Leveraging Foundation Models for Histological Grading in Cutaneous Squamous Cell Carcinoma using PathFMTools

Abdul Rahman Diab ABDUL@DS.DFCI.HARVARD.EDU
Dana-Farber Cancer Institute

Emily E. Karn EMILYEKARN@GMAIL.COM
Brigham and Women's Hospital

Renchin Wu RENCHIN_WU@DFCI.HARVARD.EDU
Dana-Farber Cancer Institute

Emily S. Ruiz ESRUIZ@BWH.HARVARD.EDU
Brigham and Women's Hospital

William Lotter LOTTERB@DS.DFCI.HARVARD.EDU
Dana-Farber Cancer Institute, Brigham and Women's Hospital, & Harvard Medical School

Abstract

Despite the promise of computational pathology foundation models, adapting them to specific clinical tasks remains challenging due to the complexity of whole-slide image (WSI) processing, the opacity of learned features, and the wide range of potential adaptation strategies. To address these challenges, we introduce **PathFMTools**, a lightweight, extensible Python package that enables efficient execution, analysis, and visualization of pathology foundation models. We use this tool to interface with and evaluate two state-of-the-art vision-language foundation models, CONCH and MUSK, on the task of histological grading in cutaneous squamous cell carcinoma (cSCC), a critical criterion that informs cSCC staging and patient management. Using a cohort of 440 cSCC H&E WSIs, we benchmark multiple adaptation strategies, demonstrating trade-offs across prediction approaches and validating the potential of using foundation model embeddings to train small specialist models. These findings underscore the promise of pathology foundation models for real-world clinical applications, with **PathFMTools** enabling efficient analysis and validation.

Keywords: computational pathology, cutaneous squamous cell carcinoma, foundation models, histological grading

Data and Code Availability The cSCC dataset utilized has not been IRB-approved for public release. The PathFMTools code is available at https://github.com/lotterlab/pathfmtools.

Institutional Review Board (IRB) This research was approved by the Mass General Brigham IRB (protocol 2023P002386)

1. Introduction

Foundation models have emerged as powerful tools in computational pathology, offering the ability to extract rich, domain-specific features from histopathology slides (Li et al., 2025; Bilal et al., 2025). By training on large collections of digitized H&E slides with self-supervised learning, these models learn generalizable morphological patterns that can be leveraged for downstream applications even when data is scarce. Nevertheless, adapting foundation models to specific research or clinical contexts presents several challenges. First, their use often involves complex, disjointed pipelines spanning data pre-processing, model inference, and downstream analysis of hundreds or thousands of gigapixel-scale whole-slide images (WSIs). Second, the feature spaces learned by foundation models remain opaque, necessitating efficient probing tools to facilitate biological validation.

Third, the wide array of potential adaptation strategies — ranging from zero-shot inference to supervised transfer learning — underscores the importance of benchmarking studies that compare these approaches in task-specific settings.

One particularly important task for many types of cancer is histological grading. Grading systems assess how differentiated tumor cells appear relative to normal cells — a key indicator of tumor aggressiveness and clinical risk. In cutaneous squamous cell carcinoma (cSCC), a highly prevalent skin cancer with over one million diagnoses annually in the United States, tumor grade is a critical component of risk stratification (Que et al., 2018; Thompson et al., 2016). Although most cSCC tumors are successfully managed with local excision, a subset of patients experience metastasis and/or disease-specific death (Schmults et al., 2013). Notably, a recent large, multi-institutional study found that among several existing risk factors, tumor grade was the strongest predictor of patient outcomes (Jambusaria-Pahlajani et al., 2025). Nevertheless, grading remains a subjective and error-prone process, exhibiting limited reproducibility and high inter- and intra-rater variability (Nash et al., 2022). Based on this, the American Joint Committee on Cancer has excluded differentiation from the 8th edition staging for cSCC despite its prognostic significance, though it is included in the popular BWH staging system. Thus, the clinical significance of cSCC grading and the associated challenges highlight an important opportunity to assess the utility of foundation models.

In this work, we introduce **PathFMTools**, a lightweight Python package that enables efficient exploration and analysis of computational pathology foundation models. We leverage the package to evaluate two state-of-the-art vision-language foundation models, CONCH (Lu et al., 2024) and MUSK (Xiang et al., 2025), on the task of cSCC grading. Using a curated cohort of 440 cSCC WSIs, we benchmark multiple adaptation strategies and explore the biological relevance of the learned features.

2. Related Work

2.1. Foundation Models in Computational Pathology

Numerous foundation models have been developed for computational pathology, and the available options differ in their training datasets, self-supervised learning strategies, and model outputs (Li et al., 2025; Bilal et al., 2025). The most common class of models operates at the patch level, where fixed-size crops (e.g. 224×224 pixels) from H&E-stained WSIs are used as inputs to generate feature embedding vectors. These embeddings can subsequently be used for a variety of downstream tasks, including slide-level prediction via aggregation across all patches.

Two predominant self-supervised learning paradigms have emerged for training such models: (1) vision-language contrastive learning and (2) vision-only self-distillation with masked data augmentation. Vision-language contrastive learning uses paired image and text inputs to learn visual and textual representations in a shared space (Radford et al., 2021). This approach enables zero-shot classification, where image embeddings are compared to text embeddings to perform inference without task-specific training. In contrast, popular self-distillation methods such as DINOv2 (Oquab et al., 2024) rely solely on image data, encouraging the model to learn stable representations under data augmentation and masking.

Given the zero-shot capabilities of vision-language models, we focus our analysis on two publicly-available pathology vision-language foundation models: CONCH and MUSK. CONCH was trained on approximately 1.17 million image-caption pairs sourced from both public and private datasets (Lu et al., 2024), and MUSK was trained on over 50 million pathology images and over one billion pathology-related text tokens from publicly-available sources (Xiang et al., 2025). Although our analyses are centered on these two models, the software package we introduce is model-agnostic and compatible with a broad range of foundation model architectures.

2.2. Software Packages for Computational Pathology

Growing interest in computational pathology has led to the development of several open-source software packages in the field. Notable examples include TIAToolbox (Pocock et al., 2022), STAMP (El Nahhas et al., 2024), TRIDENT (Zhang et al., 2025), and PathML (Rosenthal et al., 2022), which vary according to functionality and design. Some earlier packages (e.g., TIAToolbox) support tasks such as cell and tissue segmentation but lack native compatibility with pathology foundation models. More recent tools (e.g., TRIDENT) enable foundation model embedding gen-

eration but place less emphasis on downstream analysis and visualization. PathFMTools offers both built-in embedding generation and exploratory analysis tailored to pathology foundation models, including unique text embedding and zero-shot classification modules for vision–language models.

2.3. cSCC Grading

Although no standardized grading system for cSCC currently exists (Nash et al., 2022), tumors are typically categorized as well, moderately, or poorly differentiated based on their histological appearance. In a recent multinational cohort study comprising over 20,000 tumors, histological grade emerged as the most prognostic factor for poor clinical outcomes (Jambusaria-Pahlajani et al., 2025), underscoring its clinical relevance in risk stratification.

Despite its importance, cSCC grading is highly subjective. Reported inter-rater agreement remains moderate, with Cohen's kappa values in the range of $\kappa \sim 0.53$-0.55 (Nash et al., 2022). This variability highlights the need for more objective, quantitative grading approaches. Recent work has begun to explore machine-learning based solutions to this problem. For example, Choudhary et al. (2023) developed and tested a weakly-supervised model trained from scratch to estimate tumor grade using 718 H&E-stained WSIs from a single clinical site.

3. Methods

3.1. The PathFMTools Package

Here, we present **PathFMTools**, a lightweight, extensible package for running pathology foundation models and interfacing with the embeddings that they produce.

3.1.1. Command line tools for efficient embedding generation

The **PathFMTools** command line interface provides a simple and scalable entry point for executing core computational pathology workflows on collections of whole-slide images, enabling users to preprocess slides and generate foundation model embeddings using a single command.

To support large-scale data processing, the pipeline includes built-in support for automatic parallelization to an arbitrary number of workers. This scalability is enabled by the package's low memory footprint, achieved through efficient chunked processing of WSIs. Additionally, the workflow can be separated into distinct stages: a CPU-intensive preprocessing step and a GPU-dependent embedding generation step, facilitating the efficient use of computational resources in environments with limited or shared GPU availability.

3.1.2. Modular Python design

The command line interface is built upon the Slide class, which is the primary component that users working within Python will use. This class provides a unified abstraction for interacting with individual WSIs, encapsulating both the raw image data and any intermediate representations and outputs produced during analysis.

The class supports core computational pathology operations, such as tissue segmentation and tiling (Figure 1). By default, Otsu segmentation (Otsu, 1979) is utilized, but the modular design enables custom methods. The segmentation and tiling operations are optimized to run on localized regions of the slide, enabling memory-efficient processing of gigapixel-scale images on standard consumer hardware.

3.1.3. Built-in analysis and visualization

In addition to preprocessing, the Slide class interfaces directly with foundation models through an encapsulated model class. This design enables patch-level embedding generation and zero-shot analysis with vision–language prompts (Figure 1), supporting downstream tasks such as clustering and multiple-instance learning. The package also implements native K-means clustering to aggregate embeddings across slides, producing both cluster centroids and discrete patch assignments.

The Slide class further provides visualization tools for exploratory analysis, including high- or low-resolution slide rendering, contextual tile inspection, random sampling, and heatmaps of zero-shot scores that map image–text similarity. These utilities support both quantitative analysis and qualitative interpretation of model behavior.

In addition to the Slide class, **PathFMTools** also provides the SlideDataset class, a custom subclass of the Pytorch Dataset object. This class enables the efficient integration of Slide objects into deep learning pipelines. It supports both pixel-based and

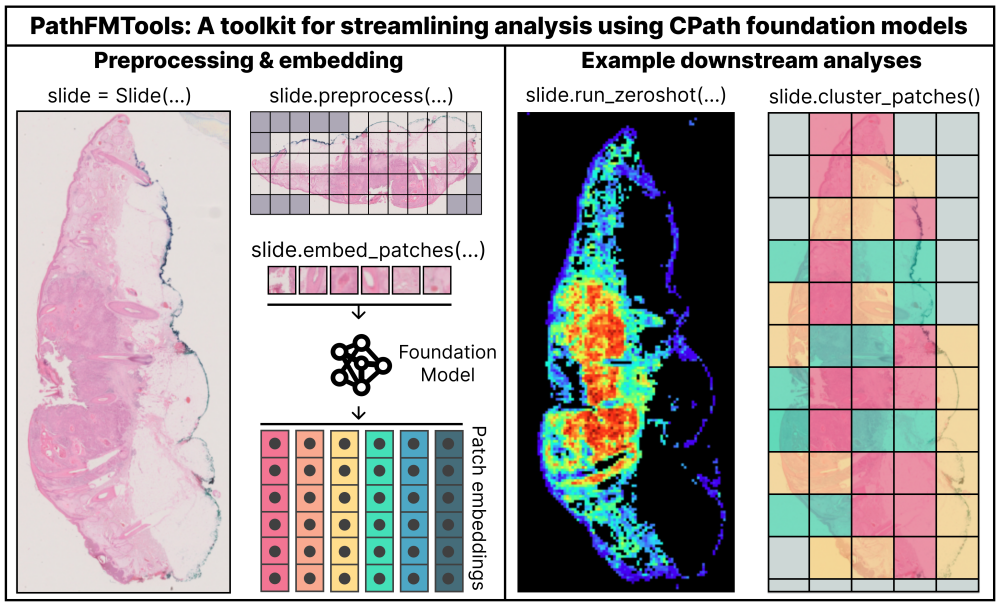

Figure 1: Overview of core interactions supported by the **PathFMTools** package. The package implements a modular end-to-end pipeline that facilitates embedding generation and exploratory analysis, including zero-shot similarity visualization and clustering of embeddings. See A.7 for code examples.

embedding-based inputs and handles memory management and indexing internally, streamlining model training workflows.

3.2. cSCC Analysis

To evaluate how pathology foundation models might support real-world clinical workflows, we conducted a series of analyses reflecting common tasks that clinicians and researchers might wish to perform using model-generated embeddings. These tasks include zero-shot classification based on textual descriptors, unsupervised clustering for morphologic feature discovery, and weakly supervised classification using slide-level labels. Each approach is designed both to probe the different ways that the user can interact with foundation models — ranging from minimal supervision to deep learning model training — and to assess the utility of such models for histological grading of cSCC slides. To evaluate each approach, we formulated a binary classification task according to whether each slide corresponded to a well-differentiated ('0') or moderately/poorly differentiated ('1') tumor, though we also consider different thresholds in Appendix A.4.

3.2.1. Patch & text embedding

Each WSI in the dataset was tiled into non-overlapping (448 × 448) pixel patches at 40× magnification. To exclude background regions, we applied Otsu thresholding (Otsu, 1979) to the distribution of mean grayscale pixel intensities across the slide's patches. Patches passing the tissue threshold were then downsampled to (224 × 224) pixels, the expected size of the studied foundation models, and passed through the respective model-specific preprocessing pipelines.

We generated patch-level embeddings using two publicly-available foundation models: CONCH and MUSK. Each model can produce two types of embeddings per patch: one optimized for downstream machine learning tasks (e.g. classification, regression), and one optimized for zero-shot tasks via alignment with pretrained language encoders. We use the machine learning-optimized embeddings for all analyses detailed in this work, except for the zero-shot approach. We generated text embeddings using the respective language encoders and tokenization pipelines provided by each foundation model.

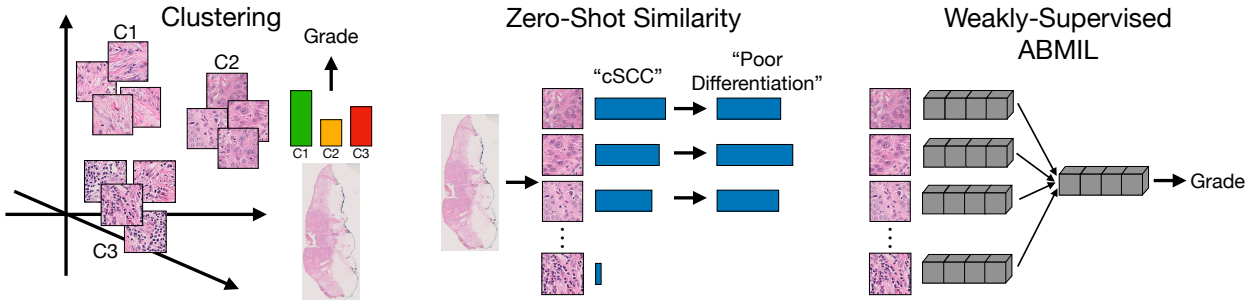

Figure 2: Summary of analytical approaches. Three different approaches were used to analyze the foundation model embeddings and their ability to assist with cSCC grading. Left: K-means clustering in the embedding space. Middle: A zero-shot classification task was formulated by first classifying each patch in a WSI as 'cSCC' or 'non-neoplastic' based on similarity between the patch embedding and text embeddings of these terms. Patches classified as cSCC were subsequently assessed for their similarity with text embeddings for well and poor differentiation. Right: Attention-based multiple instance learning (ABMIL) based on the patch embeddings was employed for a weakly-supervised benchmark.

3.2.2. Zero-shot histological grading

To assess the zero-shot capabilities of the foundation models in estimating WSI-level grading, we formulated a two-stage inference procedure using the language-aligned embeddings from CONCH and MUSK. In the first stage, we identified candidate cancer regions within each WSI by computing the cosine similarity between each patch embedding and the language embeddings of the phrases "cutaneous squamous cell carcinoma" and "non-neoplastic". Each patch was thus assigned two similarity scores (logits), and we retained only those patches for which the similarity to "cutaneous squamous cell carcinoma" exceeded that to "non-neoplastic", indicating that the patch was more likely to be neoplastic according to the model. Formally, for a slide S_i with m patches $S_i = \{p_1^{(i)}, \ldots, p_m^{(i)}\}$, we define the set of cancer-predicted patches as

$$C_i := \{p_j^{(i)} \in S_i \ : \ P(\text{cancer} \mid p_j^{(i)}) > 0.5\}$$

In the second stage, we computed a grading score for each of retained cancer patches C_i that reflects its differentiation on a continuum from well-differentiated to poorly-differentiated. We first computed the cosine similarity between each patch embedding and the language embeddings for the phrases "poorly differentiated" and "well differentiated". The grading score was then formulated as the difference in the two similarity scores for each patch. Formally given a patch k with embedding $z_k^{(i)}$ and text embeddings z_{poor} and z_{well}, the patch-level score $s_k^{(i)}$ is defined as

$$s_k^{(i)} := \cos(z_k^{(i)}, \ z_{poor}) - \cos(z_k^{(i)}, \ z_{well})$$

Intuitively, a high value for $s_k^{(i)}$ indicates that the pixel embedding for patch k is more aligned with features associated with poor differentiation. Finally, we aggregate the patch-level scores within each slide into a slide-level score s_i by taking the maximum over the cancer patch grading scores

$$s_i := \max_{k \in C_i} \ s_k^{(i)}$$

This max-pooling approach reflects the clinical intuition that tumor grade is often determined by the most poorly-differentiated region within the tissue. We also evaluated alternative approaches for computing slide-level scores, including varied patch-level aggregation methods and prompt ensembling, which we present in Appendix A.6.

3.2.3. K-means clustering of patch embeddings

We applied K-means clustering with 25 clusters, selected based on silhouette analysis, to all patch embeddings pooled across the dataset, assigning each patch to one of 25 discrete groups. Each slide was then represented by a 25-dimensional normalized histogram, where the j^{th} component quantifies the proportion of patches from that slide assigned to cluster

j. This cluster histogram encodes the distribution of morphologic patterns captured by the foundation model across each slide. To evaluate the predictive utility of these representations, we applied two strategies:

- Univariate AUROC: For each component of the 25-D summary vectors, we compute the area under the receiving operator characteristic curve (AUROC) for predicting the binarized histological grade. This provides an estimate of the standalone predictive power of individual clusters.
- Multivariate logistic regression: We train an ℓ_2-regularized logistic regression model using the full 25-dimensional cluster histogram as input. This evaluates the collective predictive power of the cluster distribution.

3.2.4. Attention-based multiple instance learning

To further assess the utility of patch-level embeddings for predicting histological grade, we implemented an attention-based deep multiple instance learning (ABMIL) framework (Ilse et al., 2018). In this setting, each WSI is treated as a bag of instances (patch embeddings), and the model is trained to aggregate information across patches to produce a slide-level prediction without requiring patch-level labels.

We used a gated single-headed attention mechanism to compute a weighted aggregation of patch embeddings into a fixed-dimensional slide-level representation (see A.2 for details). This aggregated embedding is passed to a lightweight multilayer perceptron consisting of two hidden layers with ReLU activations (64 and 32 neurons, respectively), followed by a sigmoid-activated scalar output node that predicts the tumor differentiation. We train the ABMIL model end-to-end using binary cross-entropy loss on the binarized grade labels (see A.3 for training details). We also visualize the attention scores assigned to individual patches.

3.2.5. Cohort

We curated the data used in this study from a U.S. academic hospital system, sampling 440 cSCC tumors from the system's tumor registry and scanning the associated H&E slides for this work. Given the cost of retrieval and digitization, our data collection procedure consisted of retrieving all moderately-differentiated and poorly-differentiated tumors available in the registry, in addition to a random subset of the well-differentiated tumors, as high-grade tumors are comparatively rare and valuable for model evaluation and development. Differentiation level was determined based on the associated pathology report, and if not explicitly mentioned in the report, well-differentiated was assumed based on clinical convention (Jambusaria-Pahlajani et al., 2025). The resulting dataset contained 239 well-differentiated tumors, 99 moderately-differentiated tumors, and 102 poorly-differentiated tumors (Table 1). The majority of slides corresponded to diagnostic biopsies, with some slides corresponding to excisional biopsies. Only one slide was scanned per sampled tumor when multiple slides were available, which was selected randomly from the available slides. The WSIs were digitized at 40x magnification using a Hamamastu NanoZoomer digital slide scanner. See A.1 for a detailed summary of the cohort, including patient and tumor characteristics.

3.2.6. Statistical analysis

AUROC was used as the primary assessment metric. Sensitivity and specificity at selected operating points for the ABMIL models are also presented in Appendix A.4. For all analyses other than zero-shot performance quantification, we conducted five-fold cross-validation at the patient level and report performance as the average across folds (see A.3 for additional details). For the logistic regression models based on K-means cluster proportions, both the clustering and regression were fit per each fold before evaluation on the test split for the fold.

4. Results

We evaluated the utility of foundation model embeddings for cSCC grading using three approaches: zero-shot classification, unsupervised clustering with K-means, and ABMIL.

4.1. Zero-shot histological grading

We leveraged the language-aligned embeddings from CONCH and MUSK to assess zero-shot grading performance without task-specific fine-tuning. Using phrase-based similarity scores aggregated at the slide level (see Methods), this approach achieved AUROC values of 0.63 and 0.60 (Figure 3), for CONCH and MUSK respectively, indicating performance moderately above chance.

4.2. K-means clustering of patch embeddings

Unsupervised clustering provided an alternative, interpretable representation of slide features by discretizing patches into clusters that capture general morphological patterns (Figure 4, 6, 7). By clustering all patch embeddings into 25 clusters, we obtained a normalized histogram of cluster proportions for each slide. We then computed univariate AUROCs using individual histogram components as scores and multivariate AUROC by training logistic regression models on the histograms.

Univariate AUROC analysis demonstrated that individual clusters can be surprisingly informative, achieving AUROCs of ∼0.75 for the most-discriminative clusters. Inspection of these clusters yielded mostly intuitive findings (Figure 4). For associations with the high-grade label, both CONCH and MUSK exhibited clusters of moderately/poorly differentiated cancer cells. CONCH also exhibited a strongly-discriminative cluster corresponding to desmoplasia, a known risk factor potentially linked to higher grade (Que et al., 2018). However, one of the MUSK clusters most associated with high-grade cSCC corresponded to sweat glands, an association that is unclear. Clusters associated with low-grade tumors frequently contained keratin fragments, consistent with the keratinization typical of well-differentiated cSCC.

Multivariate logistic regression on all 25 cluster frequencies did not substantially exceed the performance of the best univariate clusters (Figure 3), though it outperformed the zero-shot strategy and was assessed in a stricter cross-validation setting.

4.3. Attention-based multiple instance learning (ABMIL)

Weakly supervised ABMIL models outperformed zero-shot and clustering baselines, reaching AUROCs of 0.81 (95% CI: 0.71–0.90) and 0.78 (95% CI: 0.68–0.87) for CONCH and MUSK, respectively (Figure 3b, Table 2). Attention heatmaps highlighted regions most influential to the predictions, revealing both overlaps and differences with informative patches identified in the zero-shot and clustering analyses (Figure 5). Training on finer-grained grade distinctions showed similar trends (Appendix A.4).

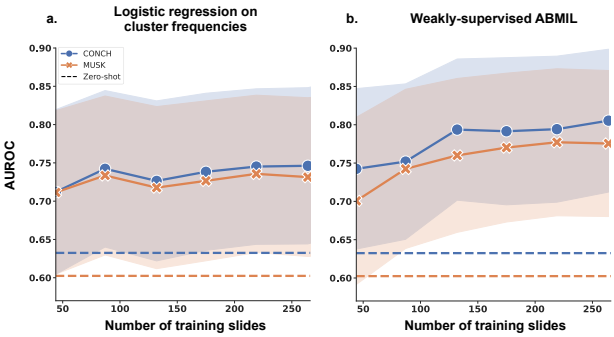

Figure 3: Grade classification performance. The dashed lines represent the zero-shot classification performance leveraging the vision-language capabilities of CONCH and MUSK. The markers represent the mean test AUROC of models trained using CONCH and MUSK patch embeddings across the 5 test splits for each number of training slides. Error bars represent DeLong 95% confidence intervals. **(a)** Performance of logistic regression trained on the cluster frequency vectors. **(b)** Performance of ABMIL models.

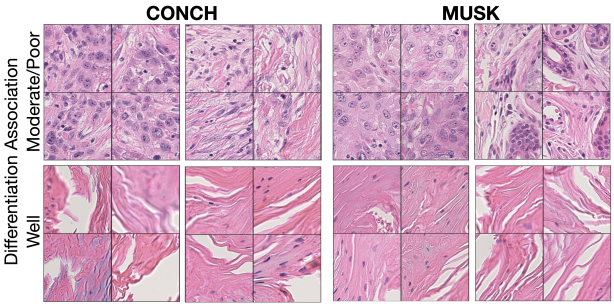

Figure 4: Examples of clusters most associated with histological grade for CONCH and MUSK. For each cluster, the patches whose embeddings are most similar to the cluster centroid are shown.

4.4. Data efficiency and model generalization

We further analyzed how model performance scaled with training data. As expected, accuracy improved with larger training sets, with ABMIL models showing greater gains than logistic regression (Figure 3). Notably, ABMIL required relatively few training

samples (<100) to surpass logistic regression performance, particularly when using CONCH embeddings.

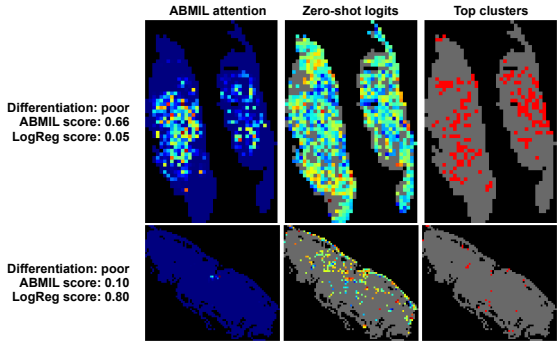

Figure 5: Comparison of slide-level predictions and spatial patterns across classification methods for two poorly differentiated cSCC tumors. Each row shows one tumor where ABMIL and logistic regression (LogReg) produced divergent scores: in the top row, ABMIL assigned a high score (0.66) while LogReg assigned a low score (0.05); in the bottom row, the reverse pattern was observed (0.10 vs. 0.80). Left: ABMIL attention heatmaps showing normalized attention weights per patch. Middle: Zero-shot logits, defined as the difference in cosine similarity between the phrases "poorly differentiated" and "well differentiated", shown only for patches classified as cSCC. Right: Patch-level cluster maps highlighting the four clusters most strongly associated with poor differentiation (red) based on univariate AUROC analysis (Methods). Black indicates background, and gray indicates tissue excluded by the corresponding method. All visualizations are based on CONCH embeddings.

5. Discussion

In this manuscript, we introduced **PathFMTools**, a modular and extensible Python package designed to facilitate the efficient execution, exploration, and visualization of computational pathology foundation models. Using this package, we systematically evaluated two vision-language foundation models, CONCH and MUSK, on the clinically valuable task of histological grading in cutaneous squamous cell carcinoma. Our analysis explored a variety of model adaptation approaches, including zero-shot inference, unsupervised clustering, and weakly-supervised learning, demonstrating clear performance differences and tradeoffs between these methodologies.

We found that zero-shot classification using vision-language embeddings with no task-specific fine-tuning was moderately effective, achieving better-than-chance performance. This highlights the flexibility of vision-language foundation models, especially in contexts where annotated data is limited; however the performance was insufficient for the tested task and formulation.

Our clustering analysis provided interpretable insights into the morphological patterns learned by foundation models. We demonstrated that individual clusters could be independently informative for predicting histological differentiation, suggesting that foundation models can effectively capture biologically meaningful features. Notably, several identified clusters corresponded to histological features that are known to be relevant to cSCC grading, such as clusters enriched for keratin fragments and poorly differentiated tumor cells. Conversely, some clusters captured imaging artifacts or less relevant features (e.g., sweat glands), underscoring the importance of quality control in computational pathology workflows and the risk of data-driven confounders.

Attention-based multiple instance learning (ABMIL) emerged as the most effective approach for leveraging foundation model embeddings, outperforming both zero-shot and clustering-based logistic regression methods even with minimal training data. Importantly, ABMIL attention heatmaps also provide interpretability at the patch level, which can be interactively visualized with **PathFMTools**.

Across all analyses, both CONCH and MUSK followed similar performance trends: moderate zero-shot accuracy, improved performance with clustering, and the strongest results with ABMIL. While the absolute performance was comparable between models, CONCH consistently outperformed MUSK at each training set size in the ABMIL setting.

A critical consideration in evaluating foundation models in downstream tasks is what constitutes "good" performance. While ABMIL demonstrated the highest overall performance, it is unclear whether the observed levels (∼0.80 AUROC) are sufficient for clinical utility. The high inter-rater variability of cSCC grading may suggest that this level is near the

inter-reader noise ceiling, but future studies directly comparing human graders and AI in the same setting would be necessary. Beyond AI standalone performance, a fundamental consideration is how AI assistance affects clinician performance and how AI fits into the clinical workflow (McNamara et al., 2024). Clinical AI deployment involves many steps, ranging from regulatory considerations and UI engineering, thus our analysis and software package should be viewed as informing research and validation, rather than actual clinical use.

Limitations Our study has several limitations. Because the analysis relies on data from a single clinical system, the observed performance may not generalize to other sites or populations. We did not explicitly evaluate sensitivity to staining differences and used each model's default preprocessing; however, our software allows users to easily define and test custom preprocessing pipelines. Accordingly, the results should be viewed primarily as a demonstration of how our tools can help researchers explore and adapt foundation models for specific clinically relevant tasks. Finally, the software is intended for technically oriented researchers and practitioners, rather than as a no-code or GUI-based tool for users with limited programming experience.

Conclusions As foundation models continue to improve and use cases expand, our findings illustrate how **PathFMTools** streamlines exploration and adaptation of these models, highlighting their promise in addressing critical challenges in histopathology, such as the subjective nature of tumor grading. Our quantitative findings can also help inform the optimal adaption strategies to other clinically-important tasks.

Acknowledgments

This research is supported by National Institute of Biomedical Imaging and Bioengineering award R21EB035247 and the National Library of Medicine award R01LM014775.

References

Mohsin Bilal, Aadam, Manahil Raza, Youssef Altherwy, Anas Alsuhaibani, Abdulrahman Abduljabbar, Fahdah Almarshad, Paul Golding, and Nasir Rajpoot. Foundation models in computational pathology: A review of challenges, opportunities, and impact, 2025. URL https://arxiv.org/abs/2502.08333.

Anirudh Choudhary, Angelina Hwang, Jacob Kechter, Krishnakant Saboo, Blake Bordeaux, Puneet Bhullar, Nneka Comfere, David DiCaudo, Steven Nelson, Emma Johnson, Leah Swanson, Dennis Murphree, Aaron Mangold, and Ravishankar K Iyer. RACR-MIL: Weakly supervised skin cancer grading using rank-aware contextual reasoning on whole slide images. *arXiv [cs.CV]*, August 2023.

Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nature Protocols*, Sep 2024. ISSN 1750-2799. doi: 10.1038/s41596-024-01047-2. URL https://doi.org/10.1038/s41596-024-01047-2.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv [cs.LG]*, February 2018.

Anokhi Jambusaria-Pahlajani, Vincent Jeanselme, David M Wang, Nina A Ran, Emily E Granger, Javier Cañueto, David G Brodland, David R Carr, Joi B Carter, John A Carucci, Kelsey E Hirotsu, Emily E Karn, Shlomo A Koyfman, Aaron R Mangold, Fabio Muradás Girardi, Kathryn T Shahwan, Divya Srivastava, Allison T Vidimos, Tyler J Willenbrink, Ashley Wysong, William Lotter, and Emily S Ruiz. riSCC: A personalized risk model for the development of poor outcomes in cutaneous squamous cell carcinoma. *J. Am. Acad. Dermatol.*, February 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv [cs.LG]*, December 2014.

Dong Li, Guihong Wan, Xintao Wu, Xinyu Wu, Ajit J. Nirmal, Christine G. Lian, Peter K. Sorger, Yevgeniy R. Semenov, and Chen Zhao. A survey on computational pathology foundation models: Datasets, adaptation strategies, and evaluation tasks, 2025. URL https://arxiv.org/abs/2501.15724.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30: 863–874, 2024.

Stephanie L McNamara, Paul H Yi, and William Lotter. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. *NPJ Digit Med*, 7(1):80, March 2024.

Jessica Nash, Kathryn T Shahwan, Catherine Chung, Nadia Abidi, Yevgeniya Gokun, Xueliang Pan, and David R Carr. Grading of differentiation in cutaneous squamous cell carcinoma: Evaluation of interrater and intrarater reliability. *J. Am. Acad. Dermatol.*, 87(4):895–897, October 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Johnathan Pocock, Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Srijay Deshpande, Giorgos Hadjigeorghiou, Adam Shephard, Raja Muhammad Saad Bashir, Mohsin Bilal, Wenqi Lu, David Epstein, Fayyaz Minhas, Nasir M Rajpoot, and Shan E Ahmed Raza. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, 2(1): 120, sep 2022. ISSN 2730-664X. doi: 10.1038/s43856-022-00186-5. URL https://www.nature.com/articles/s43856-022-00186-5.

Syril Keena T Que, Fiona O Zwald, and Chrysalyne D Schmults. Cutaneous squamous cell carcinoma: Incidence, risk factors, diagnosis, and staging. *J. Am. Acad. Dermatol.*, 78(2):237–247, February 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jacob Rosenthal, Ryan Carelli, Mohamed Omar, David Brundage, Ella Halbert, Jackson Nyman, Surya N. Hari, Eliezer M. Van Allen, Luigi Marchionni, Renato Umeton, and Massimo Loda. Building tools for machine learning and artificial intelligence in cancer research: Best practices and a case study with the pathml toolkit for computational pathology. *Molecular Cancer Research*, 20(2):202–206, 02 2022. ISSN 1541-7786. doi: 10.1158/1541-7786.MCR-21-0665. URL https://doi.org/10.1158/1541-7786.MCR-21-0665.

Chrysalyne D Schmults, Pritesh S Karia, Joi B Carter, Jiali Han, and Abrar A Qureshi. Factors predictive of recurrence and death from cutaneous squamous cell carcinoma: a 10-year, single-institution cohort study. *JAMA Dermatol.*, 149(5): 541–547, May 2013.

Agnieszka K Thompson, Benjamin F Kelley, Larry J Prokop, M Hassan Murad, and Christian L Baum. Risk factors for cutaneous squamous cell carcinoma recurrence, metastasis, and disease-specific death: A systematic review and meta-analysis. *JAMA Dermatol.*, 152(4):419–428, April 2016.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, pages 1–10, January 2025.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating

data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Appendix A.

A.1. Cohort Summary

Table 1: Cohort Summary. IQR: interquartile range.

		Value (%)
Total	Tumor count	440
	Patient count	412
Patient Sex	Male	280 (64)
	Female	160 (36)
Age	Median (IQR)	74 (66, 82)
Tumor Location	Head/neck	215 (49)
	Trunk/extremeties	224 (51)
	Unknown	1 (<1)
Diameter	Median (IQR), mm	13 (8, 23)
Invasion Depth	Dermis	217 (49)
	Subcutaneous fat	47 (11)
	Beyond subcut. fat	41 (9)
	Unknown	135 (31)
Differentiation	Well	239 (54)
	Moderate	99 (23)
	Poor	102 (23)

A.2. ABMIL approach

Given a set of n d-dimensional patch embeddings $\{z_1, \ldots, z_n\} \in \mathbb{R}^d$ for a given slide, the model computes attention weights $\alpha_i \in [0, 1]$ as

$$\alpha_i = \frac{\exp\left(w^T \tanh(V z_i^T) \odot \sigma(U z_i^T)\right)}{\sum_{j=1}^{n} \exp\left(w^T \tanh(V z_j^T) \odot \sigma(U z_j^T)\right)}$$

where σ is the sigmoid function, $V, U \in \mathbb{R}^{h \times d}$ and $w \in \mathbb{R}^h$ are learned parameters, $h = 256$ is the attention hidden dimensionality, and d is the dimension of the embedding produced by the foundation model.

The slide-level representation is then computed as the attention-weighted sum of patch embeddings

$$z_{slide} := \sum_{i=1}^{n} \alpha_i z_i \in \mathbb{R}^d$$

A.3. Hyperparameters & cross validation details

We trained the ABMIL models using the Adam optimizer (Kingma and Ba, 2014) and a cyclic learning rate that varied linearly from 5×10^{-5} to 5×10^{-4}, using a half-cycle length of 2 epochs. Logistic regression models were trained using ℓ_2 regularization with a penalty scaling factor of 1 and a maximum of 1,000 iterations.

To evaluate model performance, we performed 5-fold cross-validation at the patient level for both the ABMIL and logistic regression models. In each fold, 20% of the data was held out as a test set, with stratified sampling to preserve the proportion of positive and negative slides across splits and prevent different tumors from a single patient from appearing in both the development and test splits of a given fold. For ABMIL, an additional 20% of the data was held out as a validation set for monitoring model convergence.

We computed AUROC scores on the test set within each fold, and reported the mean across folds as the primary performance estimate. To quantify variability, we computed 95% DeLong confidence intervals for the AUROC point estimates. Furthermore, to assess data efficiency, we evaluated model performance as a function of training set size. For both ABMIL and logistic regression, we trained models on progressively larger subsets of the training data, using increments of 10% of the full dataset.

A.4. cSCC grading performance

When considering individual cSCC grades, the ABMIL models achieved the highest accuracy when distinguishing well- from poorly differentiated tumors, whereas classification between moderately and poorly differentiated tumors was more challenging but still yielded reasonable performance. Table 2 summarizes AUROC values across different grading cutoffs.

Positive class	Negative class	AUROC	
		CONCH	MUSK
Moderate	Well	**0.74**	0.71
Poor	Well	**0.87**	0.82
Poor	Moderate	**0.73**	0.66
Moderate & Poor	Well	**0.81**	0.78

Table 2: AUROC of ABMIL models using CONCH and MUSK embeddings for binary cSCC grading using different groups. The higher AUROC for each comparison is shown in bold.

Table 3 summarizes the sensitivity and specificity of the CONCH and MUSK models at three operating points: high-sensitivity (targeting 90%), high-specificity (targeting 90%), and balanced (target-

ing equal sensitivity and specificity). For each fold, thresholds were calibrated on the validation set and then applied to the corresponding test set. Point estimates and 95% confidence intervals (Adjusted Wald method) were obtained by aggregating predictions across all test folds.

Operating Point	CONCH	MUSK
High Sensitivity		
Sens	0.91 (0.89–0.92)	0.84 (0.83–0.85)
Spec	0.45 (0.43–0.46)	0.50 (0.48–0.52)
High Specificity		
Sens	0.59 (0.57–0.61)	0.46 (0.44–0.48)
Spec	0.89 (0.88–0.90)	0.84 (0.83–0.86)
Balanced		
Sens	0.74 (0.72–0.76)	0.74 (0.72–0.76)
Spec	0.74 (0.72–0.75)	0.64 (0.62–0.66)

Table 3: Sensitivity and specificity for CONCH and MUSK models at three operating points. Thresholds were calibrated per fold on validation data and evaluated on held-out test folds.

A.5. Runtime analysis

All experiments were conducted on a server equipped with an NVIDIA H100 GPU, an Intel Xeon Silver 4410Y CPU, running Linux, with data stored on a network file system (NFS). Table 4 summarizes the average time per slide for each major processing stage, measured across all slides in the cohort.

Step	Per slide (mean ± SD)
Segmentation	4.2 ± 3.8 s
Tiling	20.1 ± 25.7 s
CONCH embedding	23.2 ± 25.5 s
MUSK embedding	46.9 ± 57.0 s

Table 4: Per-slide runtime (mean ± SD) for each stage of the slide processing pipeline. Times were averaged across all slides.

A.6. Sensitivity to prompt choice for zero-shot analysis

The zero-shot results presented in the main text followed a clinically motivated strategy where we first identified patches with a higher logit for "cutaneous squamous cell carcinoma" than for "non-neoplastic", and then computed a slide-level score as the maximum logit difference between "poorly differentiated" and "well differentiated" across these patches. Here, we also present results using alternative aggregation methods rather than the maximum. Specifically, we (1) averaged the logit differences across all identified patches and (2) averaged them across the 10% of identified patches with the largest logit difference. Both methods yielded results comparable to the clinically motivated approach, with CONCH achieving AUROCs of 0.66 and 0.64, and MUSK achieving 0.58 and 0.61, respectively.

We also tested prompt ensembling, averaging logits over ten semantically varied prompt phrasings for each class (Table 5). Ensembling did not improve performance (AUROC ∼0.55 for both CONCH and MUSK), potentially reflecting the consistent terminology used in cSCC diagnosis.

A.7. Example code

With PathFMTools, common computational pathology workflows can be performed using a few lines of code, as highlighted in the following example:

```
# Instantiate one Slide object per WSI
slides = [Slide(fp) for fp in PATHS]
for slide in slides:
    # Apply preprocessing (segmentation + tiling)
    slide.preprocess(PATCH_SIZE, SEG_METHOD)
    # Run foundation model inference on slide patches
    slide.embed_patches(MODEL)
    # Run zero-shot classification on slide patches
    slide.classify_patches(MODEL, PROMPTS)
# Perform K-means clustering on all patches from
# all WSIs jointly
slide_group = SlideGroup(slides)
kmeans = KMeansPatchClusterer.fit_predict(slide_group, k=K)
```

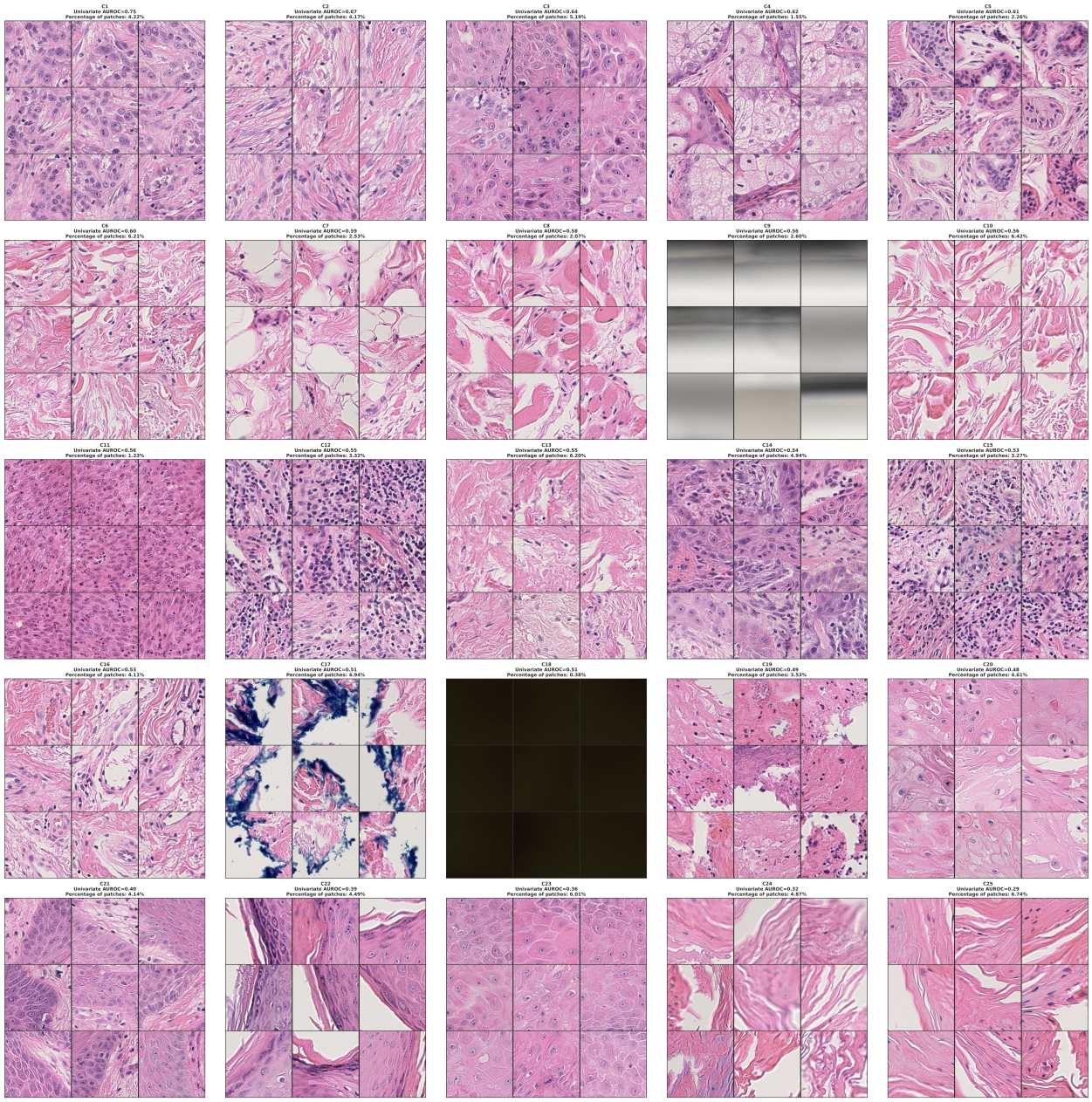

Figure 6: Representative patches from each of the 25 CONCH embedding clusters. The outer grid represents individual clusters, each containing an inner 3×3 grid of example patches nearest to the cluster centroid. Clusters are ordered left to right, top to bottom by decreasing univariate AUROC for predicting cancer grade (Methods), with AUROC values indicated above each cluster. The text annotation also specifies the percentage of total patches assigned to each cluster.

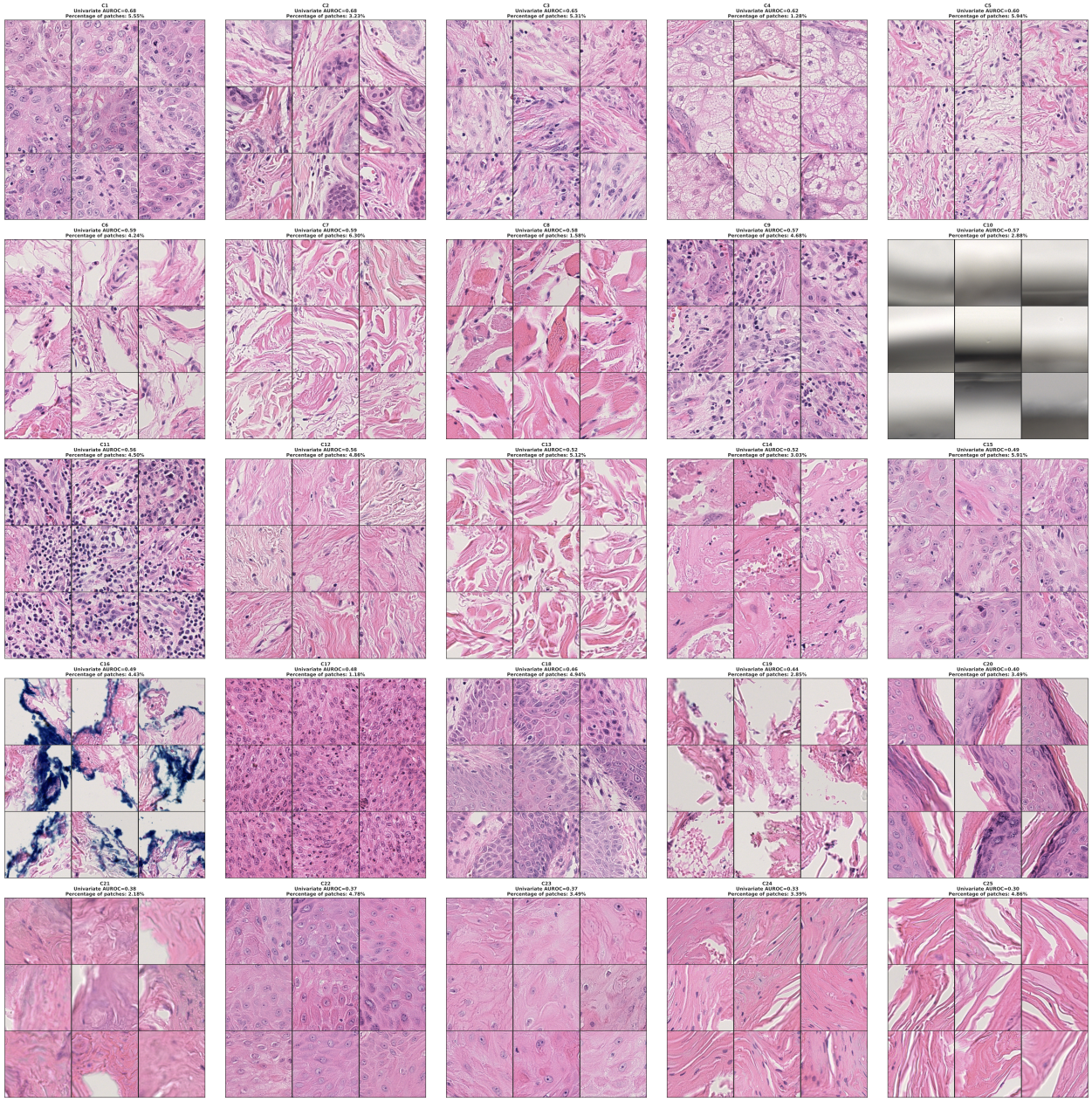

Figure 7: Representative patches from each of the 25 MUSK embedding clusters. The outer grid represents individual clusters, each containing an inner 3×3 grid of example patches nearest to the cluster centroid. Clusters are ordered left to right, top to bottom by decreasing univariate AUROC for predicting cancer grade (Methods), with AUROC values indicated above each cluster. The text annotation also specifies the percentage of total patches assigned to each cluster.

Label	**Prompt variants**
Cutaneous squamous cell carcinoma	cutaneous squamous cell carcinoma; cSCC; cutaneous SCC; squamous cell carcinoma of the skin; squamous carcinoma; cutaneous squamous carcinoma; H&E image of cutaneous squamous cell carcinoma; histopathology image of skin squamous cell carcinoma; tumor region consistent with cSCC; slide patch showing squamous carcinoma.
Non-neoplastic	non-neoplastic; benign tissue; normal skin; non-tumorous region; non-cancerous tissue; normal cells; H&E image of normal skin; tissue without neoplastic changes; non-lesional tissue patch; normal tissue.
Well differentiated	well differentiated; well-differentiated carcinoma; well-differentiated squamous cell carcinoma; well-organized tumor architecture; low-grade carcinoma; low grade; well differentiated tumor cells; tumor with minimal atypia; slide showing well-differentiated tumor cells; well-differentiated tumor.
Poorly differentiated	poorly differentiated; poorly-differentiated carcinoma; high-grade squamous cell carcinoma; undifferentiated tumor cells; disorganized tumor architecture; marked cellular atypia; aggressive poorly differentiated SCC; slide showing poorly differentiated cells; high grade; poorly differentiated tumor cells.

Table 5: Prompt variants used for each class.