ANNOTATEANYCELL: OPEN-SOURCE AI FRAMEWORK FOR EFFICIENT ANNOTATION IN DIGITAL PATHOLOGY

A PREPRINT

Shourya Verma*[1], Aditya Malusare*[2,5], Mengbo Wang[1,5], Luopin Wang[1,5], Arpan Mahapatra[1]

Abigail English[3], Abigail Cox[3], Meaghan Broman[3], Simone De Brot[4], Grant Burcham[3]

Deborah Knapp[5], Deepika Dhawan[5], Mario Sola[3], Vaneet Aggarwal[1,2,5], Ananth Grama[1]

Nadia Atallah Lanman[3,5]

* Equal Contribution
[1] Computer Science, [2] Industrial Engineering, [3] Comparative Pathobiology, Purdue University;
[4] Veterinary Pathology, University of Bern; [5] Purdue Institute for Cancer Research

November 2, 2025

ABSTRACT

Digital pathology has grown exponentially in recent years, which has created an urgent need for methods to analyze large histopathological image slides. We present an open-source semi-supervised deep learning framework to efficiently analyze and annotate whole slide images in digital pathology. Our software "AnnotateAnyCell" is an intelligent interactive web-based interface, which facilitates exploration and labeling of cellular structures using active contrastive learning, and dimensionality reduction techniques. We integrate an active learning mechanism that continually improves the predictive model based on new annotations, showing improvements in classification accuracy over iterations of human-in-the-loop feedback learning. AnnotateAnyCell aims to enable rapid annotation of large-scale histopathological datasets by assisting experts with quantitative and contextual information at a cellular level leading to efficient and accurate labeling of cells. This framework has the potential to accelerate cancer biology and personalized medicine research by dramatically reducing the time and effort required for cellular analysis in tissue samples. Github: https://github.com/shouryaverma/AnnotateAnyCell/tree/main

Keywords Digital Pathology, Cell Annotation, Active Learning

1 Introduction

The integration of artificial intelligence and machine learning into digital pathology has transformed histopathological analysis, driven by advances in whole slide imaging technologies and widespread adoption of digital workflows in clinical laboratories. This transition has generated vast repositories of high-resolution histopathological images, creating unprecedented opportunities for computational analysis while presenting significant technical challenges. Modern whole slide images present unique computational difficulties due to their extraordinary scale and complexity. A typical WSI at 40× magnification contains over 100,000 × 100,000 pixels with hundreds of thousands of individual cells, often exceeding several gigabytes per slide. This volume far surpasses what can be systematically analyzed through manual examination, creating an urgent need for automated analysis systems capable of extracting clinically relevant information at scale.

However, developing robust computational approaches for WSI analysis poses several fundamental challenges. First, the dimensionality problem extends beyond image size, and diagnostically relevant information is encoded in subtle morphological variations requiring multi-scale analysis pipelines. Second, histopathological images exhibit extreme heterogeneity due to biological variation, tissue preparation protocols, and scanner-specific artifacts. Third, cellular segmentation and classification in complex tissue environments remains computationally demanding, particularly with overlapping cells and poor image quality regions common in clinical samples.

Current state-of-the-art approaches rely on fully supervised deep learning requiring extensive manual annotations. This process is labor-intensive and certified pathologists spend hundreds of hours manually delineating cellular boundaries and assigning class labels. Models trained on these datasets frequently exhibit poor generalization across institutions and imaging platforms, where a model trained at one hospital may perform poorly at another due to variations in fixation protocols or scanner calibration, presenting a major barrier to clinical deployment. Semi-supervised learning methodologies offer promising solutions by leveraging abundant unlabeled data alongside limited expert annotations. However, their adoption in digital pathology has been limited due to several technical barriers: most semi-supervised methods require substantial computational infrastructure and GPU resources that exceed typical clinical laboratory capabilities; existing platforms lack intuitive interfaces that integrate seamlessly into pathologist workflows, requiring extensive computational expertise; and many approaches fail to provide real-time feedback loops where model improvements are immediately visible to experts, breaking the iterative refinement cycle critical for building trust in AI-assisted diagnostics. Our framework addresses these adoption barriers through a lightweight web-based architecture deployable on standard hardware, an embedding-based visualization interface that requires no machine learning expertise, and asynchronous model retraining with live progress updates that maintain expert engagement throughout the annotation process.

Our framework addresses these challenges through a novel approach combining deep learning with an interactive web-based interface designed for pathologist workflows. The core innovation involves performing annotations within a learned embedding space rather than directly on high-resolution pixels. By applying dimensionality reduction techniques like UMAP to deep feature representations, we transform complex WSI data into intuitive 2D visualizations where morphologically similar cellular structures naturally cluster together. This embedding-based approach offers key advantages: it reduces cognitive load by presenting cellular populations in structured, navigable formats; captures complex non-linear relationships between cellular characteristics beyond pixel-level similarities; and facilitates active learning strategies that intelligently select the most informative samples for expert review. By integrating these innovations with domain-specific pathology considerations, our framework represents a significant step toward making large-scale histopathological analysis both computationally tractable and clinically viable.

2 Related Work

Cell annotation tools. Digital pathology has benefited from increasingly sophisticated annotation tools designed to handle gigapixel whole slide images. Desktop applications like QuPath [Bankhead et al., 2017] and Orbit [Stritt et al., 2020] provide efficient tile-based visualization with manual annotation capabilities, while commercial platforms including HALO (Indica Labs) and Visiopharm dominate clinical deployment with FDA-cleared algorithms, validated workflows, and enterprise support infrastructure. Web-based platforms such as Cytomine [Marée et al., 2016] and Digital Slide Archive [Gutman et al., 2017] enable collaborative annotation through RESTful APIs and distributed architectures, though they require substantial server infrastructure. Recent AI-assisted tools have dramatically improved annotation efficiency: Quick Annotator [Miao et al., 2021] demonstrated significant speedups over manual annotation by training U-Net models concurrently during annotation and using UMAP-based patch selection for active learning, while PatchSorter [Walker et al., 2024] achieved improvements through deep learning-based object labeling.

Active learning in digital pathology. Active learning has emerged as a critical approach to address the substantial annotation burden in digital pathology. Recent work has demonstrated significant efficiency gains through intelligent sample selection strategies. DADA (Diversity-Aware Data Acquisition) combined uncertainty and diversity clustering to achieve comparable AUC with 96% fewer patches for tumor-infiltrating lymphocyte classification [Meirelles et al., 2023]. Human-in-the-loop systems like Quick Annotator [Miao et al., 2021] have reduced annotation time through iterative review-and-revise workflows where models provide suggestive annotations that pathologists can rapidly correct. Uncertainty quantification methods, particularly Monte Carlo dropout [Dolezal et al., 2022], have enabled high-confidence predictions robust to domain shift. Despite these advances, most active learning systems operate at the patch or region level rather than individual cells, and few leverage modern embedding spaces for direct interactive annotation. AnnotateAnyCell extends these approaches by combining active learning with contrastive learning-based embeddings in UMAP space, enabling pathologists to annotate cells interactively across learned feature representations rather than spatial coordinates alone.

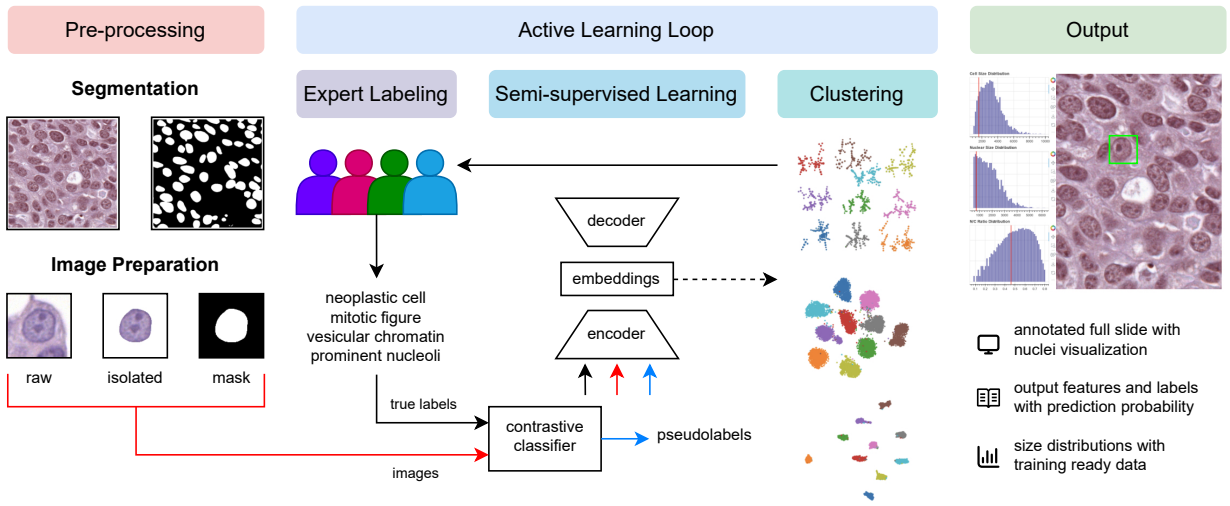

Figure 1: Overview of the "AnnotateAnyCell" framework showing the complete pipeline from image preprocessing to output generation. (a) Pre-processing stage: Raw H&E-stained tissue images undergo segmentation using Cellpose models to generate cellular and nuclear masks. Each detected cell is extracted as three complementary representations: raw image patches, isolated cellular regions, and semantic masks. (b) Active Learning Loop: Expert pathologists (represented by colored figures) provide initial annotations for nuclear morphology features including neoplastic cells, mitotic figures, vesicular chromatin, and prominent nucleoli. These labeled samples train a contrastive classifier that generates embeddings and pseudolabels for unlabeled data. A convolutional autoencoder processes the multi-modal inputs (raw, isolated, mask) to learn compact latent representations. (c) Output: The system produces annotated whole slide images with nuclei visualization, quantitative feature distributions, and prediction probabilities for downstream analysis. The iterative process refines the model's understanding of cellular morphologies through expert feedback.

3 Methods

Our framework implements a semi-supervised active learning pipeline (Figure 1) for efficient annotation of cellular structures in whole-slide histopathology. The workflow consists of four iterative stages that refine both model understanding of morphologies and annotation quality. Firstly, high-resolution H&E-stained tissue images are segmented using state-of-the-art deep learning models to extract individual cells, generating multi-modal representations including raw patches, isolated regions, and semantic masks. Next, extracted cells are presented to pathologists within a learned embedding space, where morphologically similar cells cluster, enabling efficient exploration and labeling of phenotypes. Expert labels and abundant unlabeled data are integrated through a contrastive learning framework that generates high-confidence pseudolabels and prioritizes the most informative samples for review. A multi-modal autoencoder processes labeled and pseudolabeled data to learn compact latent embeddings capturing fine-grained morphology and higher-level context, producing updated visualizations that guide the next iteration. This iterative loop continues until annotation coverage is sufficient, progressively improving both model accuracy and expert efficiency. The following subsections detail each pipeline component.

Dataset and Sample Collection. We evaluated our framework on canine invasive urothelial carcinoma (IncUC) specimens, a spontaneous tumor model that recapitulates human muscle-invasive bladder cancer biology and serves as a comparative oncology platform for translational research. Tissue samples were collected from client-owned dogs diagnosed with naturally occurring IncUC at the Purdue University Veterinary Teaching Hospital following informed owner consent under Purdue Animal Care and Use Committee protocols. Formalin-fixed paraffin-embedded tissue blocks were sectioned at 4 μm thickness and stained with hematoxylin and eosin using standardized automated protocols (Leica Autostainer XL). Slides were digitized at 40× magnification (0.25 μm/pixel resolution) using an Aperio AT2 scanner, generating whole slide images with typical dimensions of 80,000×100,000 pixels. Cases were selected based on diagnostic quality (minimal artifacts, adequate fixation, uniform staining) and tumor cellularity (>30% neoplastic nuclei) to ensure sufficient cellular diversity for annotation. We selected 8 representative slides spanning histological grade (low, intermediate, high), mitotic activity (0-15 mitoses per 10 high-power fields), and architectural patterns (papillary, solid, nested) to capture the morphological spectrum of canine IncUC. Expert pathologists (board-certified veterinary pathologists with 5-15 years of experience) collaboratively defined annotation categories based on diagnostically relevant nuclear morphology features: mitotic figures (canonical condensed chromatin patterns and bizarre/atypical configurations), nucleolar characteristics (absent, single, multiple, prominent), chromatin texture

(vesicular vs. hyperchromatic), and nuclear shape (circular, oval, irregular). These features were selected for their established prognostic significance in both canine and human urothelial carcinoma, relevance to grade assignment, and feasibility for reliable expert annotation with moderate inter-observer agreement.

User Interface. The interactive user interface for our active learning system was built with BokehJS, a JavaScript library with Python bindings that supports customizable client-side UI elements. This framework integrates the computational backend with the visualization front end, enabling real-time multi-user collaborative annotation. The server architecture maintains persistent user sessions with checkpoint functionality, allowing experts to save progress and revert to earlier states, which is useful for extended sessions and comparative analyses across annotation strategies or expert interpretations. Main Annotation Interface. The primary workspace 3 comprises three coordinated panels. The central panel displays the UMAP embedding space as an interactive scatter plot where each point represents a cellular tile. Points are color-coded by annotation status (blue: unlabeled, green: labeled, orange: selected). Hovering over a point reveals contextual information: the original image tile, segmented nuclear/cellular regions, and spatial coordinates in the whole slide image (WSI). The left panel guides users through a stepwise protocol with interactive analytics. Histograms of nuclear area and nuclear to cytoplasmic ratios and key pathological features, update dynamically with selections, with red markers highlighting specific measurements relative to population statistics. The right panel presents 512×512 high-resolution previews with annotation controls for morphological classes (e.g., mitotic figures, chromatin patterns, nucleoli, nuclear shape). The taxonomy, developed with pathologists, is easily customizable and ensures coverage of clinically relevant morphologies with consistent terminology. Each annotation records confidence scores and timestamps for downstream analysis of inter-observer variability and annotation dynamics. Progress tracking reports dataset coverage (e.g., "300/300 labeled points") in real time. Experts can trigger model retraining at any time using the "Update Clusters" button, which executes asynchronously. Training progress is displayed live, and once complete, the embedding visualization updates automatically, revealing refined clusters and novel morphological relationships. After annotation, the system provides a results-focused interface 4, reusing the three-panel layout. The left panel summarizes annotations quantitatively (e.g., counts per morphology type). The central panel displays the full WSI with cyan markers for all annotated cells, supporting pan/zoom exploration of tissue architecture. The right panel enables detailed inspection of individual annotations. Data can be exported via CSV download, while advanced filters highlight cells matching specific morphological criteria for hypothesis generation and large-scale pattern discovery. Our design balances analytical depth and usability, separating annotation from analysis to optimize cognitive load while maintaining consistent visual patterns. User studies with expert pathologists showed significant improvements in efficiency and consistency compared to manual methods, accelerating dataset curation for computational pathology while preserving annotation quality.

Image Pre-processing. Preparing high-quality input data requires precise image preprocessing to segment, extract, and normalize cellular and nuclear features from whole-slide images (WSIs). Digital microscopy captures large high-resolution H&E-stained slides, which undergo segmentation to delineate cells and nuclei for morphology, size, and shape analysis. We use pre-trained Cellpose models Stringer et al. [2021]: cyto3 for cell segmentation and nuclei for nuclear segmentation. Each produces segmentation masks and boundary outlines, which are filtered to remove artifacts, partial objects, and merged cells. We then integrate cell and nuclear masks, enforcing containment (each nucleus within a cell body). For cells without detected nuclei, the entire cell is treated as a nucleus to maintain downstream consistency. From segmented regions, we extract tile patches: 128×128 pixels for model training and 512×512 pixels for expert inspection. Three versions of each tile are generated: (1) raw WSI image, (2) semantic mask, and (3) isolated cell/nucleus image with surrounding tissue removed. Size constraints (1,000-10,000 pixels) filter artifacts and select valid cells. All tiles are bilinearly interpolated to standard dimensions. Metadata including nuclear-to-cytoplasmic ratio, nuclear and cellular areas, and centroid coordinates, is also stored for analysis and annotation context. Finally, UMAP is applied to the 128×128 patches for dimensionality reduction, enabling efficient visualization of the feature space and clustering of morphologically similar cells.

Active Learning Loop. To minimize annotation burden while maximizing expert input, we developed a semi-supervised active learning pipeline that iteratively selects the most informative samples for expert review. The process begins with unsupervised initialization, where UMAP embeddings of 128×128 cellular tiles create a 2D representation of the morphological feature space. This interactive visualization allows pathologists to explore cellular diversity and annotate an initial set of representative tiles. Annotations are stored in a MongoDB database with secure access, version control, and support for multi-expert collaboration. Next, a contrastive learning model is trained with a convolutional backbone, projection head, and classification head. The objective combines supervised classification loss on labeled data with unsupervised contrastive loss on both labeled and unlabeled samples, embedding similar morphologies close together while separating dissimilar ones. The trained model generates class-balanced pseudolabels, selecting the top-N highest-confidence samples per class to avoid bias toward dominant morphologies. For subsequent annotation rounds, uncertainty sampling with diversity promotion identifies the most informative examples. We then train a convolutional autoencoder on both annotated and pseudolabeled data, processing raw images, nuclear features, and semantic masks

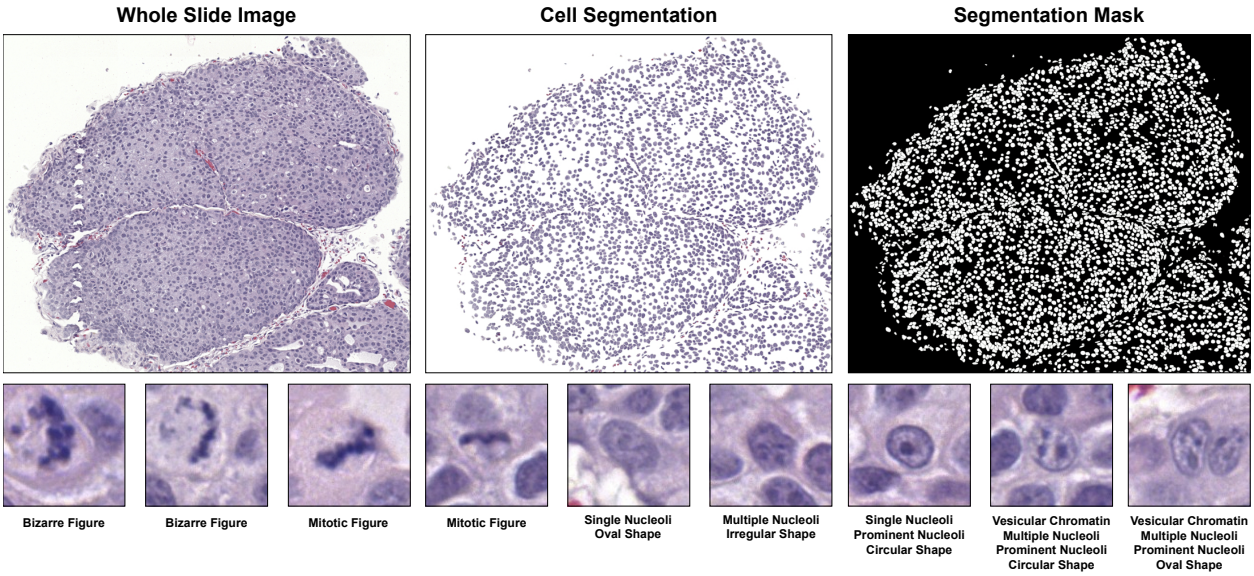

Figure 2: Computational workflow from raw H&E-stained tissue to individual cell extraction. Left: Representative whole-slide image (WSI). Center: Cell segmentation results using pretrained Cellpose models Stringer et al. [2021]. Right: Binary segmentation mask showing all detected valid cells. (Bottom row) Nine examples of 128×128 pixel cellular tiles illustrating morphological diversity across expert-defined annotation categories. Examples include bizarre figures showing irregular mitotic configurations, mitotic figures demonstrating canonical mitotic chromatin condensation patterns, nucleolar variations (single vs. multiple nucleoli) combined with nuclear shape classifications (circular, oval, irregular), and chromatin texture patterns (vesicular chromatin).

simultaneously. With skip connections and a variational bottleneck, the autoencoder captures fine-grained morphology and contextual structure in its latent space. UMAP projection of these embeddings yields an updated visualization that clusters morphologically similar cells while preserving expert-defined boundaries. Experts use this refined space to validate or correct pseudolabels and annotate uncertain regions. This iterative loop progressively refines the nuclear morphology space, improving classification accuracy, annotation efficiency (measured as performance gain per label), and discovery of novel phenotypes. The following subsections detail key pipeline components.

Pseudolabel Generation. Our semi-supervised learning framework integrates supervised and unsupervised techniques to generate reliable pseudolabels for unlabeled data. It consists of three components: a convolutional classifier for feature extraction, a contrastive learning module for representation enhancement, and a selective pseudolabeling strategy for label propagation. The base model is a CNN that extracts features from input tiles using four convolutional blocks (Conv2D with 3×3 kernels, ReLU activation, max-pooling), progressively increasing channel depth (16→128) while reducing spatial dimensions. This is followed by four fully connected layers (1024→128 neurons) with batch normalization and 50% dropout to prevent overfitting on limited labeled data. The final layer maps to expert-defined nuclear morphology classes.

To improve generalization, we extend this with a contrastive learning framework. A projection head maps CNN features into a 128-dimensional normalized embedding space, where embeddings of similar samples are pulled closer and dissimilar samples are pushed apart. The supervised classification objective is combined with InfoNCE loss, defined as:

$$L_{\text{InfoNCE}} = -\log \frac{\exp(sim(q, k_+)/\tau)}{\sum_{i=0}^{K} \exp(sim(q, k_i)/\tau)} \tag{1}$$

where q is the query embedding, k_+ the positive key, k_i negative keys, $sim(u, v)$ the cosine similarity, τ a temperature parameter, and K the number of negatives. The total objective is a weighted sum of cross-entropy (CE) loss on labeled samples and InfoNCE loss on both labeled and unlabeled data:

$$L_{\text{total}} = \alpha L_{CE} + \beta L_{\text{InfoNCE}} \tag{2}$$

with α and β controlling loss balance.

Figure 3: Main annotation interface of the "AnnotateAnyCell" system demonstrating the embedding-based annotation workflow. Left panel: Displays workflow guide for pathologists, nuclear size distribution histogram, and nuclear-to-cell ratio distribution with red indicators showing the currently selected cell's measurements. Center panel: Interactive UMAP embedding visualization where each point represents a cellular tile extracted from the whole slide image. Cells are color-coded based on annotation status (blue: unlabeled, green: labeled, orange: selected) and naturally cluster based on morphological similarities. Right panel: Shows a high-resolution preview of the selected cell tile (Tile 710) with morphological annotation options including mitotic figures, chromatin patterns, nucleoli characteristics, and nuclear shape classifications. The interface includes progress tracking (300/300 labeled points), version control, and model retraining capabilities ("Update Clusters" button).

For pseudolabeling, the trained model assigns labels to unlabeled data based on confidence scores. Instead of a fixed threshold, we use a class-balanced selection strategy, choosing the top-N most confident predictions per class (e.g., $N = 20$), ensuring balanced representation across morphology types and mitigating bias toward dominant classes. Unselected samples retain a null label and contribute only to unsupervised learning. Experimental validation Su et al. [2015] shows that this approach outperforms both fully supervised baselines and naive pseudolabeling strategies, effectively combining expert-labeled data, contrastive learning, and class-balanced pseudolabeling to achieve robust nuclear morphology classification with limited initial annotations.

Autoencoder Model. The autoencoder is central to our semi-supervised pipeline, enabling unsupervised feature learning, dimensionality reduction, and data visualization. It processes preprocessed image tiles to learn compact representations of cellular and nuclear morphologies that integrate expert-defined labels with intrinsic data structure. We implement a multi-modal convolutional autoencoder with three input channels: raw cellular images, extracted nuclear regions, and semantic segmentation masks. Each modality passes through a parallel encoder (four Conv2D layers: $3\times16 \rightarrow 16\times32 \rightarrow 32\times64 \rightarrow 64\times128$, ReLU activations). Skip connections between encoder and decoder layers preserve spatial detail. Encoded features are flattened, passed through fully connected layers (1024→512, batch normalization), and concatenated with one-hot pseudolabels. The integrated embedding is further processed through additional fully connected layers (1024→512). Following the variational autoencoder (VAE) framework, mean and logvar projections map to a 256-dimensional latent space, with the reparameterization trick applied during training. This bottleneck enforces a smooth latent representation, improving interpolation and clustering. Decoders mirror the encoder structure with transposed convolutions, reconstructing modality-specific inputs.

The training objective combines reconstruction and feature prediction losses. Reconstruction is defined as:

$$L_{\text{recon}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{3}$$

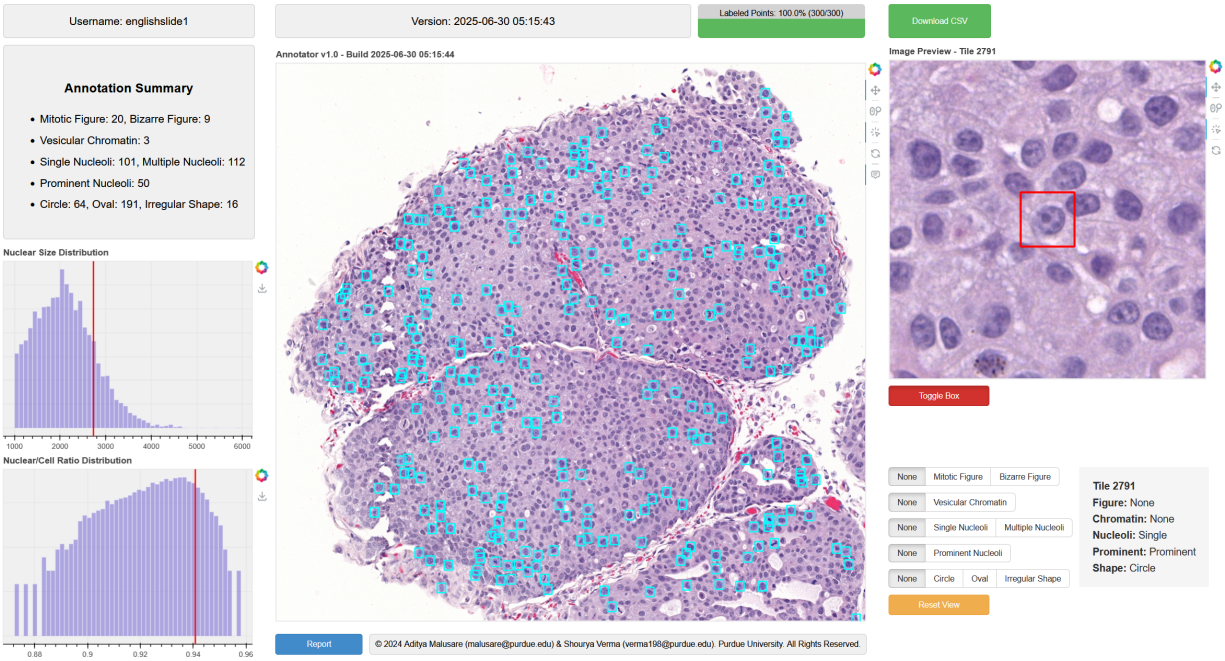

Figure 4: Results and output interface showing the final annotated dataset and analysis capabilities. Left panel: Displays annotation summary statistics including counts for different morphological features and the same distribution histograms as the annotation interface. Center panel: Whole slide image view with cyan markers indicating the spatial locations of all annotated cells across the tissue sample, enabling pathologists to assess spatial distribution patterns and tissue architecture context. Right panel: Individual cell inspection tool showing detailed annotations for the selected cell (Tile 2791) with its specific morphological classifications. The interface includes data export functionality (Download CSV), version information, and the ability to filter and visualize cells based on specific morphological criteria. This output interface enables analysis of annotation results and supports research workflows requiring quantitative cellular phenotype data.

where x_i is the input and $\hat{x}_i$ the reconstruction. Feature prediction uses cross-entropy on pseudolabels from the contrastive classifier:

$$L_{\text{feat}} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{c=1}^{C} y_{j,c}\log(\hat{y}_{j,c}) \tag{4}$$

The total loss is a weighted sum:

$$L_{\text{total}} = \alpha L_{\text{recon}} + \beta L_{\text{feat}} \tag{5}$$

with α, β controlling trade-offs.

The encoder transforms $128\times128\times9$ input tensors into 256-dimensional embeddings that capture both supervised annotation signals and unsupervised data structure. These embeddings are projected into two dimensions via UMAP for interpretable visualization of local and global morphology organization. This multi-modal variational autoencoder improves representation quality, reconstruction fidelity, and clustering performance compared to standard architectures. Within the active learning loop, it supports efficient annotation by integrating multiple modalities, reducing dimensionality, and continuously updating the visualization interface for expert review.

4 Results

Figure 5 illustrates sequential snapshots of the learned embedding space at 0, 50, 100, 200, and 300 labeled cells, showing the progression of expert annotations (green) within the unlabeled cellular population (blue). At initialization (0 points), unsupervised UMAP projection of raw 128×128 pixel cellular tiles creates a 2D morphological landscape where all cells appear unlabeled. With 50 labeled points, annotations distribute sparsely across the embedding space as pathologists explore cellular diversity and establish initial class boundaries. At 100-200 points, annotation density increases within coherent morphological neighborhoods, reflecting the active learning strategy's intelligent sample

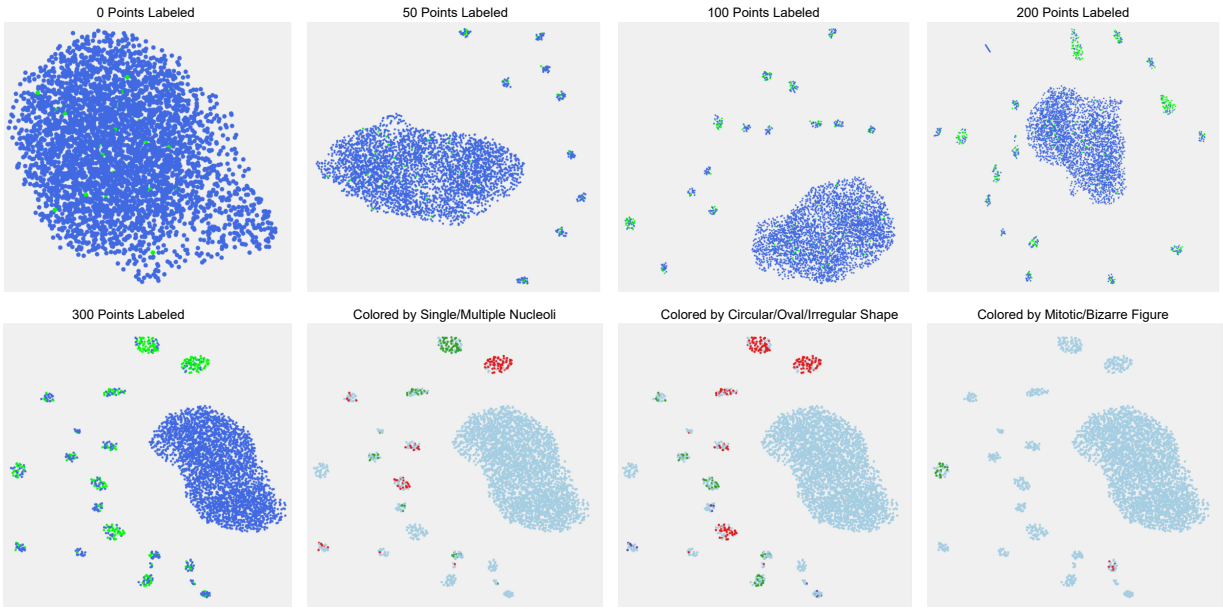

Figure 5: Progressive evolution of UMAP embedding space during active learning annotation and feature-specific clustering patterns.

selection through uncertainty sampling with diversity promotion. At 300 points, labeled cells (green) populate discrete clusters and boundary regions, with a large central cluster representing unlabeled and unpseudolabeled cells that have not yet been confidently assigned to morphological categories, and peripheral clusters capturing both annotated dominant morphologies and rare cellular phenotypes. The persistent large unlabeled cluster represents unlabeled and unclassified cells, the aim is to break the large cluster into as many sub-clusters as possible through our learning framework. (Bottom row) Feature-specific visualization of the final 300-point embedding colored by morphological classifications, revealing feature-dependent clustering quality.

Figure 6 summarizes spatial agreement maps, temporal annotation efficiency, feature-level agreement metrics, annotator reliability, inter-feature correlations, and disagreement structures. Spatial distributions of agreement (green) and disagreement (brown) reveal localized cell-level ambiguities, with disagreement intensity encoded by saturation. Annotation efficiency analysis shows that clustering-based active learning substantially outperforms sequential annotation, requiring on average 47 minutes compared to 63 minutes for 300 cells (25% reduction; 16 minutes absolute; 9.4 vs. 12.6 sec/cell). These gains were consistent across different volumes, with 22% improvement at 50 cells (18 vs. 23 minutes), 30% at 100 cells (26 vs. 37 minutes), and 25% at 150 cells (33 vs. 44 minutes), all scaling linearly with sample size. Both workflows exhibited linear growth without evidence of fatigue effects, though clustering showed narrower confidence intervals, reflecting more consistent performance. Efficiency improvements stem from embedding-based grouping of morphologically similar cells, allowing annotators to evaluate consecutive sets of 5-10 related examples, thereby reducing cognitive load and context switching. Interactive UMAP visualization and real-time statistics further enhance contextual awareness, while uncertainty- and diversity-based active learning ensures that the most informative and novel morphologies are prioritized. Consequently, the observed 25% time savings likely underestimates the true efficiency advantage when accounting for greater information content per annotation hour.

Agreement rates varied substantially across morphological features. Chromatin texture and prominent nucleoli achieved perfect concordance (100%), confirming their unambiguous diagnostic criteria. Nucleoli classification showed high agreement (95%, Cohen's $\kappa = 0.95$), while shape was more variable (81%, $\kappa = 0.36$), and mitotic activity was most challenging (64%, $\kappa = 0.58$). Reliability across 11 annotators ranged from 0.65-0.95, with Users 2, 8, 9, and 10 achieving the highest scores (≥ 0.95), most clustered between 0.80-0.95, and User 6 performing lowest (0.65). Decision consistency ranged from 0.52-0.82, with User 5 the most stable (0.82) and Users 1 and 2 the least (0.52-0.57). Spatial quality scores were high overall (0.80-1.00), with User 5 demonstrating perfect uniformity (1.00) and User 6 showing spatial bias (0.80). Integrated scatter plots revealed distinct annotator profiles: high performers (Users 5, 9, 10) combined strong consistency (0.75-0.82) with near-perfect spatial coverage (0.95-1.00), while Users 2 and 6 emerged as candidates for retraining. These results indicate that while chromatin and nucleoli are robust features suitable for

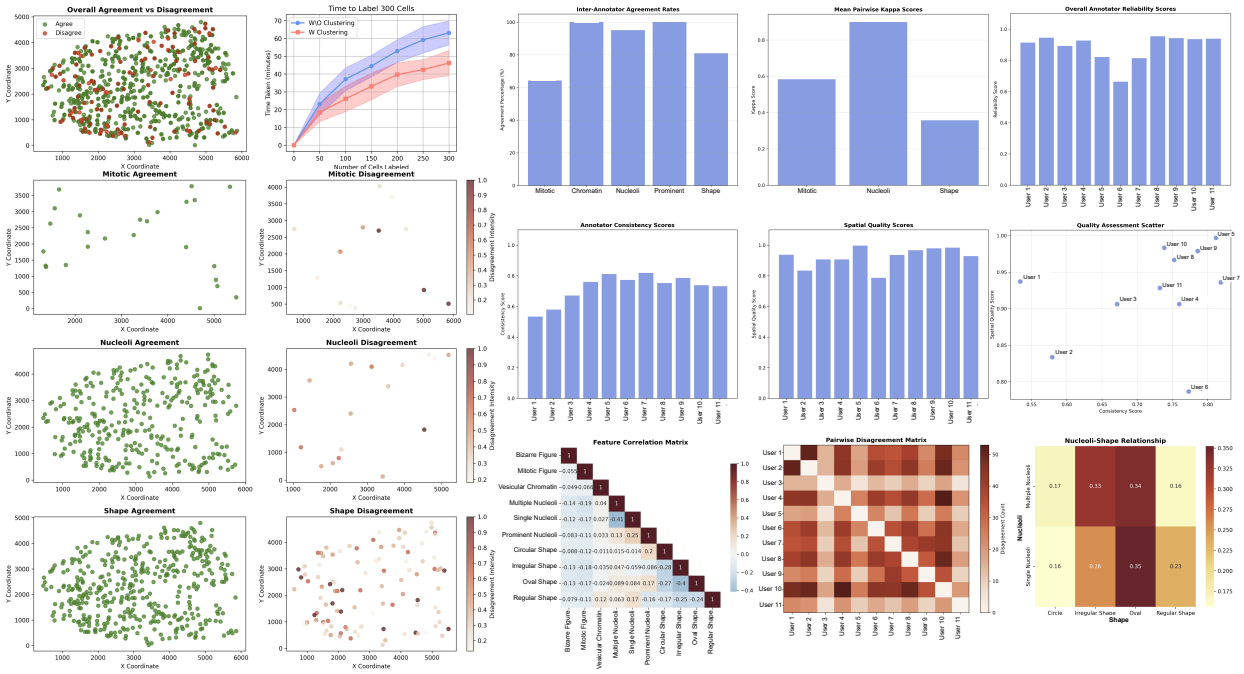

Figure 6: Comprehensive inter-annotator agreement analysis, annotation workflow efficiency, and quality control metrics across morphological features.

clinical and computational pipelines, mitotic activity and shape assessment require standardized protocols, adjudication, or noise-robust learning approaches.

Spatial annotation patterns provided further insights into feature-specific reliability. Mitotic figures were sparse (38 total sites) with 30 concordant annotations and 8 disagreements (21%), the latter ranging from 0.2-1.0 in intensity and reflecting difficulty distinguishing mitotic events from apoptotic bodies or fragments. The sparse distribution suggests potential false negatives and supports the need for multiple independent reviews. Nucleoli achieved the highest density (approx 350 sites) with uniform coverage and only 4% disagreements (15 sites), most representing borderline single vs. multiple nucleoli classifications. This low rate confirms nucleolar annotation as highly reliable and reproducible. Nuclear shape annotations, by contrast, showed the highest absolute disagreement (approx 80 sites, 17%) across the full intensity range. Moderate disagreements (0.4-0.6) clustered in certain regions, while sporadic high-intensity disagreements (≥ 0.8) reflected fundamentally divergent classification strategies. These findings underscore the interpretive subjectivity of shape classification along continuous morphological spectra. Importantly, disagreement distributions across all features showed no regional clustering, indicating that disagreements arise from intrinsic cell-level ambiguity rather than slide-level technical artifacts.

Correlation and crosstab analyses further demonstrated that nuclear features are largely independent dimensions of morphology. Pairwise correlations were weak overall ($|r| < 0.3$), with the strongest observed between single and prominent nucleoli ($r = 0.25$). Oval shape showed weak positive correlations with single ($r = 0.084$) and prominent nucleoli ($r = 0.17$), while irregular shape correlated negatively with circular ($r = -0.27$) and oval ($r = -0.40$). Multiple nucleoli were mutually exclusive with single nucleoli ($r = -0.41$). Mitotic figures ($r < 0.20$) and vesicular chromatin ($r < 0.10$) exhibited minimal associations, supporting their independence. Crosstab analysis highlighted biologically meaningful but non-redundant associations: multiple nucleoli were enriched in oval (0.34) and irregular (0.33) nuclei, while single nucleoli also occurred frequently in these shapes (0.35 and 0.26, respectively). Both nucleolar states were elevated in irregular nuclei, suggesting nuclear irregularity is an independent atypia marker. These results support a multidimensional model in which features contribute complementary diagnostic information.

Inter-annotator disagreement patterns confirmed systematic variability. Users 1, 4, and 10 exhibited the highest discord counts (>40 with multiple annotators), while Users 3 and 9 showed strong concordance and User 8 maintained consistently low-to-moderate disagreement. Asymmetric disagreement patterns (e.g., User 1 vs. 4, User 1 vs. 10) revealed entrenched divergences, with clustering analysis suggesting two "schools" of annotation: a concordant

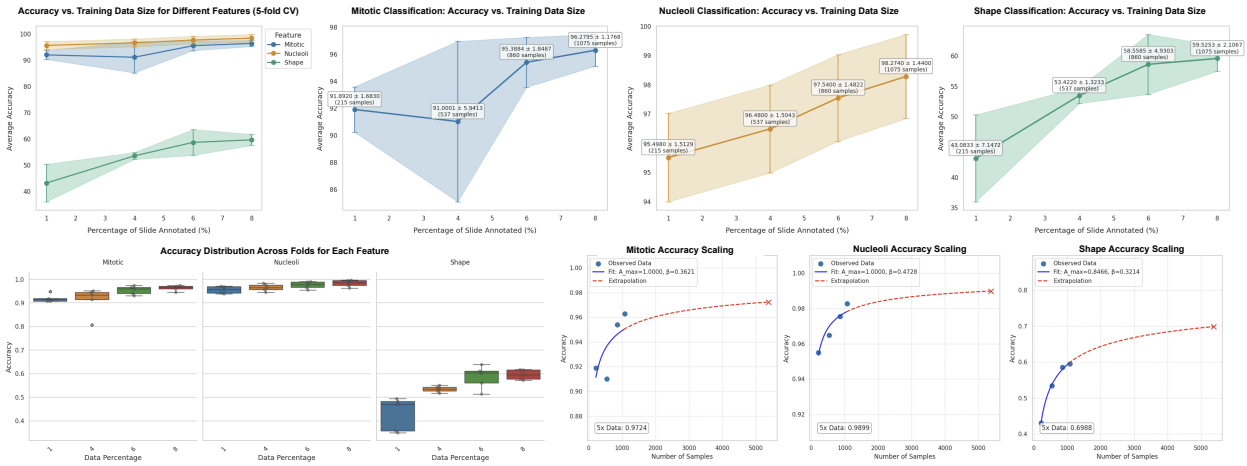

Figure 7: Performance scaling and cross-validation stability of contrastive learning-based classification across morphological features.

group (Users 2, 3, 6, 9) and a discordant group (Users 1, 4, 10). These results emphasize the importance of targeted calibration, consensus reviews, and benchmark case evaluations to harmonize annotation practices. For machine learning applications, features with near-perfect agreement (chromatin, nucleoli) provide robust ground-truth labels, while ambiguous features (mitotic figures, nuclear shape) require consensus labeling or algorithms tolerant to label noise. Incorporating disagreement intensity as a spatial uncertainty measure offers a principled way to weight or filter annotations, preventing penalization of models for inherently ambiguous cases.

Figure 7 summarizes the scaling experiment, which evaluated classification of mitotic figures, nucleoli, and nuclear shape across four training sizes (1%, 4%, 6%, and 8% of slide coverage; 215, 537, 860, and 1075 labeled samples) using 5-fold cross-validation. The semi-supervised contrastive learning framework leveraged InfoNCE loss to jointly exploit labeled and unlabeled data, enabling robust representation learning with sparse supervision. Performance curves show that nucleoli classification achieved the highest accuracy (98.3% at 1075 samples), mitotic classification reached 96.3%, and shape classification plateaued at 59.5%. Box plots and swarm plots confirm fold-to-fold stability, while power-law fits highlight distinct scaling regimes: nucleoli and mitotic figures exhibit early saturation with diminishing returns, whereas shape maintains a steep learning trajectory, suggesting substantially more labeled data would be required to approach parity.

Mitotic classification demonstrated strong baseline performance (91.9% $\pm$ 1.7% at 215 samples), a transient dip at 4% coverage (91.0% $\pm$ 5.9%), and recovery to 95.4% $\pm$ 1.8% at 6% before plateauing at 96.3% $\pm$ 1.2% with 1075 samples. The mid-range instability likely reflects class imbalance or sampling variability, as indicated by wide confidence intervals. Above approx 800 samples, performance stabilized with narrow bounds, yielding accuracy far exceeding the 64% inter-annotator agreement. Nucleoli classification was superior, starting at 95.5% $\pm$ 1.5% with only 215 samples and improving monotonically to 98.3% $\pm$ 1.4%. The consistently narrow confidence intervals mirror the 95% inter-annotator agreement, reflecting stable class boundaries and high self-similarity. The final accuracy approaches the theoretical ceiling set by annotator disagreement, indicating expert-level reliability. Shape classification proved most difficult, beginning at 43.1% $\pm$ 7.1% and rising to 59.5% $\pm$ 2.1% at full data. Despite the largest absolute gain (16.4 points), performance remained far below that of other features. Wide variability at low data volumes and persistent limitations at scale reflect the subjectivity of categorical shape labels, their dependence on continuous morphological spectra, and complex contextual factors. The 17% inter-annotator disagreement underscores these challenges, suggesting that regression-based shape descriptors, multi-scale contextual modeling, or consensus labeling may be required to improve reliability.

Scaling behaviors highlight fundamental differences in data efficiency. Nucleoli required minimal supervision, reaching 95.5% with only 215 samples and saturating quickly, with an overall gain of just 2.8 points across a five-fold increase in labels. Mitotic classification improved more substantially (4.4 points), showing continued benefit with additional data but clear diminishing returns beyond 1000 samples. Shape classification displayed the steepest gradient (43.1%-59.5%) without evidence of saturation, and extrapolation suggests thousands more labeled samples or methodological innovations would be necessary to achieve clinical-grade accuracy. These trends were consistent across folds and unaffected by overfitting, reflecting the stabilizing role of contrastive pretraining on large unlabeled datasets. Practically,

these results suggest that clinically acceptable performance (>90%) can be achieved with only 200-500 samples for nucleoli, 800-1000 samples for mitotic figures, and potentially >5000 samples for shape. Within the semi-supervised framework, however, each labeled example is effectively amplified by thousands of unlabeled samples, reducing annotation burden by an order of magnitude compared to fully supervised paradigms.

5 Discussion

The results validate the contrastive learning framework described in the Methods. The InfoNCE loss successfully leverages unlabeled data to establish morphological feature spaces where similar cells cluster, enabling robust classification from sparse annotations. The class-balanced pseudolabel generation strategy appears effective, as performance improves smoothly with additional labeled data without evidence of biased pseudolabels degrading model quality. The multi-modal autoencoder processing raw images, nuclear features, and semantic masks likely contributes to strong baseline performance, as the 128-dimensional normalized embedding space captures diverse morphological characteristics. The integration of labeled and pseudolabeled data through the variational bottleneck creates smooth latent representations that support both supervised classification and unsupervised structure learning, explaining the strong generalization despite limited supervision. The results justify the iterative active learning loop, where experts annotate informative samples identified through uncertainty sampling with diversity promotion. The steep learning curves, particularly for mitotic and shape features, indicate model predictions improve substantially with targeted annotations, creating positive feedback where updated embeddings reveal refined morphological relationships guiding subsequent annotation rounds.

For clinical deployment, the differential performance across features suggests tiered confidence thresholds. Nucleoli and mitotic classifications (96-98% accuracy) achieve reliability suitable for screening applications or diagnostic support, while shape classifications (59% accuracy) require human verification or should be reported with appropriate uncertainty quantification. Clinical workflows might automatically accept high-confidence nucleoli/mitotic predictions while flagging ambiguous cases and all shape classifications for expert review. The data efficiency demonstrated for nucleoli achieving 95.5% accuracy with 215 samples supports rapid adaptation to new institutions or tissue types. A pathologist could annotate 200-500 representative cells from local specimens, enabling the semi-supervised framework to generate institution-specific models capturing local preparation protocols and scanner characteristics while leveraging general morphological knowledge from unsupervised pretraining on large unlabeled datasets. The moderate ongoing improvements at higher training sizes suggest that institutions with resources for extensive annotation efforts would realize incremental quality gains, but resource-constrained settings can achieve clinically meaningful performance with minimal annotation investment. This accessibility democratizes computational pathology, enabling smaller institutions to deploy AI-assisted diagnostics without prohibitive annotation infrastructure.

References

Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7 (1), December 2017. ISSN 2045-2322. doi:10.1038/s41598-017-17204-5. URL `http://dx.doi.org/10.1038/s41598-017-17204-5`.

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. ISSN 2041-1723. doi:10.1038/s41467-022-34025-x. URL `http://dx.doi.org/10.1038/s41467-022-34025-x`.

David A Gutman, Mohammed Khalilia, Sanghoon Lee, Michael Nalisnik, Zach Mullen, Jonathan Beezley, Deepak R Chittajallu, David Manthey, and Lee AD Cooper. The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research. *Cancer research*, 77(21):e75–e78, 2017.

Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, and Louis Wehenkel. Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, 32(9):1395–1401, January 2016. ISSN 1367-4803. doi:10.1093/bioinformatics/btw013. URL `http://dx.doi.org/10.1093/bioinformatics/btw013`.

André L S Meirelles, Tahsin Kurc, Jun Kong, Renato Ferreira, Joel Saltz, and George Teodoro. Effective and efficient active learning for deep learning-based tissue image analysis. *Bioinformatics*, 39(4), March 2023. ISSN 1367-4811. doi:10.1093/bioinformatics/btad138. URL `http://dx.doi.org/10.1093/bioinformatics/btad138`.

Runtian Miao, Robert Toth, Yu Zhou, Anant Madabhushi, and Andrew Janowczyk. Quick annotator: an open-source digital pathology based rapid image annotation tool. *The Journal of Pathology: Clinical Research*, 7(6):542–547, 2021.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

Manuel Stritt, Anna K. Stalder, and Enrico Vezzali. Orbit image analysis: An open-source whole slide image analysis tool. *PLOS Computational Biology*, 16(2):e1007313, February 2020. ISSN 1553-7358. doi:10.1371/journal.pcbi.1007313. URL `http://dx.doi.org/10.1371/journal.pcbi.1007313`.

Hang Su, Zhaozheng Yin, Seungil Huh, Takeo Kanade, and Jun Zhu. Interactive cell segmentation based on active and semi-supervised learning. *IEEE transactions on medical imaging*, 35(3):762–777, 2015.

Cédric Walker, Tasneem Talawalla, Robert Toth, Akhil Ambekar, Kien Rea, Oswin Chamian, Fan Fan, Sabina Berezowska, Sven Rottenberg, Anant Madabhushi, Marie Maillard, Laura Barisoni, Hugo Mark Horlings, and Andrew Janowczyk. Patchsorter: a high throughput deep learning digital pathology tool for object labeling. *npj Digital Medicine*, 7(1), June 2024. ISSN 2398-6352. doi:10.1038/s41746-024-01150-4. URL `http://dx.doi.org/10.1038/s41746-024-01150-4`.