strategies (Suppl. Fig. S2) shows robust reconstruction capabilities, with however, notable differences between the individual markers.

3.2 Biological discovery and clinical diagnostics tasks

VirTues enables comprehensive analysis of tissues across biological scales. At the cellular level, cell summary tokens support tasks such as cell type classification. Niche summary tokens facilitate tissue structure analysis, while tissue summary tokens enable clinical diagnostics like the prediction of cancer type and grade. We thus evaluate these multi-scale representations in the following using cell, niche and tissue level classification tasks.

Cell level classification. Identifying the cell types in a tissue sample offers critical insights into its functional dynamics and cellular makeup. Moreover, it helps in dissecting the complex interactions between different cell types within a TME, which can help guiding systemic therapies geared towards the tumor environment (such as immunotherapy), treatments targeting the tumor cells themselves, and ultimately refine a broad range of systemic anti-cancer treatment regimens. Therefore, we benchmark the performance of VirTues against CA-MAE (Kraus et al., 2024) in identifying the most common cell type within a patch. This evaluation spans two datasets, Cords et al. (2023) as well as Danenberg et al. (2022), and for Cords et al. (2023), is performed at two levels of class granularity. Metdata of the datasets allows the pixel-wise assignment of cell types. We determine patch labels as the mode of all pixel labels within a patch. For the prediction task, we perform linear probing using a logistic regression model with an L-BFGS solver and L2 regularization with coefficient $\lambda = 1.0$. This ensures the evaluation focuses on the quality of the learned representations rather than the complexity or configuration of the classifier.

We report the results in terms of class-wise recall and F1-scores in Fig. 3 and Suppl. Fig. S3. VirTues outperforms the baseline in cell type multi-class classification for both breast cancer (distinguishing stromal, ER+, NK, ER-, B cell, myeloid and T cell populations) and lung cancer tissues (tumor, immune, fibroblast, T cell and vessel cells). Despite significant class imbalance, the model demonstrates particular strength compared to the baseline in identifying rare cell populations, maintaining robust performance across varying cell type frequencies.

Niche level classification Multi-cellular structures present within the niches of the TME often recur across tumors and identifying such structures gives insights into the functional state of the TME providing prognostic value. For instance, Danenberg et al. (2022) identify recurrent multicellular structures within breast tumor microenvironments, which correlate with the hazard ratio. We thus evaluate VirTue's niche level representations by predicting in form of a binary classification task the presence of three such structures, suppressive expansion, tertiary lymphoid structures (TLS)-like

regions, PDPN$^+$ regions, within a niche. For these, we train gated attention-based multiple instance learning (ABMIL) models on the cell summary tokens of VirTues.

We report the results in form of accuracy and macro-average F1-score in Fig. 3 averaged over five seeded runs and bench-marked against CA-MAE (Kraus et al., 2024) and ResNet (Sorin et al., 2023).

Tissue level classification Akin to a pathologist, AI models should evaluate tissue characteristics comprehensively for robust clinical classifications and predictions. We therefore evaluate VirTues' tissue level representations obtained from fine-tuned ABMIL models on critical clinical tasks such as ER status determination, tumor grade classification and PAM50 subtyping in breast cancer, as well as the prediction of cancer type, cancer relapse and grade classification in lung cancer.

Results in form of accuracy and macro-average F1-score are shown in Fig. 3.

Further details. For a more detailed analysis of the results of cell, niche and tissue level classification tasks, we refer to the appendix.

3.3 Clinical information retrieval

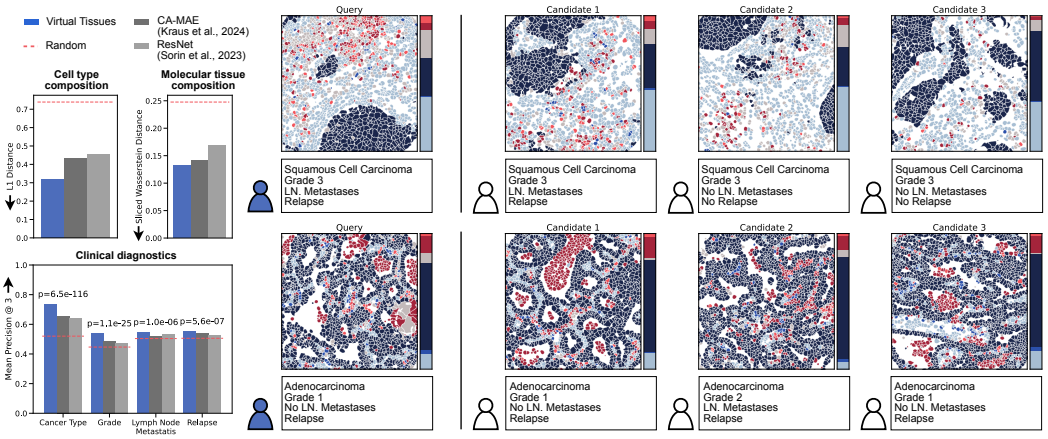

Figure 4: Quantitative and qualitative retrieval results for VirTues database built from Cords et al. (2023) dataset. Each retrieved case includes a visualization of the cell type distribution and clinical annotations, enabling direct comparison of tissue architecture and outcomes.

We test VirTues database for the Cords et al. (2023) dataset and assess its retrieval results. For this, we consider only images of Cords et al. (2023) associated with the two most frequent cancer types: Adenocarcinoma and Squamous cell carcinoma. Two case examples shown in Fig. 4 demonstrate VirTues' ability to retrieve relevant matches. To quantitatively evaluate the efficacy of the retrieval mechanism with VirTues' embeddings, we compare it against the two baseline embedding methods, CA-MAE (Kraus et al., 2024) and ResNet (Sorin et al., 2023), as well as randomized retrieval. Specifically, we compute two alternative distance metrics between the reference image and its closest retrieved match, then compare their average values to those from randomized retrieval. The first metric evaluates similarity based on cell type composition by calculating the proportion of each coarse cell type within the image and computing the L1-distance between the resulting proportion vectors. The second metric assesses molecular tissue composition by treating each image as a set of pixel vectors and calculating the sliced Wasserstein distance (Bonneel et al., 2015) between these sets. This evaluation (see Fig. 4) demonstrates superior performance in matching both cell type composition and molecular tissue structure compared to existing methods (i.e., ResNet (Sorin et al., 2023), CA-MAE (Kraus et al., 2024)) and random retrievals.

Further, we evaluate the similarity of the retrieved archival patients also based on their associated clinical records. Concretely, we perform a McNemar test for each clinical label to compare the number of hits among the top three results achieved by VirTues with the number of hits for random

retrieval. The model achieves significant improvements in critical clinical diagnostic tasks (see Fig. 4), including cancer type classification (P = 6.5e-116), grade determination (P = 1.1e-25), lymph node metastasis prediction (P = 1.0e-06) and relapse prediction (P = 5.6e-07).

This systematic evaluation of VirTues' retrieval capabilities demonstrates not only its ability to identify similar tissue architectures but also the clinical relevance of these similarities, as evidenced by the concordance of retrieved cases' clinical features and outcomes. Thus, by integrating molecular profiles, tissue architecture and clinical outcomes in a unified retrieval framework, VirTues provides the foundation for a data-driven comparison of tissue phenotypes across patient cohorts.

3.4 ZERO-SHOT CAPABILITIES

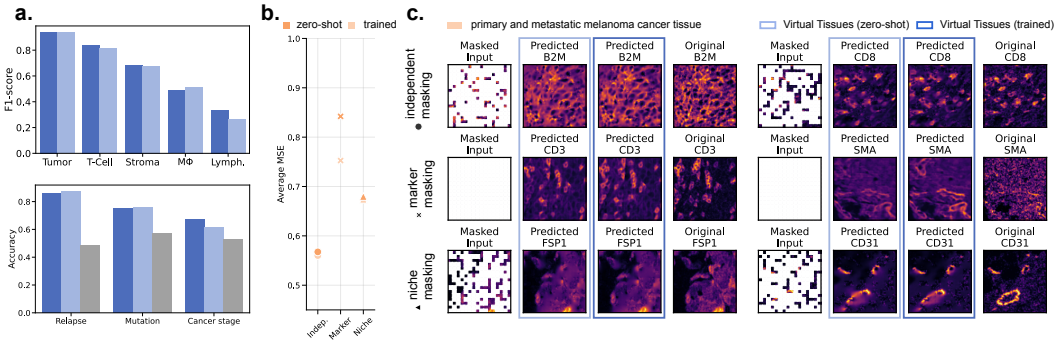

Figure 5: Zeroshot reconstruction, cell type classification and tissue level classification results on Hoch et al. (2022) using markers present during training.

The rapid advancement of precision oncology and the emergence of new molecular disease subtypes require computational methods that can immediately analyze novel cancer types without the time-consuming process of collecting large training datasets and retraining models. This is particularly critical in clinical settings, where prompt analysis of rare cancers or newly characterized disease subtypes can directly impact treatment decisions. We therefore examine VirTues's ability to generalize to a new dataset and cancer type by retraining VirTues only on Cords et al. (2023), Danenberg et al. (2022) and Jackson et al. (2020) and running reconstruction, cell level and tissue level experiments on Hoch et al. (2022). The results of these experiments, presented in Fig. 5 and Suppl. Fig. S4 and analyzed in detail in the appendix, demonstrate emergent zero-shot capabilities of our model.

4 CONCLUSION AND OUTLOOK

The development of the Virtual Tissues platform represents a new development in computational pathology, introducing a foundation model framework that addresses key challenges in analyzing multiplex imaging data while enabling new capabilities for biological discovery and clinical applications. Our results demonstrate that VirTues achieves three critical objectives: universal tissue representation across cancer types, diseases and organs, flexible incorporation of new molecular markers and interpretable multi-scale analysis from molecular to tissue levels. The system also offers clinical decision support by identifying similar cases based on molecular and cellular patterns, though prospective clinical validation would be needed to confirm its utility in clinical decision-making. Current limitations of our study include performance degradation for out-of-distribution markers, and the need for further validation across larger, more diverse datasets and new experimental protocols. Addressing these limitations, expanding to other imaging modalties or molecular data types, improving generalization for rare markers, and integrating generative AI approaches to improve virtual multiplexing are promising directions for future research.

REFERENCES

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel Vision Transformers: An Image Is Worth C x 16 x 16 Words. In *International Conference on Learning Representations*

(ICLR), 2023.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *International Conference on Machine Learning (ICML)*, volume 139, pp. 813–824. PMLR, 2021.

Sarah Black, Darci Phillips, John W Hickey, Julia Kennedy-Darling, Vishal G Venkataraaman, Nikolay Samusik, Yury Goltsev, Christian M Schürch, and Garry P Nolan. Codex multiplexed tissue imaging with dna-conjugated antibodies. *Nature Protocols*, 16(8):3802–3835, 2021.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities. *arXiv Preprint arXiv:2409.11654*, 2024.

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

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):4294, 2023.

Nicolas Damond, Stefanie Engler, Vito RT Zanotelli, Denis Schapiro, Clive H Wasserfall, Irina Kusmartseva, Harry S Nick, Fabrizio Thorel, Pedro L Herrera, Mark A Atkinson, et al. A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metabolism*, 29(3): 755–768, 2019.

Esther Danenberg, Helen Bardwell, Vito RT Zanotelli, Elena Provenzano, Suet-Feung Chin, Oscar M Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O Ellis, et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature Genetics*, 54(5):660–669, 2022.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

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 (ICLR)*, 2021.

Charlotte Giesen, Hao AO Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, 2014.

Ankit Gupta, Zoe Wefers, Konstantin Kahnert, Jan N Hansen, William D Leineweber, Anthony Cesnik, Dan Lu, Ulrika Axelsson, Frederic Ballllosera Navarro, Theofanis Karaletsos, et al. SubCell: Vision foundation models for microscopy capture single-cell biology. *bioRxiv*, 2024.

Douglas Hanahan. Hallmarks of cancer: new dimensions. *Cancer Discovery*, 12(1):31–46, 2022.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5): 646–674, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.

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):eabk1692, 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning (ICML)*, pp. 2127–2136. PMLR, 2018.

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

Kian Kenyon-Dean, Zitong Jerry Wang, John Urbanik, Konstantin Donhauser, Jason Hartford, Saber Saberian, Nil Sahin, Ihab Bendidi, Safiye Celik, Marta Fay, et al. ViTally Consistent: Scaling Biological Representation Learning for Cell Microscopy. *arXiv preprint arXiv:2411.02572*, 2024.

Diederik P Kingma. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11757–11768, 2024.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xFormers: A modular and hackable Transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022.

Sabrina M Lewis, Marie-Liesse Asselin-Labat, Quan Nguyen, Jean Berthelet, Xiao Tan, Verena C Wimmer, Delphine Merino, Kelly L Rogers, and Shalin H Naik. Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods*, 18(9):997–1012, 2021.

Jia-Ren Lin, Yu-An Chen, Daniel Campton, Jeremy Cooper, Shannon Coy, Clarence Yapp, Juliann B Tefft, Erin McCarty, Keith L Ligon, Scott J Rodig, et al. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature Cancer*, 4(7):1036–1052, 2023a.

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, 2023b.

Thierry M. Nordmann, Andreas Mund, and Matthias Mann. A new understanding of tissue biology from MS-based proteomics at single-cell resolution. *Nature Methods*, 21:2220–2222, 2024.

Darci Phillips, Magdalena Matusiak, Belén Rivero Gutierrez, Salil S Bhate, Graham L Barlow, Sizun Jiang, Janos Demeter, Kimberly S Smythe, Robert H Pierce, Steven P Fling, et al. Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. *Nature Communications*, 12(1):6726, 2021.

Andrea J Radtke, Ekaterina Postovalova, Arina Varlamova, Alexander Bagaev, Maria Sorokina, Olga Kudryashova, Mark Meerson, Margarita Polyakova, Ilia Galkin, Viktor Svekolkin, et al. Multi-omic profiling of follicular lymphoma reveals changes in tissue architecture and enhanced stromal remodeling in high-risk patients. *Cancer Cell*, 42(3):444–463, 2024.

Alessandra Rigamonti, Marika Viatore, Rebecca Polidori, Daoud Rahal, Marco Erreni, Maria Rita Fumagalli, Damiano Zanini, Andrea Doni, Anna Rita Putignano, Paola Bossi, et al. 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, 2024.

Mark Sorin, Morteza Rezanejad, Elham Karimi, Benoit Fiset, Lysanne Desharnais, Lucas JM Perus, Simon Milette, Miranda W Yu, Sarah M Maritan, Samuel Doré, et al. Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature*, 614(7948):548–554, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4285–4294, 2023.

Apostolia M Tsimberidou, Michael Kahle, Henry Hiep Vo, Mehmet A Baysal, Amber Johnson, and Funda Meric-Bernstam. Molecular tumour boards—current and future considerations for precision oncology. *Nature Reviews Clinical Oncology*, 20(12):843–863, 2023.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1096–1103, 2008.

Xiao Qian Wang, Esther Danenberg, Chiun-Sheng Huang, Daniel Egle, Maurizio Callari, Begoña Bermejo, Matteo Dugo, Claudio Zamagni, Marc Thill, Anton Anton, et al. Spatial predictors of immunotherapy response in triple-negative breast cancer. *Nature*, 621(7980):868–876, 2023.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, pp. 10524–10533. PMLR, 2020.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pp. 1–8, 2024.

A APPENDIX

A.1 ARCHITECTURE HYPERPARAMETERS AND IMPLEMENTATION DETAILS.

Following He et al. (2022), we setup the encoder-decoder framework in an asymmetric fashion, where the size of the encoder is deeper than the decoder, allowing the major workload of the model to rely on the encoder rather than the decoder. We construct a deep encoder consisting of 16 transformer blocks with alternating marker and channel attention in contrast to the shallow decoder comprised of only 4 transformer blocks with full attention. Both encoder and decoder use 8 attention heads, 2D rotatory position embeddings (Su et al., 2024) to encode spatial positions and pre-layer normalization (Xiong et al., 2020).

To efficiently implement marker and channel attention, we reduce these mechanisms to full attention by merging either the spatial or channel axis of the batched token tensor with the batch axis, allowing subsets of tokens that attend to each other to be treated as independent sequences. During inference without masking, this reduction leverages built-in, hardware-optimized implementations of standard self-attention. However, during training, channel-wise independent masking with varying ratios and channel dropout lead to token sequences in the marker and channel attention blocks having variable lengths. This variability poses a technical challenge because efficient built-in PyTorch attention mechanisms require uniform sequence lengths within a batch. To avoid the computational overhead of adding padding tokens, we employ a dynamic re-packaging strategy in conjunction with xFormers' (Lefaudeux et al., 2022) support for block-diagonal masked self-attention. Non-masked tokens within a batch are repacked into a single sequence, preserving coherent subsequences of tokens that belong to the same sample and channel or spatial position. A block-diagonal mask is generated dynamically to indicate the subsequences, specifying which tokens can attend to each other. The repacked sequence and associated mask are processed using xFormers' masked self-attention implementation.

A.2 ADDITIONAL PRETRAINING DETAILS

Data preprocessing For each image, intensity values are clipped channel-wise at the 99th percentile, followed by a shifted logarithm transformation with a size factor of 1, as commonly applied to scRNA-seq count data (Heumos et al., 2023). Additionally, a Gaussian blur filter with a kernel size of 3 and unit variance is used to smooth each image. Finally, each image is self-standardized channel-wise (Sypetkowski et al., 2023; Kraus et al., 2024).

Data augmentation. Before training, we first randomly sample from each tissue image $4N$ subimages of dimension 256×256, where N is the number of 128×128 crops within the tissue image. During training, each size 128×128 is subsampled uniformly at random from a randomly chosen subimage. Such a hierarchical two-step subsampling method approximates sampling crops uniformly at random from the whole image, while avoiding an I/O-bottleneck while training. We further apply random rotations and flips to each selected crop. Moreover, to ensure VirTues learns representations robust to varying combinations of markers, and enhance its ability to generalize to unseen datasets and markers, we also randomly drop up to 25% of input channels randomly, and exclude them from the training sample.

Optimization. We train VirTues for 3000 epochs using AdamW (Kingma, 2014) with an effective batch size of 128, achieved by accumulating gradients over 8 mini-batches. Each epoch involves iterating over one random crop from each training image. A weight decay, applied to all weights except biases and Layer Normalization terms, follows a cosine schedule from 0.04 to 0.4. The learning rate is initialized with a linear warmup over 10 epochs, followed by a cosine schedule from 2×10^{-4} to 2×10^{-6}. Training employs automatic mixed precision. Gradients are clipped to a maximum norm of 1.0.

A.3 BASELINES FOR DOWNSTREAM TASKS

We compare VirTues primarily with two baselines: (1) ResNet (Sorin et al., 2023) and (2) CA-MAE (Kraus et al., 2024). We utilize a pretrained ResNet (Sorin et al., 2023) based on the

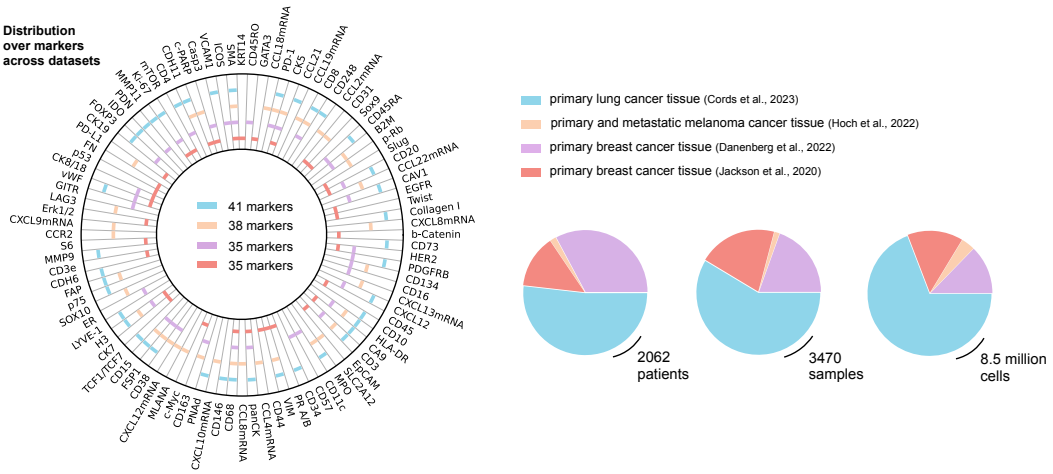

Figure S1: Overview of marker panels used in pretraining datasets and their respective sizes in terms of number of patients, images and cells measured.

approach described by Sorin et al. (2023). Specifically, each channel is individually embedded using the ResNet50 (He et al., 2016) architecture, pretrained on ImageNet-1K (Deng et al., 2009), where each channel is duplicated thrice to match the input dimension. The resulting ResNet50 embeddings are then projected channel-wise to their 16 principal components and concatenated to generate the crop representation. We specifically choose the top 16 prinicipal components, so that the final crop representation retains approximately the same dimension as VirTues'. Since ResNet is a convolutional neural network that generates niche level representations directly and is thus unable to embed patches at the cellular level. Hence, we compare against ResNet (Sorin et al., 2023) only for niche-level and tissue-level tasks. Furthermore, since it utilizes a pretrained network and is channel-agnostic, we also compare against it for zero-shot tasks. Secondly, we use a channel-agnostic masked autoencoder proposed by Kraus et al. (2024). This model adopts a multi-channel tokenization strategy and an encoder-decoder framework similar to VirTues, but with key differences: each channel is assigned a separate decoder, marker identities are not encoded in the tokenization, and full attention is utilized. These design choices restrict the model's capability to scale to a large number of channels, imposes efficiency issues and hinders the model's ability to zero-shot to unseen markers or datasets. We pretrain CA-MAE (Kraus et al., 2024) with the reconstruction objective and procedure described in Kraus et al. (2024), setting the patch size to 8 to capture information at the cellular scale. Further, we train CA-MAE (Kraus et al., 2024) for each dataset separately addressing scaling issues and mitigating the computational bottlenecks caused by the unequal representation of channels in datasets, which would otherwise lead to a disproportionate increase in model parameters without a corresponding increase in data. CA-MAE (Kraus et al., 2024) can be used to generate both cell-level and niche-level representations. For cell-level representations, we average the embedded tokens along the channel dimension. For self-supervised niche-level representations, we take the average of all embedded tokens of the crop.

A.4 MAE Reconstruction Losses

We report mean squared reconstruction errors per dataset, marker and masking strategy in Suppl. Fig. S2.

A.5 Further Analysis of Biological Discovery and Clinical Diagnostic Tasks

Cell type classification. For breast cancer tissue, VirTues achieves F1-scores of 0.55 for stromal cells and 0.49 for ER- cells, surpassing CA-MAE (Kraus et al., 2024) by 4.94% and 18.27% respectively. The performance advantage is even more pronounced for rare cell types such as NK and B cell, which cover only 1.17% and 3.27% of the test set. We first notice that VirTues is able to achieve an F1-score of 0.51 for NK, while CA-MAE (Kraus et al., 2024) fails to identify any

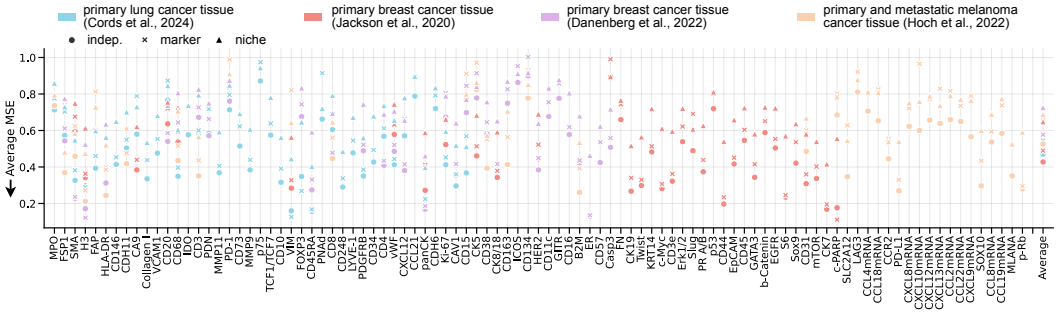

Figure S2: Overview of reconstruction errors per dataset, marker and masking strategy in terms of mean squared error computed over all masked tokens' pixels.

NK cell. In case of B cells, VirTues achieves an F1-score of 0.41, surpassing CA-MAE (Kraus et al., 2024) by 66.26%. In the coarse-grained lung cancer cell type analysis, VirTues maintains its superior performance, achieving F1-scores of 0.91 for tumor cells and 0.75 for immune cells, representing improvements of 4.05% and 48.83% over CA-MAE (Kraus et al., 2024) respectively. For the challenging and underrepresented vessel cell identification task, VirTues achieves an F1-score of 0.36, outperforming the baseline by a significant margin of 324.38%. This advanced ability to predict rare cell types, become even more apparent in the results of the fine-grained cell type classification on Cords et al. (2023) (see Suppl. Fig. S3), where the F1-scores of VirTues exceed those of CA-MAE (Kraus et al., 2024) in all classes, with significant margins especially in the rare cell types such as the different cancer-associated fibroblasts (CAFs).

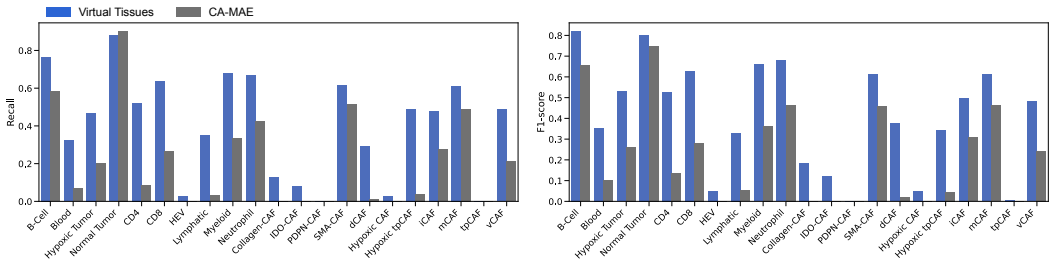

Figure S3: Classification results for fine-grained cell type prediction on Cords et al. (2023).

Niche level classification. For suppressed expansion regions, binary classification based on VirTues learned representation achieves an accuracy of 0.85, surpassing CA-MAE (Kraus et al., 2024) by 4.11% ($P < 0.006$), and ResNet (Sorin et al., 2023) by 3.45% ($P < 0.005$). The VirTues-based model maintains strong performance across other tissue structures, with accuracies of 0.81 for TLS-like regions and 0.78 for PDPN+ regions, consistently outperforming CA-MAE (Kraus et al., 2024) by 15.01% ($P < 0.006$) and 11.49% ($P < 0.005$), and ResNet (Sorin et al., 2023) by 7.28% ($P < 0.005$) and 4.24% ($P < 0.005$) for TLS-like and PDPN+ respectively. P values are computed using a Mann-Whitney U test.

Tissue level classification. For breast cancer, VirTues representations achieve accuracies of 0.89 for ER Status and 0.68 for cancer grade prediction, representing improvements of 7.26% ($P < 0.005$) and 32.16% ($P < 0.005$) over CA-MAE (Kraus et al., 2024), and 1.88% ($P < 0.199$) and 8.86% ($P < 0.02$) over ResNet (Sorin et al., 2023) respectively. Similarly, in lung cancer analysis, the model demonstrates superior performance in cancer type prediction (0.87 accuracy, 11.72% over CA-MAE (Kraus et al., 2024) ($P < 0.006$) and 4.36% over ResNet (Sorin et al., 2023) ($P < 0.005$)) and cancer grade prediction (0.63 accuracy, 16.21% over CA-MAE (Kraus et al., 2024) ($P < 0.006$), and 5.37% over ResNet (Sorin et al., 2023) ($P < 0.005$)).

A.6 Model and Training Details for Niche and Tissue Classification Tasks

The ABMIL classifiers aggregate the input representations using gated-attention computed over four heads, each with a hidden dimension of 256. The aggregated representation is projected either to a single binary logit value or class logits followed by a sigmoid activation resp. softmax to compute the class probabilities. The models are trained using the Adam (Kingma, 2014) optimizer with learning rate of 10^{-4} and a batch size of 32. We use early stopping, with patience of 5 epochs over the training loss. For ResNet (Sorin et al., 2023), we employ for niche level classification logistic regression instead of ABMIL, as ResNet (Sorin et al., 2023) directly provides aggregated niche-level representations, and for tissue level classification tasks these niche levels representations are used instead of cell summary representations. To train and evaluate the classifiers, we precompute training and testing sets of embedded cell or niche representations: Each training and testing image is divided in non-overlapping crops of size 128×128, and excess pixels at the borders of the images are disregarded. Each crop, representing a single niche, is embedded individually using VirTues and the baseline methods.

A.7 Zero-shot Experiments

To assess VirTues' ability to generalize to a new dataset, we retrain the model on Cords et al. (2023), Danenberg et al. (2022) and Jackson et al. (2020) and subsequently evaluate the performance on (1) reconstruction, (2) cell type classification and (3) tissue level classification tasks on the withheld dataset Hoch et al. (2022). We benchmark against the model pretrained on the complete collection of datasets.

For reconstruction experiments, we employ the same three masking strategies, independent, marker and niche, as the for the non-zeroshot evaluation. However, since the marker panel in the study of Hoch et al. (2022) partially overlaps with the markers of the training datasets, we distinguish two cases: (1) reconstruction of markers present during pretraining and (2) reconstruction of unseen markers. For the first case, examples of different marker reconstructions are visible in Fig 5c. Quantifying the average MSE over all markers (see Fig 5b) shows that the reconstruction performance for independent and niche masking is relatively stable across the trained (many-shot) and the zero-shot setting. When masking an entire marker and thus any information of inter-marker correlation, the performance significantly drops. This suggest that in independent and niche masking, spatial attention can leverage local spatial context and marker attention dynamically picks up correlations between markers to reconstruct missing regions. However, when an entire marker is masked, reconstruction relies solely on learned relations between markers from the training data. Unsurprisingly, in zero-shot settings, these may not fully generalize to novel cancer types where marker relationships could differ from the training distribution.

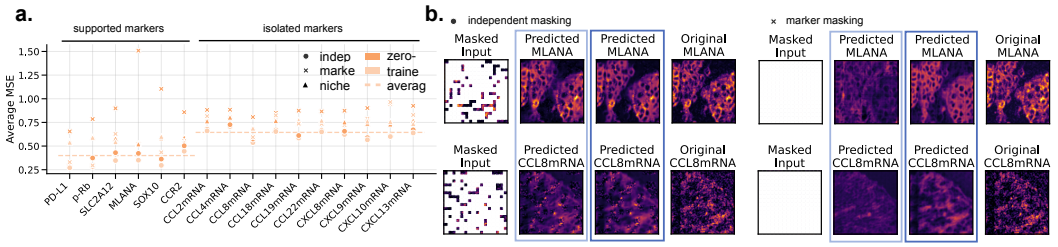

Figure S4: Zeroshot reconstruction results on Hoch et al. (2022) for markers unseen during training.

For the reconstructions of markers unseen during training, we observe similar trends (see Suppl. Fig. S4c). For independent and niche masking, the zero-shot model achieves almost the same reconstruction quality as the model instance whose training included Hoch et al. (2022). In contrast, in the setting where the entire marker is masked, VirTues can reconstruct only the overall expression of the marker, however not its intensity. To investigate further the importance of the PLM, which serves in VirTues' architecture as a crucial building block to enable the integration of new markers, we separate the markers of Hoch et al. (2022) unseen during training into two groups: The first group consist of markers, which are well supported among markers seen during training, i.e., they have close neighboring markers in the PLM embedding space, whereas the second contains the

isolated markers. The average squared reconstruction errors, reported in Suppl. Fig. S4a thereby increase from 0.40 in supported markers to 0.67 in isolated markers.

For the cell type classification and tissue level classification tasks, we use annotations provided in Hoch et al. (2022). To compute the embeddings we only use markers of Hoch et al. (2022) which were present during training and follow otherwise apart from the model instance the same experimental setup as in the many-shot experiments. Strikingly, quantitative evaluation (Fig. 5a) shows strong performance in these biological discovery and clinical diagnostic tasks even in the zero-shot setting, i.e., for melanoma samples cell type classification and clinical feature prediction performance levels approach those achieved on a trained model instance. At the cellular level, we compare the performance of VirTues (zero-shot) with VirTues (many-shot) on cell type classification, distinguishing lymphocytes, macrophages, stromal, T cell and tumor cells. VirTues (zero-shot) achieves F1-scores of 0.94 for tumor cells and 0.81 for T cells, as compared to VirTues (many-shot) achieving F1-scores of 0.94 and 0.84 respectively. Moreover, we notice that in macrophages, VirTues (zero-shot) outperforms VirTues (many-shot) achieving F1-scores of 0.51 and 0.48 respectively. At the tissue level, VirTues shows stronger performance than ResNet (Sorin et al., 2023) in general. VirTues (zero-shot) demonstrates superior performance, for example, in relapse prediction (0.88 accuracy, improving performance by over 79.49% as compared to ResNet (Sorin et al., 2023)), and mutation prediction (0.76 accuracy, with relative performance improvement of 32.59% over ResNet (Sorin et al., 2023)). Benchmarking here is reduced to ResNet (Sorin et al., 2023) in the tissue level as CA-MAE (Kraus et al., 2024) is not designed to operate in the zero-shot setting. As mentioned before, ResNet (Sorin et al., 2023) is restricted to niche- and tissue-level tasks.