Do Histopathological Foundation Models Eliminate Batch Effects? A Comparative Study

Jonah Kömen
Machine Learning Group
Technische Universität Berlin
BIFOLD*
Berlin, Germany

Hannah Marienwald
Machine Learning Group
Technische Universität Berlin
BIFOLD*
Berlin, Germany

Jonas Dippel
Machine Learning Group
Technische Universität Berlin
BIFOLD*
Aignostics
Berlin, Germany

Julius Hense
Machine Learning Group
Technische Universität Berlin
BIFOLD*
Berlin, Germany

Abstract

Deep learning has led to remarkable advancements in computational histopathology, e.g., in diagnostics, biomarker prediction, and outcome prognosis. Yet, the lack of annotated data and the impact of batch effects, e.g., systematic technical data differences across hospitals, hamper model robustness and generalization. Recent histopathological foundation models — pretrained on millions to billions of images — have been reported to improve generalization performances on various downstream tasks. However, it has not been systematically assessed whether they fully eliminate batch effects. In this study, we empirically show that the feature embeddings of the foundation models still contain distinct hospital signatures that can lead to biased predictions and misclassifications. We further find that the signatures are not removed by stain normalization methods, dominate distances in feature space, and are evident across various principal components. Our work provides a novel perspective on the evaluation of medical foundation models, paving the way for more robust pretraining strategies and downstream predictors.

1 Introduction

Increasingly complex tasks in digital histopathology are solved by deep learning (DL) models, e.g., disease diagnosis or subtyping [1, 2, 3, 4, 5, 6, 7], prediction of clinical or molecular biomarkers [8, 9, 10, 11], and outcome prognosis [12, 13, 14, 15]. While the performance of supervised end-to-end models is limited by the lack of labeled training data, self-supervised learning is currently adopted as a promising solution [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Pretrained a plethora of unlabeled histopathology images, foundation models extract features that can be utilized to achieve excellent prediction performances for a wide variety of downstream tasks, outperforming models trained from scratch or pretrained on natural images [20].

However, previous research [27] revealed significant differences in digital whole-slide images (WSIs) across tissue source sites (TSS), i.e., hospitals, labs, or biobanks. These disparities stem from differences in fixation and staining protocols, WSI scanners, and patient demographics, among others.

---

*Berlin Institute for the Foundations of Learning and Data, Berlin, Germany.
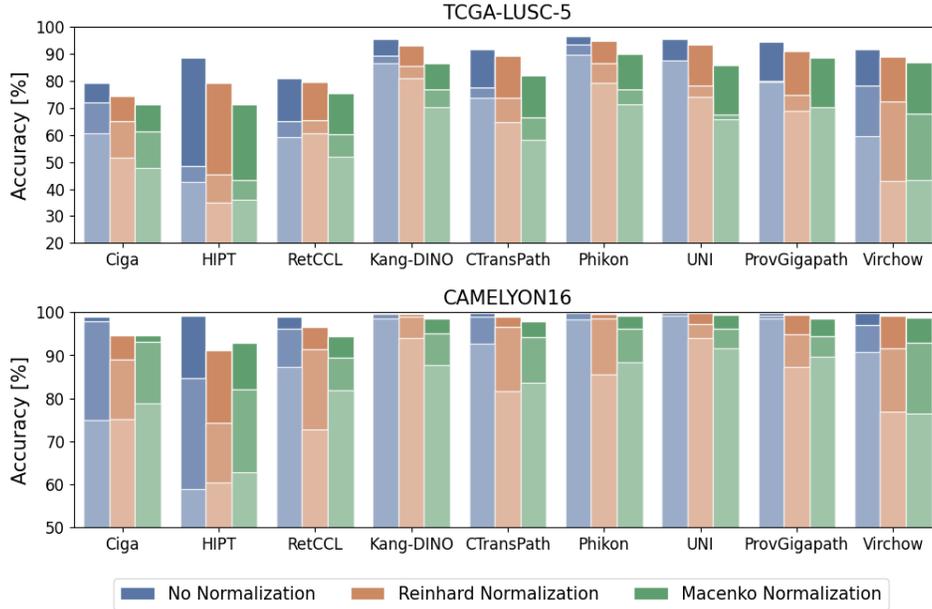
Figure 1: Accuracy scores for predicting the TSS of a patch from the feature embedding of different histopathological foundation models. We considered features of raw and stain-normalized patches [29, 30] from two datasets: TCGA-LUSC-5 (top) with 5 and CAMELYON16 (bottom) with 2 tissue source sites. The most transparent bars represent the nearest centroid classifier, the medium transparent k-nearest neighbors, and the non-transparent bars linear probing models. *Higher accuracy indicates stronger site signatures.* The experimental details are described in Section 4.1.

Such site-specific peculiarities, also called site signatures, may correlate with labels in downstream tasks, which creates potential batch effects. It was shown that DL models directly trained on WSI patches indeed employ such batch effects for solving supervised learning tasks, leading to biased predictions and compromised generalizability across tissue source sites. This lack of robustness is potentially dangerous in high-risk diagnostic systems arising in digital histopathology.

As histopathological foundation models are trained on an excess of diverse multi-site data sets, employ augmentation strategies during pretraining, and have been reported to generalize better to unseen data [24, 25, 26, 28], one may hypothesize that they are fully immune to data biases. In this work, we show that they are in fact *not*: all tested models are still susceptible to batch effects. Our key findings include:

- The feature embeddings of all foundation models contain site-specific signatures that can be accurately retrieved using linear methods. Interestingly, we observe the best-performing models according to the literature to exhibit the highest source site prediction accuracy. Figure 1 shows a preview of this result.
- Site-specific information acts as a batch effect. Correlation between site signatures and downstream task labels may result in biased predictions and misclassifications. Commonly applied stain normalization techniques like Reinhard [29] and Macenko [30] do not sufficiently reduce such biases.
- Distances between patches in feature space are more affected by discrepancies from the TSS rather than biological differences, which renders downstream algorithms resorting to the proximity of points, such as clustering, nearest neighbor search, or similar, as ill-posed.
- The linear directions of the largest variance of the features exhibit high site separability, confirming that batch effects are very accessible, easily misleading supervised downstream algorithms.
- Mitigating batch effects in feature space or during the training of foundation models is a promising research direction for improving the robustness and generalization of downstream predictors.
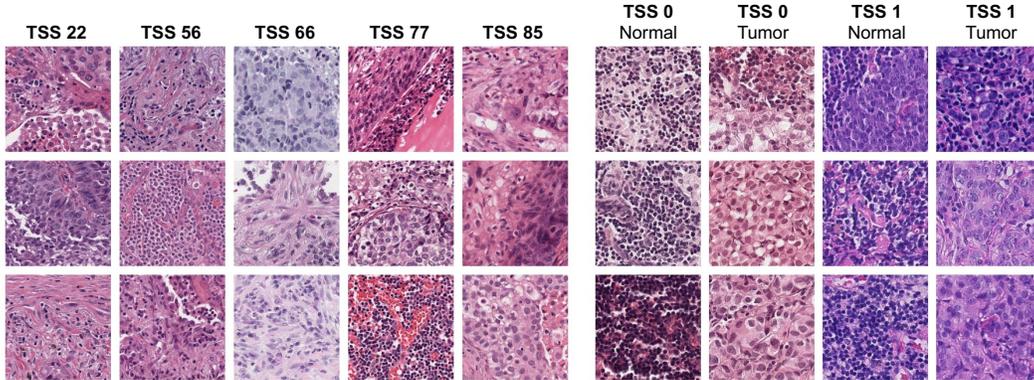
Figure 2: Arbitrarily chosen patches from the TCGA-LUSC-5 dataset (left) and the CAMELYON16 dataset (right). Differences in staining are apparent, e.g., see *TSS 66*, *TSS 0*, or *TSS 1*.

Our findings are based on several experiments that we describe in the remainder of the paper. After discussing relevant related work in Section 2, we provide a brief overview of the tested foundation models and the datasets we employ in Section 3. In Section 4, we expose batch effects in histopathological foundation models by identifying that the issuing site can be easily predicted from feature vectors and that this biases downstream tasks. We further characterize this batch effect by examining distances between patches in feature space and relating feature variance with site separability in Section 5.

2 Related work

Previous studies showed that supervised models are confounded by staining, scanner, and site-specific signatures [27, 31]. When the labels are correlated with these metadata, models possibly exploit them as shortcut/Clever Hans features, leading to catastrophic failure on new uncorrelated cases [32, 33, 34, 35].

Self-supervised learning models, trained on an abundance of diverse unlabeled data with different metadata, are proposed as a solution and observed to be more robust in some domains [36, 37]. In histopathology, scaling dataset and model sizes recently led to significant performance improvements on downstream tasks [24, 25, 23, 26]. However, to the best of our knowledge, it has not been investigated whether this also results in foundation models that eliminate batch effects. Existing works on evaluating histopathological foundation models focused on performance assessments on a variety of (clinical) downstream tasks [20, 28]. Further, Vaidya et al. [38] showed performance disparities across different demographic groups on subtyping and mutation prediction, which foundation models helped to reduce.

In contrast to performance or distribution shift evaluations, we systematically analyze to what extent hospital signatures are still encoded in foundation model representations and whether they lead to wrong predictions when correlated with prediction labels.

3 Datasets and models

Datasets

We focus on studying batch effects originating from site-specific signatures in the data, e.g., variations induced by differing cutting, staining, or scanning protocols across hospitals. We used carefully composed subsets of two public multi-site histopathology datasets for our experiments:

- **TCGA-LUSC-5**. The TCGA Lung Squamous Cell Carcinoma (TCGA-LUSC) project provides H&E-stained slides containing lung squamous cell tumors [39] from various tissue source sites. We selected five of the most contributing sites: the Mayo Clinic (*TSS 22*), the International Genomics Consortium (*TSS 56*), Indivumed (*TSS 66*), the Prince Charles

3

Hospital (*TSS 77*), and Asterand Bioscience (*TSS 85*). We sampled one diagnostic primary tumor FFPE slide ("-DX1") from up to 50 unique patients per TSS, respectively.

- **CAMELYON16**. The CAMELYON16 dataset [1] comprises 400 sentinel lymph node slides from women with breast cancer, some of which contain precisely annotated regions of lymph node metastasis regions of different sizes. The dataset was issued by two hospitals in the Netherlands: the Radboud University Medical Center (*TSS 0*) and the University Medical Center Utrecht (*TSS 1*).

We extracted patches from the slides of 256×256 pixels without overlap at 20x magnification (0.5 microns per pixel). We identified and excluded background patches via Otsu's method [40] on slide thumbnails and applied a patch-level minimum standard deviation of 8. Each patch was labeled with the tissue source site of its slide. For CAMELYON16, we additionally labeled those patches fully inside the metastasis annotations as 'tumor' and those fully outside as 'normal'. The resulting dataset statistics and example patches are displayed in Table 1 and Figure 2.

We additionally aimed to assess the effect of stain normalization in combination with the foundation models, which is arguably the most popular approach intending to mitigate batch effects in histopathology (see e.g. [41]). We assessed the widely used Reinhard [29] and Macenko [30] techniques, and applied them to all patches in our datasets. As normalization targets, we used the average statistics of 500 patches randomly sampled from TCGA-LUSC-5. Example patches are displayed in Figure 7.

As we were primarily interested in non-informative signatures like scanning and staining characteristics, we aimed for homogeneous distributions of biological metadata (e.g., diagnosis, resection site, or stage) across tissue source sites. In both datasets, all slides come from the same organ and disease entity, respectively. Additionally, we report the statistics of available clinical metadata per tissue source sites for TCGA-LUSC-5 in Table 4. Even though the disease stages vary to some degree across TSS, the sites display similar distributions of clinical metadata.

Table 1: Dataset statistics, i.e., the average number of patches per slide after preprocessing. For CAMELYON16 only patches annotated as 100% tumorous or normal are considered.

Dataset	TCGA-LUSC-5					CAMELYON16			
Subset	TSS 22	TSS 56	TSS 66	TSS 77	TSS 85	TSS 0 Normal	TSS 0 Tumor	TSS 1 Normal	TSS 1 Tumor
# Slides	37	41	38	44	50	150	95	89	65
# Patches per slide	7,351	2,577	2,520	11,320	6,983	6,663	9,090	9,733	10,715
# Tumor patches per slide	-	-	-	-	-	-	508	-	993

Foundation models

We used multiple publicly available histopathological foundation models (see Table 2 for an overview). They cover a broad spectrum of architectures (convolutional vs. ViTs), pretraining frameworks (SimCLR, DINO, SRCL, iBOT, DINOv2), dataset diversity (6k–1,5M WSIs), and model sizes (11M–1100M parameters). We passed all patches from our datasets through all foundation models to obtain real-valued feature vectors from the [CLS] token representations. The patches were resized to the size expected by the respective model.

Table 2: Overview of analyzed histopathological foundation models.

Model		Pretraining objective	#Pretraining WSIs	Architecture	#Parameters	Feature dimension
Ciga	[16]	SimCLR	(400k patches)	ResNet-18	11M	512
Hipt	[17]	DINO	10k	ViT-S/16	21M	384
RetCCL	[18]	cluster-guided contrastive	34k	ResNet-50	25M	2048
Kang-DINO	[20]	DINO	36k	ViT-S/8	21M	384
CTransPath	[19]	SRCL	34k	SWIN-T	28M	768
Phikon	[21]	iBOT	6k	ViT-B/16	86M	768
UNI	[24]	DINOv2	100k	ViT-L/16	303M	1024
ProvGigapath	[23]	DINOv2	171k	ViT-g/14	1100M	1536
Virchow	[25]	DINOv2	1.5M	ViT-H/14	631M	1280

4 The role of batch effects in histopathological foundation models

First, we analyzed whether site signatures contained in patches are preserved in the features extracted by histopathological foundation models and how they affect downstream prediction tasks. We further assessed the effectiveness of stain normalization to mitigate such batch effects.

4.1 Experiment 1: Source site prediction

We evaluated to what extent the tissue source site of a patch can be predicted from its feature vector. We sampled roughly 50,000 patches per TSS, split them into 60% training, 10% validation, and 30% test data at patient level, and trained classifiers to predict the TSS of a patch. As classifiers, we chose the nearest centroid classifier (NCC), k-nearest neighbors (KNN), and linear probing (LP). We ran this experiment for all foundation models and repeated it with the features of the stain-normalized patches. Further details are described in Appendix A.2.1.

The results are displayed in Figure 1. For all foundation models, the TSS of the patches were recovered from their feature vectors with very high accuracy. An almost perfect classification was achieved via LP on CAMELYON16, and over 90% accuracy for most foundation models on TCGA-LUSC-5. This implies that the feature representations of almost all patches carry distinct signatures of their TSS that can be identified via linear probing, regardless of the foundation model used. Even with NCC and KNN, which base their decisions on proximity of features, performances of over 90% (CAMELYON16) and 75% (TCGA-LUSC) were reached for most foundation models. This already indicates that these signatures rather than biological signals regularly dominate the distances in feature space. This is further explored in Section 5.

More recent models — trained on more data and showing better downstream performances [21, 23, 24, 28] — generally appeared to contain stronger TSS signatures. This points towards a trade-off between foundation model performance and susceptibility to batch effects. At the same time, we observed differences between recent models: NCC and KNN extracted less site information from Virchow than from UNI, ProvGigapath, or Phikon. This indicates that site signatures may be more dominant and available in the latter models, though present in all.

The stain normalization methods reduced the TSS prediction accuracy across all models, classifiers, and datasets. However, the LP accuracies were still over 80% in TCGA-LUSC-5 and over 97% in CAMELYON16 for most foundation models. Moreover, stain normalization seemed to have a stronger impact on NCC and KNN. We infer that Macenko and Reinhard target more dominant parts of the TSS signature, while not affecting more subtle features that can still be extracted via LP.

Therefore, despite large-scale and multi-site pretraining, histopathological foundation models do not fully eliminate batch effects, but still encode accessible TSS signatures in their feature vectors. Image-based stain normalization techniques slightly mitigate but do not remove these signatures.

4.2 Experiment 2: Effect of site-specific signatures on downstream tasks

We examined the impact of TSS signatures on a downstream task by creating a biased dataset with varying correlations between the classification label and the issuing TSS. The experiment is conducted on CAMELYON16 and stipulates the classification of patches as 'tumor' or 'normal'. For each TSS,

Table 3: Four compositions of training data for the cancer prediction task. Each row presents the numbers of normal and cancer patches with their corresponding source site. The split identifiers (*Split i – x/y*) provide information about the varying ratios x and y of cancerous patches per TSS 0 and TSS 1, respectively. The allocation of test data remains fixed as 0/1.

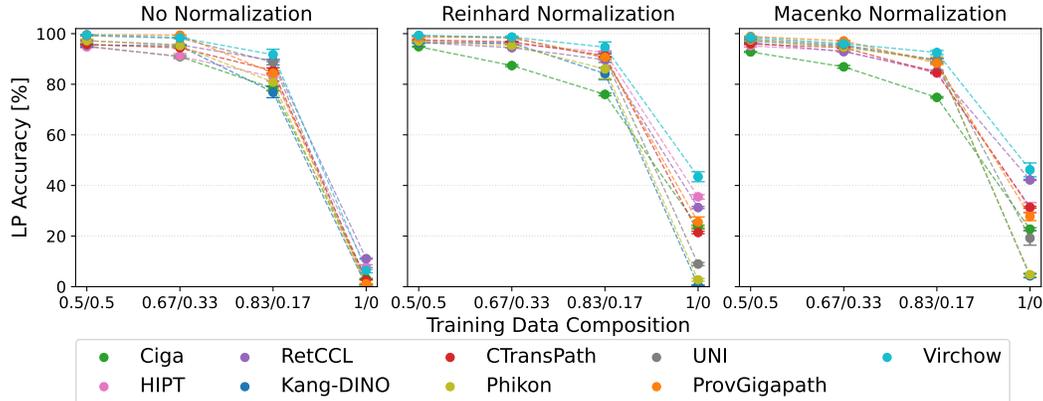|		TSS 0		TSS 1		$\sum$
		# normal	# cancer	# normal	# cancer	
Split 1 – 0.5/0.5	Training	7,500	7,500	7,500	7,500	30,000
Split 2 – 0.67/0.33	Training	5,000	10,000	10,000	5,000	30,000
Split 3 – 0.83/0.17	Training	2,500	12,500	12,500	2,500	30,000
Split 4 – 1/0	Training		15,000	15,000		30,000
	Test	5,000			5,000	10,000

5

Figure 3: Cancer classification accuracy in % using LP on CAMELYON16 for different training data compositions (x/y). The test data composition remains 0/1 across splits. We assessed the performances for features of unnormalized, Reinhard, and Macenko normalized patches, and report mean and standard deviation over 5 repetitions.

we sampled 14 slides and extracted $2,500$ patches each. We created four slide-level training splits as summarized in Table 3. While the first split simulates a perfectly balanced data set with equally distributed tumor and normal patches, the training data of the fourth split consists of cancerous patches solely from TSS 0, whereas TSS 1 only provides healthy tissue patches. Notably, the test data, which does not change per split, is built vice versa: normal patches come from TSS 0 and cancerous ones from TSS 1. We trained LP classifiers per split and foundation model to assess whether the classifiers are affected by TSS signatures. We also repeated the experiments with the features of the stain-normalized patches (further details in Appendix A.2.2).

The results, depicted in Figure 3, demonstrate that the correlation between TSS signatures and downstream task labels can lead to biased predictions and misclassifications. As the correlation increases, tumor classification accuracy drops from nearly perfect ($> 90\%$ on uncorrelated data) to below random ($< 20 - 50\%$ on fully correlated data). Significant performance drops can already be observed for the 0.83/0.17 split. This indicates that the downstream model appoints TSS information as a proxy for cancer prediction when they are in concordance.

In comparison, the accuracy of more recent models like Virchow and UNI was slightly higher and more stable than of earlier models like Ciga, particularly on the 0.67/0.33 and 0.83/0.17 splits, despite stronger encoding of TSS signatures as shown in Section 4.1. This indicates that more recent models profit from other factors aiding their generalizability, for which further investigation will be required.

Stain normalization increased the tumor prediction performance for imbalanced splits, especially on the more extreme 0.83/0.17 and 1/0 settings. However, the respective accuracies are still far below those of the balanced 0.5/0.5 split, and even below random performance for the 1/0 scenario. This shows that stain normalization has a limited impact on batch effect mitigation in feature space.

A potential explanation for our results is shortcut learning [34, 32, 33, 35], which occurs when models choose features not just for their predictive power but also for their availability. Hermann et al. [34] observed that models prefer more available features, even when they are less predictive. For instance, in natural images, models focus on object scale, texture, or background, as these are easily accessible [42]. Translated to the histopathological domain, this might indicate that foundation models are easily misled by predominant batch effects like staining or thickness rather than biological cell alterations because of their availability, and base their predictions on those instead. Section 5 demonstrates that TSS signatures are more available than clinical information and dominate the variance in the features, which further exacerbates shortcut learning.

In practice, many histopathological datasets suffer from batch effects, especially cohorts coming from multiple hospitals (see e.g. [27]). Therefore, our results imply that models built on top of histopathological foundation models are not necessarily able to generalize across sites and their reported performances should be interpreted with care, as they may give correct predictions for the wrong reasons [34, 32, 33]. For training more generalizable and unbiased models, data balancing,
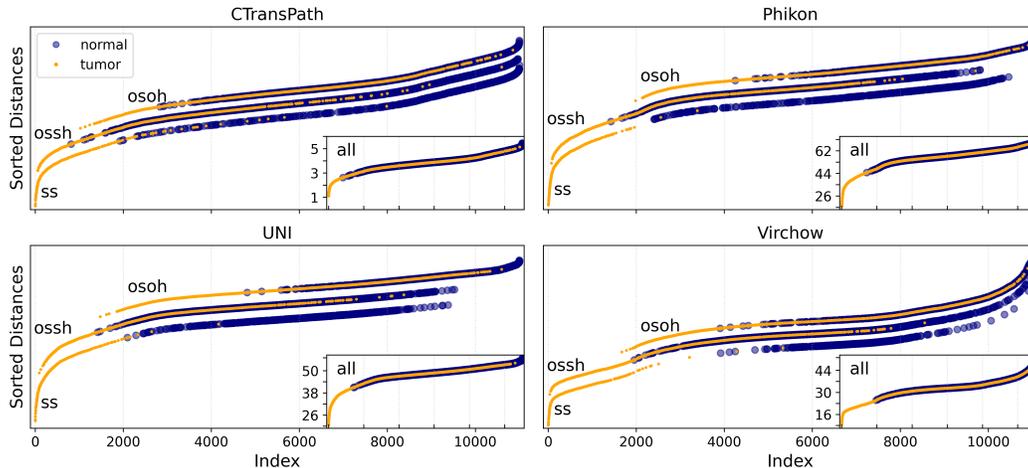
Figure 4: Ordered Euclidean distances between the feature of a randomly drawn cancerous reference patch and the features of *ss* (bottom line), *ossh* (middle line), and *osoh* (top line) on CAMELYON16. Because the distances lay all on the same line (see *all*), small offsets were added for clearer visualization such that *ss*, *ossh*, and *osoh* become distinguishable.

as proposed in [27], could be applied to ensure that batch effects do not confound the prediction task. Furthermore, new pretraining strategies accounting for batch effects could lead to more robust downstream predictors.

5 Characterization of batch effects in histopathological foundation models

After establishing the presence and impact of batch effects in histopathological foundation models, we conducted multiple experiments to characterize them in greater depth. For that, we focus on four foundation models with different model sizes, dataset sizes, and pretraining frameworks: CTransPath, Phikon, UNI, and Virchow.

5.1 Experiment 3: Visualization of distances in feature space

First, we analyzed how batch effects impact the distances between feature vectors of patches. We sampled a tumor reference patch from a randomly chosen cancerous slide from CAMELYON16. Subsequently, we drew $1,000$ patches from the same slide (*ss*), along with $5 \times 1,000$ patches from five randomly chosen other cancerous slides from the same hospital (*ossh*), and $5 \times 1,000$ patches from five cancerous slides from the other hospital (*osoh*). Again, we only considered patches annotated as 100% normal or 100% tumor, and sampled equally from both classes for each slide.

In Figure 4, we depict the ordered Euclidean distances in the feature space to the reference patch. Most notably, there is a clear distinction between the distances to the closest patches from the same TSS compared to the ones from other TSSs. While patches from the same slide are nicely separated according to their cancer annotation, distances to normal or cancer patches largely intersect for patches from other slides. Even worse, the transition between cancer to normal patches occurs at disparate distances for *ss*, *ossh*, and *osoh*, rendering cancerous patches of other sites as equidistant to normal patches of the same TSS. This implies that batch effects outweigh the biological signal of cancerous or normal morphology.

In general, the experiment shows that analyses and classification methods solely based on distances between the feature vectors from histopathological foundation models are likely to fail on multi-site data due to the strong encoding of batch effects. For example, a distance-based clustering approach is unlikely to correctly identify cancer patches of different sites as belonging to the same cluster.
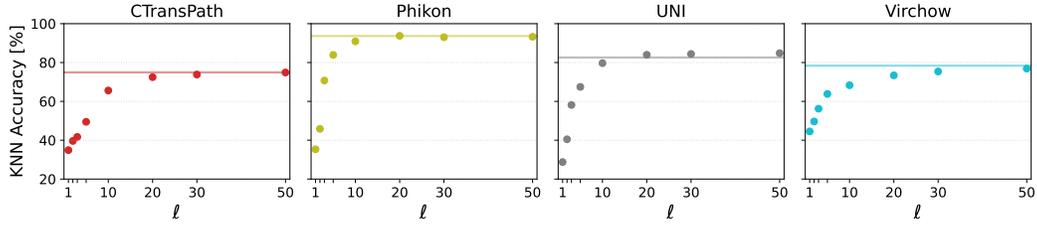
Figure 5: Accuracy on site prediction task as described in Section 4.1 on TCGA-LUSC-5 using KNN based on features projected onto the first ℓ PCs (dots) and on non-reduced features (line).
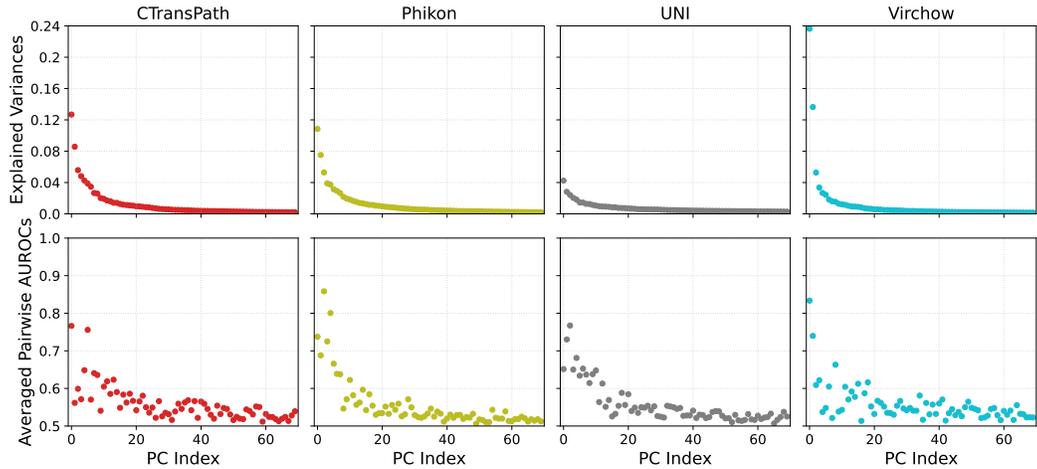

Figure 6: Estimated explained variance per principal component (top) and the TSS separability of the one-dimensional feature projections on the same PCs calculated by a one-vs-one AUROC across tissue source sites (bottom) on TCGA-LUSC-5.

5.2 Experiment 4: Site prediction performance on reduced feature data

As a next step, we tried to understand how TSS signatures are encoded in the feature space of the foundation models. We specifically aimed to answer how prevalent, available, and concentrated the signatures are, and whether it is possible to isolate them. We approached these questions through the lens of dimensionality reduction and applied PCA to the feature vectors for the whole TCGA-LUSC-5 dataset per foundation model.

As a first step, we investigated to what extent site-specific information is embedded in the first principal components of the features. For that, we revisited the source site prediction task from Section 4.1 using the same dataset and split. However, we trained a KNN classifier on the reduced features projected onto the first $\ell = \{1, 2, 3, 5, 10, 20, 30, 50\}$ principal components (PCs).

The results are displayed in Figure 5. For all models, high classification accuracy was already achieved with only the first few components of the data. For UNI and Phikon, we observed a steep initial increase in performance when adding principal components, almost attaining the performance on non-reduced features with only 10 PCs. For CTransPath and Virchow, this increase is less steep, still with only 20-30 PCs full feature vector performance is obtained.

This shows the high availability of TSS signatures in the feature vectors, as they are already encoded in the largest principal components. We argue that supervised downstream models are prone to utilizing these signatures as shortcuts instead of intended biological features for solving their tasks.

8

5.3 Experiment 5: Component-wise separability

While the previous experiment established that a low-dimensional representation of the features already contains sufficient information to predict their submitting site, we now examine the contribution of each principal component individually to the site-wise class separability. The features were projected onto each PC separately and we analyzed how well the projected one-dimensional features can be linearly separated. We measured their separability via a multi-class one-vs-one area under the receiver operating characteristic curve (AUROC) using the one-dimensional feature projections as scores and the TSS as labels[2]. This measure naturally handles class imbalances, guarantees comparability across models as it is always in $[0.5, 1]$, and is nonparametric since it does not require any further classification algorithm.

Figure 6 shows the linear class separability together with the explained variance ratio of each PC (normalized eigenvalues). We observe that many of the directions in which the data varies the most simultaneously exhibit a high site separability, making them dominant and available factors in the feature vectors. At the same time, we see that TSS signatures exist in various principal components, which makes it hard to isolate them. The pattern of how the batch effects spread across different components varies for different models.

We infer that removing site signatures is arguably difficult, as they are prevalent in various components and encoded deep within the feature space.

6 Conclusion

In this study, we showed that histopathological foundation models do not fully eliminate batch effects. Instead, their feature vectors contain strong site-specific information, which can negatively affect the classifier performance on downstream tasks with imbalanced datasets. Image-based stain normalization methods slightly mitigate but do not remove these signatures, limiting their effect on downstream performance. Furthermore, the site signatures dominate patch distances in feature space and are prevalent in various principal components, which makes distance and supervised learning-based methods susceptible to batch effects.

Our study comes with some limitations. The datasets we used are small, and their batch effects are known to be stark, meaning that we did not explicitly assess more subtle effects. Additionally, even though we tried sampling the data points to have similar clinical metadata across subsets, the site signatures may be partly caused by unknown biological differences for which different embeddings are desirable. We also did not identify which factors cause the batch effects to what extent, evaluate how batch effects affect generalization performance on previously unseen tissue source sites, or investigate what other factors aid or hinder the models' generalization capabilities. We leave these open questions for future work.

Regardless, we think that our results have interesting implications. First, we advocate for the assessment of batch effects to become an additional cornerstone of evaluating histopathological foundation models. Second, we think that it should be taken into account for the development of future foundation models. Third, our study opens up a promising new direction of research: developing methods for eliminating batch effects directly from feature vectors instead of tackling them on image level or during foundation model training.

---

[2]Instead of prediction probabilities usually required for the computation of the AUROC, we use the one-dimensional projected features as decision values. Since the orientation of these values is unknown, we compute the maximum of the one-vs-one AUROCs of both possible true label assignments. Recall that decision values above a threshold are predicted as positive so that it matters whether the reference label is set to be 0 or 1 for the computation of the ROC albeit not affecting the actual site prediction task.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

[2] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018.

[3] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[4] Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21(2):222–232, 2020.

[5] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[6] Ming Y Lu, Tiffany Y Chen, Drew F K Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

[7] Jonas Dippel, Niklas Prenißl, Julius Hense, Philipp Liznerski, Tobias Winterhoff, Simon Schallenberg, Marius Kloft, Oliver Buchstab, David Horst, Maximilian Alber, Lukas Ruff, Klaus-Robert Müller, and Frederick Klauschen. Ai-based anomaly detection for clinical-grade histopathological diagnostics. *NEJM AI*, 0(0):AIoa2400468, 2024.

[8] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, 2020.

[9] Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Masaru Ishii, Albrecht Stenzinger, Andreas Hocke, Carsten Denkert, Klaus-Robert Müller, and Frederick Klauschen. Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence*, 3(4):355–366, 2021.

[10] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4):686–696, 2021.

[11] Salim Arslan, Julian Schmidt, Cher Bass, Debapriya Mehrotra, Andre Geraldes, Shikha Singhal, Julius Hense, Xiusi Li, Pandu Raharja-Liu, Oscar Maiques, et al. A systematic pan-cancer study on deep learning-based prediction of multi-omic biomarkers from routine pathology images. *Communications Medicine*, 4(1):48, 2024.

[12] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, Olivier Elemento, Andrew G Nicholson, Jean-Yves Blay, Françoise Galateau-Sallé, Gilles Wainrib, and Thomas Clozel. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, 2019.

[13] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020.

[14] Charlie Saillard, Benoit Schmauch, Oumeima Laifa, Matahi Moarii, Sylvain Toldo, Mikhail Zaslavskiy, Elodie Pronier, Alexis Laurent, Giuliana Amaddeo, Hélène Regnault, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology*, 72(6):2000–2013, 2020.

[15] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.

[16] Ozan Ciga, Tony Xu, and Anne Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7, 2022.

[17] Richard Chen, Chengkuan Chen, Yicong Li, Tiffany Chen, Andrew Trister, Rahul Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, 2023.

[19] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[20] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[21] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv preprint medRxiv:2023.07.21.23292757v2*, 2023.

[22] Nanne Aben, Edwin D de Jong, Ioannis Gatopoulos, Nicolas Känzig, Mikhail Karasikov, Axel Lagré, Roman Moser, Joost van Doorn, Fei Tang, et al. Towards large-scale training of pathology foundation models. *arXiv preprint arXiv:2404.15217*, 2024.

[23] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015), 2024.

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

[25] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(7):1–12, 2024.

[26] Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Timo Milbich, Stephan Tietz, Simon Schallenberg, Gabriel Dernbach, Andreas Kunft, Simon Heinke, Marie-Lisa Eich, Julika Ribbat-Idel, Rosemarie Krupar, Philipp Anders, Niklas Prenißl, Philipp Jurmeister, David Horst, Lukas Ruff, Klaus-Robert Müller, Frederick Klauschen, and Maximilian Alber. Rudolfv: A foundation model by pathologists for pathologists, 2024.

[27] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12(1):4423, 2021.

[28] Gabriele Campanella, Shengjia Chen, Ruchika Verma, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Kuan-lin Huang, Ricky Kwan, et al. A clinical benchmark of public self-supervised pathology foundation models. *arXiv preprint arXiv:2407.06508*, 2024.

[29] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[30] Marc Macenko, Marc Niethammer, J. Marron, David Borland, John Woosley, Xiaojun Guan, Charles Schmitt, and Nancy Thomas. A method for normalizing histology slides for quantitative analysis. In *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, 2009.

[31] Frederick Klauschen, Jonas Dippel, Philipp Keyl, Philipp Jurmeister, Michael Bockmayr, Andreas Mock, Oliver Buchstab, Maximilian Alber, Lukas Ruff, Grégoire Montavon, and Klaus-Robert Müller. Toward explainable artificial intelligence for precision pathology. *Annual Review of Pathology: Mechanisms of Disease*, 19(1):541–570, 2024.

[32] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[33] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[34] Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[35] Jacob Kauffmann, Jonas Dippel, Lukas Ruff, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. The clever hans effect in unsupervised learning. *arXiv preprint arXiv:2408.08041*, 2024.

[36] Fernando Navarro, Christopher Watanabe, Suprosanna Shit, Anjany Sekuboyina, Jan C Peeken, Stephanie E Combs, and Bjoern H Menze. Self-supervised pretext tasks in model robustness & generalizability: A revisit from medical imaging perspective. In *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.

[37] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[38] Anurag Vaidya, Richard J Chen, Drew FK Williamson, Andrew H Song, Guillaume Jaume, Yuzhe Yang, Thomas Hartvigsen, Emma C Dyer, Ming Y Lu, Jana Lipkova, et al. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30(4):1174–1190, 2024.

[39] Cesare Gridelli, Antonio Rossi, David Carbone, Juliana Guarize, Niki Karachaliou, Tony Mok, Francesco Petrella, Lorenzo Spaggiari, and Rafael Rosell. Non-small-cell lung cancer. *Nature Reviews Disease Primers*, 1:15009, 2015.

[40] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics (SMC)*, 9(1):62–66, 1979.

[41] Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, and Tapio Seppänen. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, 102:101997, 2024.

[42] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations (ICLR)*, 2019.

A Appendix

A.1 Dataset details

Example patches and metadata statistics of TCGA-LUSC-5 are provided in Figure 7 and Table 4, respectively.
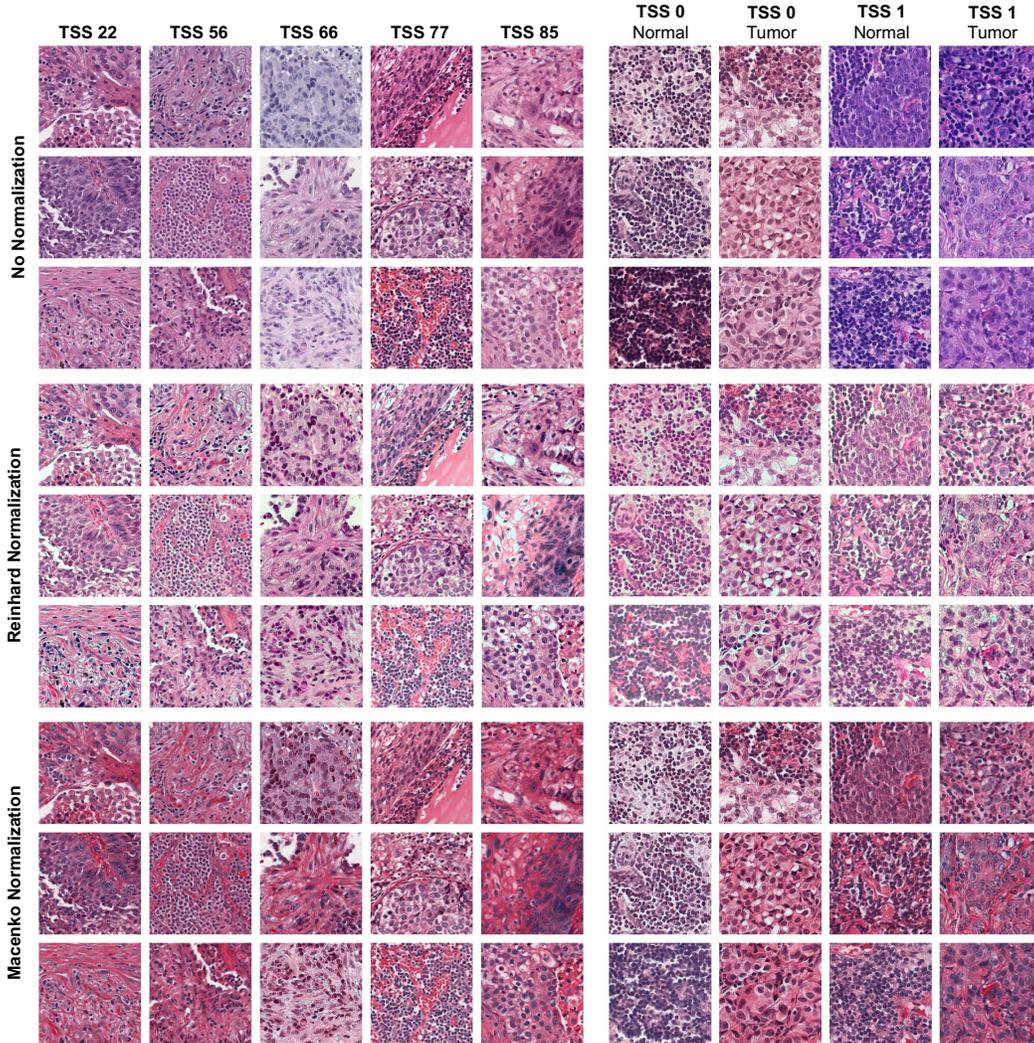

Figure 7: Arbitrarily chosen patches from the TCGA-LUSC-5 dataset (left) and the CAMELYON16 dataset (right) with Reinhard [29] and Macenko [30] stain normalization.

A.2 Experimental details

A.2.1 Experiment 1: Source site prediction

For both datasets, we randomly selected $n_i = \frac{50,000}{m_i}$ patches per slide of TSS i with m_i being the number of slides of site i. If a slide contained less than n_i patches, we drew all the available patches. This resulted in approximately 250,000 patches for TCGA-LUSC-5 and 100,000 for CAMELYON16 with roughly balanced site labels. We split each dataset into 60% training, 10% validation, and 30% test data at patient level, ensuring that no patches from the same slide appear in both training and validation or test set.

Table 4: Selected metadata statistics of the TCGA-LUSC-5 dataset. We chose *primary_diagnosis*, *site_of_resection_or_biopsy*, *ajcc_pathologic_t*, and *ajcc_pathologic_stage* as possibly relevant columns in the clinical metadata of the TCGA LUSC project. We report the number of slides per stage and subset. SCC = squamous cell carcinoma; NOS = not otherwise specified.

Primary diagnosis	TSS 22	TSS 56	TSS 66	TSS 77	TSS 85
Papillary SCC	0	0	0	1	2
SCC, NOS	36	39	38	40	45
SCC, keratinizing, NOS	0	1	0	3	2
SCC, large cell, nonkeratinizing, NOS	1	1	0	0	1

Site of resection or biopsy	TSS 22	TSS 56	TSS 66	TSS 77	TSS 85
Lower lobe, lung	13	16	16	12	13
Lung, NOS	9	0	0	8	0
Main bronchus	0	0	0	1	0
Middle lobe, lung	0	0	3	1	2
Overlapping lesion of lung	1	1	0	1	0
Upper lobe, lung	14	24	19	21	35

AJCC T stage	TSS 22	TSS 56	TSS 66	TSS 77	TSS 85
T1	6	0	4	5	2
T1a	3	3	0	0	6
T1b	4	6	0	0	10
T2	9	3	28	23	1
T2a	8	14	0	6	13
T2b	2	7	0	0	8
T3	4	8	0	7	9
T4	1	0	6	3	1

AJCC stage	TSS 22	TSS 56	TSS 66	TSS 77	TSS 85
'_	1	0	0	0	0
Stage IA	10	8	3	3	13
Stage IB	15	15	14	14	11
Stage IIA	5	8	0	4	12
Stage IIB	1	6	6	13	9
Stage IIIA	5	4	3	7	5
Stage IIIB	0	0	11	3	0
Stage IV	0	0	1	0	0

We used three classifiers of increasing expressivity to predict the source site of the patches' feature vectors: nearest centroid classifier (NCC), k-nearest neighbors (KNN), and linear probing (LP). For KNN, we used the $k = 5$ nearest neighbors from the train set to predict the source site of a test patch. For LP, we used the Adam optimizer, an initial learning rate of 0.001, and the cross-entropy loss with a batch size of 128 to train the single-layer classification head. We trained for 20 epochs and selected the model with the lowest validation loss, which we then applied to the held-out test set. We computed the accuracy score by comparing the class with the highest predicted logit with the sample label.

A.2.2 Effect of site-specific signatures on downstream tasks

For each TSS, we sampled 7 normal and 7 tumor slides (total: 28 slides), ensuring that each contained at least 2,500 patches that are 100% normal or 100% tumor according to the pathologist annotation, respectively. For each normal slide, we then sampled 2,500 patches that are 100% normal, and for each tumor slide 2,500 patches that are 100% tumorous. From this pool of patches, we created splits as described in Table 3. All splits were created on slide level, and all 2,500 patches from a selected slide were used: for instance, if Split 2 contains 10,000 cancer patches from TSS 0, it means that they come from 4 different tumor slides contributing 2,500 distinct tumor patches, each. The patches used for the test set were the same across different training splits. We also added 5,000 patches from different slides as validation sets, having the same TSS-label compositions as the training splits, respectively. We highlight that in any case, patches from the same slide only appeared in either train, validation, or test set, but never more than one of the subsets.

We applied linear probing (LP) to predict whether a patch is cancerous or normal. We used the Adam optimizer, an initial learning rate of 0.001, and the cross-entropy loss with a batch size of 128 to train the single-layer classification head. We trained for 20 epochs and selected the model with the lowest validation loss, which we then applied to the held-out test set. We computed the accuracy score by comparing the class with the highest predicted logit with the sample label. We repeated the training process 5 times on the same splits, and report the mean and standard deviation of the accuracy scores.

A.2.3 Compute resources

Data processing and feature extraction with the foundation models were conducted on our internal cluster equipped with GPUs having at least 24GB GPU memory, each. The datasets required roughly 1TB of disk memory. As the extracted feature vectors only made up a small fraction of the total memory used, all experiments could be executed on a standard CPU.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The contributions, claims, and the scope of the paper are made clear in the Abstract and Introduction (Section 1).

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of our work are extensively discussed in the Conclusion (Section 6).

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [NA]

Justification: [NA]

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: All datasets, models, and experiments are clearly described in the respective Sections 3, 4, and 5.

Guidelines:

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

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

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

Answer: [No]

Justification: Our code is relatively straightforward, as we conduct simple experiments, and it is not needed to reproduce the experiments. Datasets and models used in this work are publicly available and properly described in Sections 3. We will happily provide all codes and datasets upon request and release them upon publication alongside documentation.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: All the details in this regard are provided in Sections A.2.1, A.2.1, 4.2, A.2.2, 5.1, 5.2, and 5.3.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [No]

Justification: Our experiments are not expected to be subject to great statistical variations due to the sufficient dataset sizes and simplicity of the experiments, with the exception of the experiment described in Section 4.2, where we report mean and standard deviation over 5 repetitions.

Guidelines:

- The answer NA means that the paper does not include experiments.

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: The compute resources are mentioned in Section A.2.3.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: (1) Our data are from public datasets that have followed the guidelines of their corresponding ethics committee. (2) We declare no known potential harmful consequence of our submitted work regarding the mentioned points in the code of ethics. (3) If applicable, we implemented all the impact mitigation measures mentioned in the code of ethics, such as disclosing essential elements for reproducibility.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

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

Answer: [NA]

Justification: We anticipate no negative societal impact generated by our work.

Guidelines:

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

11. **Safeguards**

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

Answer: [NA]

Justification: We are aware of no potential risk for any misuse of our work.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: (1) We cited the open datasets and analyzed foundation models properly in Section 3. (2) If some parts of other repositories have been used in our codes, we have cited the repository properly and indicated the corresponding license.

Guidelines:

- The answer NA means that the paper does not use existing assets.

20

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

13. **New Assets**

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

    Answer: [NA]

    Justification: We do not release any new assets, except for our codes, which will be made publicly accessible upon publication alongside documentation.

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing. The human data are acquired from public datasets, properly cited in the manuscript (please see points 5, 9, 12 of this checklist.).

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

Justification: Please see point 14.

Guidelines:

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