Current Pathology Foundation Models are unrobust to Medical Center Differences

Edwin D. de Jong[1], Eric Marcus[2], and Jonas Teuwen[2]

[1]Independent research performed in November 2024, after affiliation with Kaiko and prior to affiliation with Aignostics; updated with additional results in January 2025

[2]The Netherlands Cancer Institute Amsterdam (NKI), Antoni van Leeuwenhoek Hospital (AvL)

Abstract

Pathology Foundation Models (FMs) hold great promise for healthcare. Before they can be used in clinical practice, it is essential to ensure they are robust to variations between medical centers. We measure whether pathology FMs focus on biological features like tissue and cancer type, or on the well known confounding medical center signatures introduced by staining procedure and other differences.

We introduce the **Robustness Index**. This novel robustness metric reflects to what degree biological features dominate confounding features. Ten current publicly available pathology FMs are evaluated. We find that all current pathology foundation models evaluated represent the medical center to a strong degree. Significant differences in the robustness index are observed. Only one model so far has a robustness index greater than one, meaning biological features dominate confounding features, but only slightly.

A quantitative approach to measure the influence of medical center differences on FM-based prediction performance is described. We analyze the impact of unrobustness on classification performance of downstream models, and find that cancer-type classification errors are not random, but specifically attributable to *same-center confounders*: images of other classes from the same medical center. We visualize FM embedding spaces, and find these are more strongly organized by medical centers than by biological factors. As a consequence, the medical center of origin is predicted more accurately than the tissue source and cancer type.

The robustness index introduced here is provided with the aim of advancing progress towards clinical adoption of robust and reliable pathology FMs.

1 Introduction

Pathology Foundation Models (FMs) have quickly become the dominant approach in current pathology AI. Following Campanella's groundbreaking work

that used weakly-supervised learning to scale up machine learning for pathology [1] and early papers applying Self-Supervised Learning (SSL) in the domain [2–4], an impressive and quickly growing series of pathology FMs have become available, with more than ten new such models published last year so far alone: [5–18].

Several of these models demonstrate remarkable capabilities and can detect patterns that human pathologists struggle to observe from H&E slides, such as Microsatellite Instability (MSI) and Immunohistochemistry (IHC) biomarkers such as Ki67 and PD-L1 [8, 19, 20].

Pathology foundation models thus hold great potential for healthcare by aiding pathologists, for example in routine or large volume, labor intensive tasks. If this promise is to be realized, it is essential that models can be trusted to provide unbiased estimates of a patient's condition.

A particular obstacle to the adoption of pathology FMs in clinical practice is the sensitivity of machine learning models (ML) to staining variations, caused by differences in the staining procedures used by different labs, the staining fluids, and imaging equipment. It is well known that these image variations can influence pathology ML models, and reduce their ability to generalize to data from laboratories not seen during training [21].

Staining procedures vary per laboratory, and are thereby associated with specific medical centers. This leads to a clear risk of bias [22]: if models are sensitive to the medical center from where images originate, then patients from different medical centers will be evaluated differently by such models, possibly leading to different diagnoses and treatment based on irrelevant technical differences between images. To ensure foundation models can be safely introduced into healthcare practice, evaluating and confirming their robustness to image variations that occur in practice is a necessary step.

To assess whether current pathology foundation models provide an objective assessment of a patient's condition, we analyze to what extent medical centers influence the embedding spaces generated by FMs. Our contributions are as follows:

- A basic yet effective description for the concept of *robustness* in medical ML is suggested, based on the distinction between biological features and confounding features. Models can vary in their robustness to image variations such as noise, color differences, augmentations, and variations between medical centers.
- We introduce a novel robustness metric: the Robustness Index, measuring the degree to which biological features dominate confounding features in the neighborhood structure of the embedding space induced by the foundation model.
- For the first time, a quantitative approach to measure the influence of medical center differences on FM-based prediction performance is described. The approach directly relates prediction errors to same-center

confounders: images from the same center as the predicted sample that have a different class, and thereby contribute to incorrect classification.

- 10 current pathology foundation models are evaluated on their medical center robustness. It is found that current pathology FMs vary widely in their robustness to medical center variations.
- We suggest that the value of an FM is determined by the relation between (A) its ability to characterize relevant biological information, enabling high prediction performance of biological information in downstream tasks, and (B) its robustness, reflected in its insensitivity to irrelevant non-biological variation such as staining and medical center differences.
- To gain further insight into the mappings learned by pathology FMs, we project embeddings to a 2D space for visualization using t-SNE [23]. We find that most foundation models show a clear clustering of medical centers in the embedding space; this shows more clearly than the clustering of biological classes.

2 Related Work

In previous work [24] presented at AMLD 2024, we visualized the embedding space of ViT models trained on TCGA using DINO. It was found that a 2D t-SNE projection of the embedding space was clustered by medical center, forming the inspiration for the current work.

In simultaneous work from the TU Berlin BIFOLD group [25], batch effects in pathology foundation models were analyzed and shown. [26] analyzes rotation invariance. Tellez [21] quantified the effects of data augmentation and stain color normalization in pathology in the context of CNNs, and [27] evaluated the effect of stain normalization in colorectal tissue classification.

[28] discusses the issue that hospitals are represented in WSI data, and looks into domain generalization for hospital-agnostic image representation learning; [29] also uses domain adaptation to reduce the influence of staining differences. [30] draws attention to the existence of medical center signatures in TCGA data specifically, and finds that deep neural networks can predict the acquisition site.

A main consequence of unrobustness is that predictive models based on unrobust representations can lead to biased predictions. [31] studies the impact of medical center signatures on deep learning model accuracy and bias. [18] notes that embeddings from SSL models tend to cluster by individual WSIs and proposes the EXAONEPath pathology FM to address this, which is included in our evaluation here.

While finalizing this paper, a new relevant article discussing the measurement and optimization of robustness in pathology became available [32].

Paper we've read, they use Macenko!

3 Robustness for Medical Foundation Models

To clarify what is meant by robustness in this work, we distinguish between *biological* features and *confounding* features. Biological features include any relevant features that reflect the true condition of the patient; the aim and promise of foundation models is to capture these. Confounding features are any irrelevant variations in the input that are not related to true biological differences between samples, but are rather caused by external influences such as staining differences, differences in image capture equipment, image processing pipelines, and noise. Given these notions, we can define robustness as insensitivity to confounding features.

3.1 Robustness Index

To gain insight into what a foundation model has learned, we can analyze the embedding space by considering the neighborhood around each embedding, i.e. the closest embeddings. We consider:

- How many of the k nearest neighbors represent the same biological class, e.g. tissue type or cancer type, in total across all samples
- How many of the k nearest neighbors represent the same medical center, in total across all samples

We define the **medical center robustness index** as the ratio between these quantities. For other biological classes (e.g. other diseases, or pharmacogenomic groups) or confounding factors (e.g. scanner type), robustness indices can be defined analogously.

Formally: we define the robustness index R_k for a given dataset D containing n samples as:

$$R_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \mathbf{1}(y_j = y_i)}{\sum_{i=1}^{n} \sum_{j=1}^{k} \mathbf{1}(c_j = c_i)} \tag{1}$$

Where:

- k is the number of nearest neighbors considered; in this work, $k = 50$
- y_j is the biological class of the j-th nearest neighbor; y_i is the biological class of the sample i
- c_j is the medical center of the j-th nearest neighbor; c_i is the medical center of the sample i
- $\mathbf{1}(\cdot)$ is the indicator function: 1 if the condition is true, 0 otherwise

The numerator represents the total number of nearest neighbors that have the same biological class across all samples, while the denominator represents the total number of nearest neighbors that are from the same medical center across all samples. Cosine distance is used as the distance metric, as cosine similarity is a common way to evaluate embedding similarity.

As an example, consider an embedding spaced dominated by a confounder, say center, and only minutely influenced by the cancer type. Then, the denominator will be large, since most samples will be surrounded by other samples of the same center; the numerator is small since the embedding space is mostly organized by the confounder.

A robustness index of 1 means that biological and confounding information are represented equally strongly. In an idealized scenario, the embedding space is blind to the confounding information and completely organized by the biological signal, yielding $R_k = R_{max}$, where R_{max} is determined by the random chance level for encountering a neighbor from the same medical center or confounding class. In practice, it may not be feasible to completely remove center information, but one would like to see $R_k >> 1$. For the dataset used in this paper, the Robustness Index is expected to vary between 1/4 for random cancer type prediction in combination with perfect medical center prediction and 4 for the converse.

4 Experimental Setup

Apart from robustness, another important measure of the quality of a foundation model is reflected in the prediction performance of the downstream models built upon it. In Section 4.1, we therefore define a basic classification task.

4.1 Classification Task: Tissue of Origin / Cancer Type

A classification task for the cancer types of five TCGA projects is defined: BReast invasive CArcinoma (BRCA), COlon ADenocarcinoma (COAD), LIver Hepatocellular Carcinoma (LIHC), LUng Squamous cell Carcinoma (LUSC) and STomach ADenocarcinoma (STAD). Note that these cancer types have a one-to-one correspondence with five different tissues of origin (breast, colon, liver, lung, stomach); so this task can equivalently be viewed as a tissue of origin classification task.

These particular five cancer types were selected in combination with five medical centers such that for each cancer type, TCGA WSI data from multiple medical centers is available and vice versa, resulting in the following selection of centers: Asterand, Greater Poland Cancer Center (GPCC), ILSBio, International Genomics Consortium (IGC), MSKCC; see table 1.

To build the dataset, for each available combination of center and cancer type, 10 WSIs are selected randomly. From each of the resulting WSIs, 10 informative foreground patches representing regular tissue were selected from a

larger randomly generated set based on visual inspection, filtering out anomalies and low-information (white) patches. This resulted in a dataset of 2000 patches in total. We name this dataset `TCGA-2k`. In all experiments, 5-fold cross-validation is used to generate validation predictions for this whole dataset.

4.2 Control Classification Task: Medical Center Predictions

To evaluate to what extent FM embeddings encode the medical center from which images originate, prediction of the medical center of the image is evaluated as a control classification task.

Table 1: Composition of the `TCGA-2k` dataset: Tissue Source Site (TSS), Short Name and Project Code Combinations

TSS Short Name	BRCA	COAD	LIHC	LUSC	STAD
Asterand	✓	✓	✓	✓	✓
GPCC	✓	✓			✓
IGC	✓	✓	✓	✓	✓
ILSBio	✓	✓	✓		✓
MSKCC	✓	✓		✓	

4.3 Downstream Task Learning Algorithm and Setup

To ensure we evaluate the quality of FM embeddings, rather than the performance learned by a complex downstream model, we use one of the simplest possible downstream model architectures: k-nearest neighbor (*knn*, with k=3) unless otherwise specified, using cosine similarity as the distance function.

For image pre-processing and obtaining embeddings from the model output, the default choices for each model are followed. For further details, see the Appendix 9.

5 Results

Ten current pathology models were selected; see Appendix 9.1 for details on the selection. For each model, embeddings were generated for all patches in the dataset. The first result subsection below describes the prediction performance of cancer type and medical center, and a quantitative evaluation of the influence of medical center differences on FM-based predictions.

5.1 Embedding Space Structure and Robustness Index

As noted above, for each sample, we can measure the number of neighbors that have the same biological or confounding class as the sample, i.e., whether the neighbor has:

- the same cancer type as the sample, or
- the same medical center

The same information can be plotted as a function of the neighbor index k, by calculating the fraction of samples that have the same biological or confounding class. This way, we can summarize and visualize the neighborhood structure of the entire embedding space. For the knn distance metric, cosine distance is used, as cosine similarity is a common way to evaluate embedding similarity.

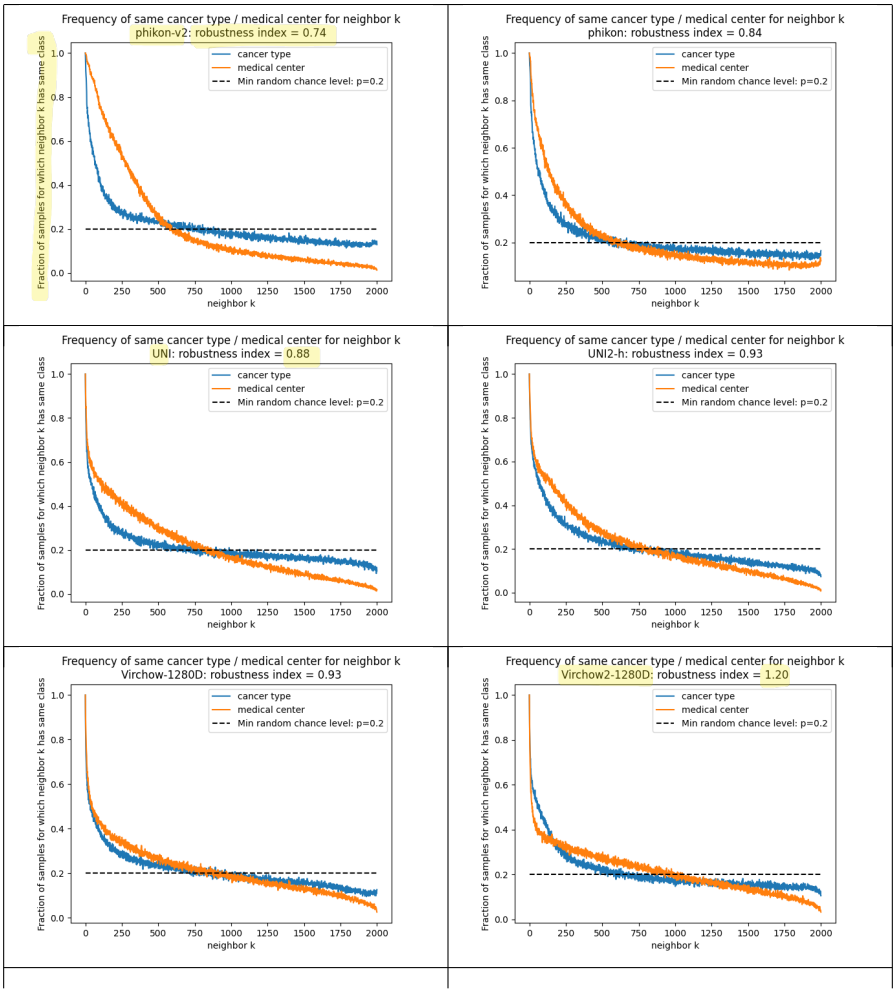


Figure 1: Fraction of samples for which the k-th neighbor has the same cancer type (blue) or medical center (orange), in order of increasing robustness. See Appendix 12 for all results. For all models, closeness in embedding space is strongly determined by whether the image comes from the same medical center. For all models except Virchow2, the medical center more strongly determines embedding proximity than the cancer type for the nearest 200 neighbors.

The robustness index can be calculated from these graphs by taking the leftmost values up to $k = 50$, taking the averages of the orange and blue lines, and then taking the ratio between these two average values. The robustness index reflects the extent to which biological factors such as cancer type dominate confounding factors such as medical center in the organization of the embedding space.

Figure 1 shows the results, and Figure 2 summarizes the resulting values of this metric for the models evaluated here. Some observations:

- According to this analysis, Phikon-v2, an expectedly improved version of Phikon, is less robust than Phikon (0.74 vs 0.84)
- Uni2-h is more robust than Uni (0.93 vs 0.88)
- Virchow2 is more robust than Virchow (1.2 vs 0.93), and than all other models. In fact, it is the only model with a robustness index above one, indicating that cancer type information dominates medical center information for the first 50 neighbors. Accordingly, this is the only model for which the blue line is above the orange line for the first 100+ neighbors
- One might expect the orange and blue lines to level off to an average value around the random chance level above a certain distance. Interestingly, this is not the case; even for the very furthest embeddings, an increase in distance still corresponds to lower probabilities of encountering the same cancer type or medical center. This shows that the organization by cancer type and medical center extends across the whole embedding space, and is a global phenomenon.

5.2 Quantification of the Influence of Medical Center Differences on FM-based Prediction Performance

Knn prediction performance was evaluated as follows:

- For all possible values of k, the accuracy of 5-class tissue type / cancer type classification was evaluated using 5-fold cross-validation (green lines).
- The accuracy of the 5-class medical center classification from which the patch originated was also evaluated (blue lines).

In addition to the above common metrics, we aim to measure the influence of the medical center on cancer type classification. To do so, we consider all samples (patches) for which the predicted cancer type class is incorrect. Given that knn operates by taking the class most common among the sample's k nearest neighbors in the training set (as determined by cross-validation here), we can identify the exact set of neighbors that contributed to the incorrect class prediction; this set consists of all neighbors that have the predicted (incorrect) class.

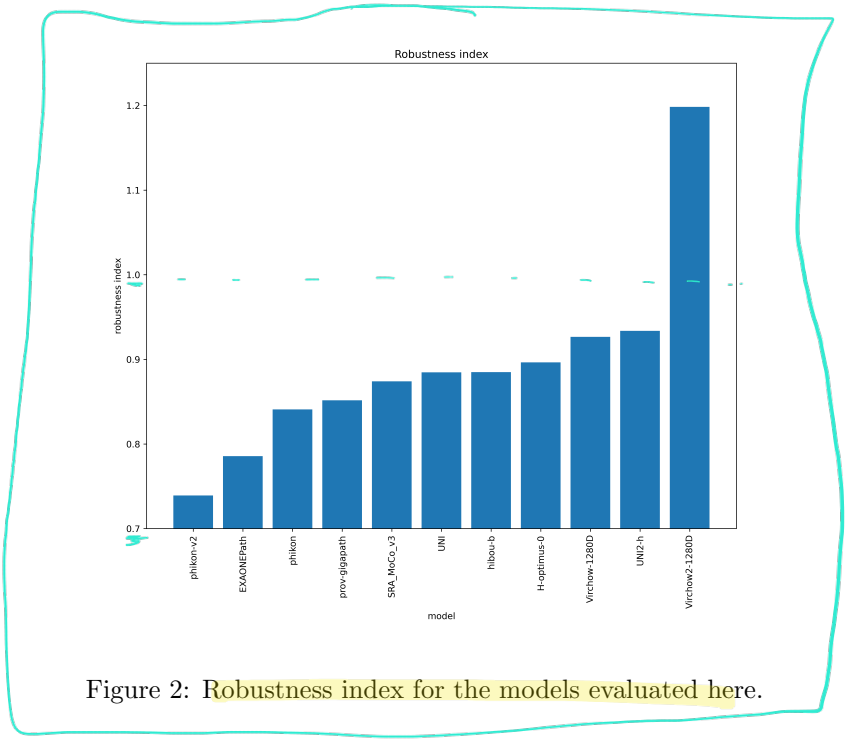


Figure 2: Robustness index for the models evaluated here.

If the FM were completely insensitive to differences between medical centers, the samples that contributed to the incorrect class prediction would be distributed randomly over the centers for that cancer type. To ensure this is the case, we restrict this analysis of the same-center confounders to the two classes that each have data for 5 centers (BRCA and COAD), so that the number of centers for this binary cancer type prediction is equal (5) for all sample points, resulting in a random chance level of occurring for each center of 1/5. Thus, the frequency of any center among these patches is expected to be 1/5.

If, on the other hand, the FM is sensitive to center differences, and tends to organize its embedding space by clustering patches from the same medical center together, then the set of neighboring patches described above will be more likely to come from the same center as the predicted sample. If the fraction of neighbors with the incorrect class that have the same center as the sample is higher than chance level (1/5), this indicates that the FM is sensitive to medical center differences, and that this sensitivity contributes to misclassification. We name such patches *same-center confounders*, as they confound the class prediction.

The figures in Figure 3 and Figure 4 show the results; see Appendix Section 11 for the complete set of these graphs. All models show clear sensitivity to the medical center; the incorrectly classified nearest neighbors are more likely to come from the same center as the predicted samples, in some cases to an extreme extent. E.g. for the Phikon-v2 model (blue line at the top in Fig. 4), the closest neighbors of the incorrectly predicted class are from the same center

in more than 95% of the cases. In other words, the embedding space is structured such that a knn classifier will effectively base its classification on whether the patch is from the same center, rather than whether it's from the same biological class.

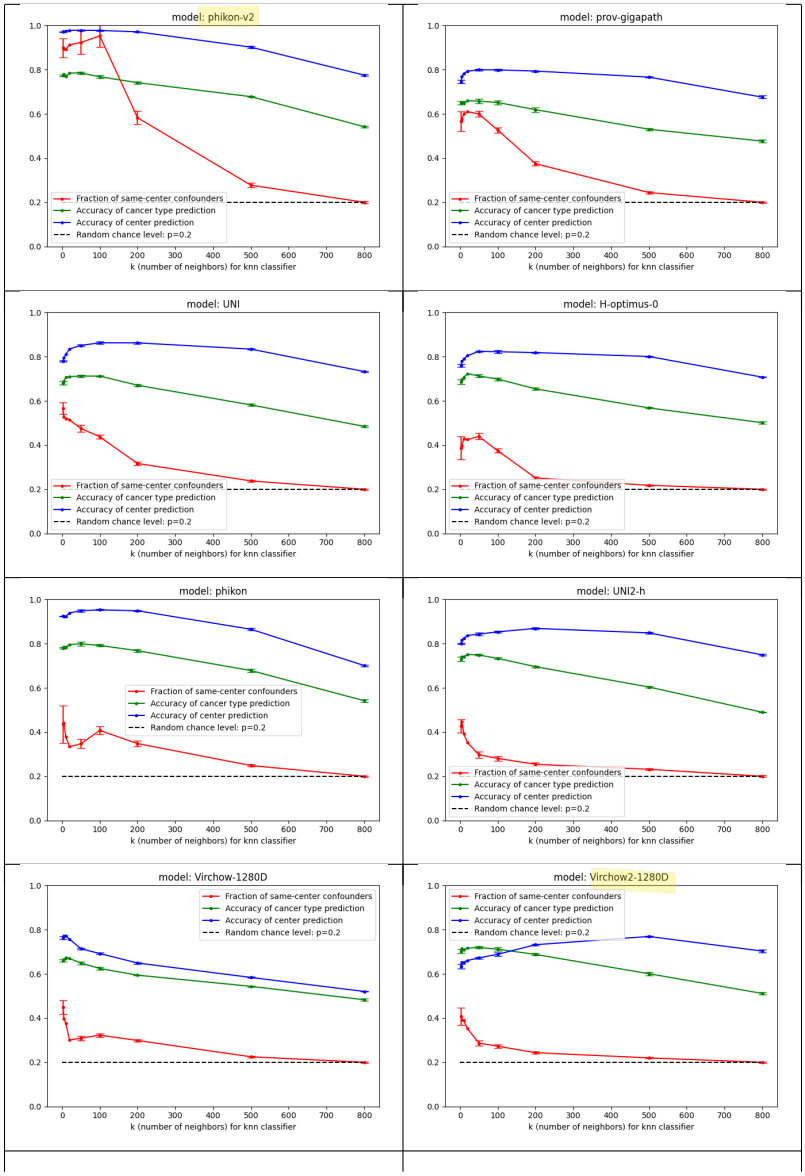


Figure 3: Fraction of same-center confounders for neighbors from the incorrectly predicted class (red); accuracy of tissue / cancer type prediction (green); accuracy of medical center prediction (blue), sorted by order of increasing center-robustness for selected FMs. All models show a substantial and significant influence of same-center confounders. See Appendix 9.3 for a complete overview.

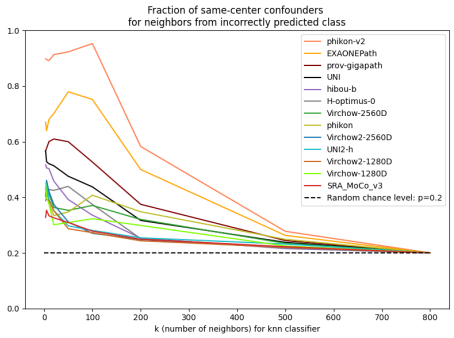


Figure 4: Fractions of same-center confounders for all models. All models are sensitive to these differences, some to a very high degree.

5.3 Visualization of the Embedding Space

To gain insight into the embedding spaces learned by the models, we use t-SNE [23] to project the high-dimensional embedding vectors to 2D. This results in 2D plots where each patch is represented by a dot in 2D space. The t-SNE method is run in an unsupervised manner; i.e. no label information about cancer type or medical center is used to obtain the 2D embeddings.

Given the 2D patch embeddings, we can color the embeddings using meta-information about the patches. Figures 5 and 6 show colorings of the 2D embeddings by cancer type (left column) and medical center (right column); note that the patch locations (the locations of the dots) in these left and right columns are identical.

The figures on the left show some degree of clustering by cancer type. No model achieves perfect separation; this may be unattainable, as patches are selected randomly from the foreground, and some patches may not contain sufficient information to identify the tissue of origin or the corresponding cancer type.

The figures on the right colored by medical center in general show increased clustering. The coloring for phikon-v2 shows extreme, almost perfect clustering by medical center; the medical center can be predicted with near-perfect accuracy based on the 2D embedding space location alone. This explains the high sensitivity to medical centers seen in the above result section 5.2.

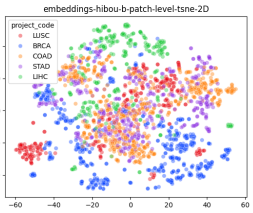
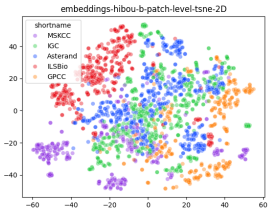
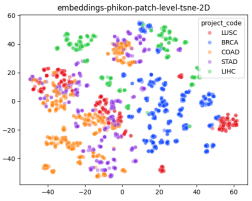
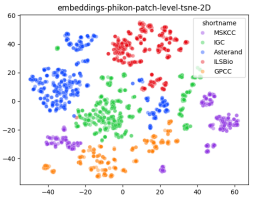
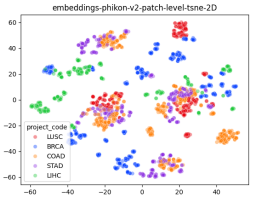
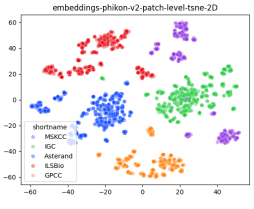
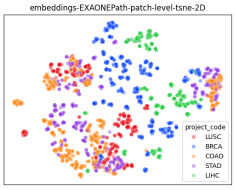
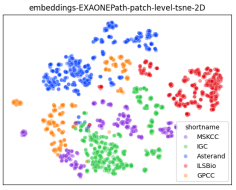
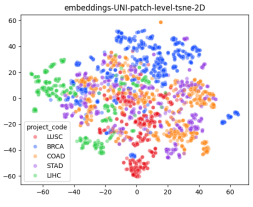
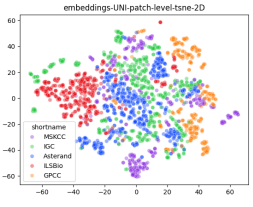
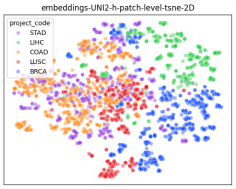
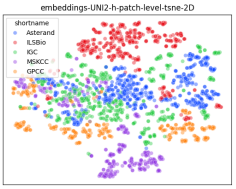
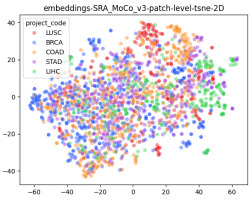
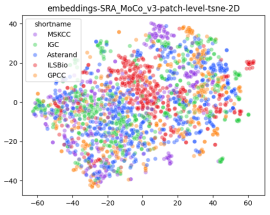
Model	Cancer Type	Medical Center
hibou-b		
phikon		
phikon-v2		
EXAONEPath		
UNI		
UNI2-h		
SRA_MoCo_v3		

Figure 5: Colorings of the t-SNE embeddings of all patches by cancer type (left) and medical center (right)

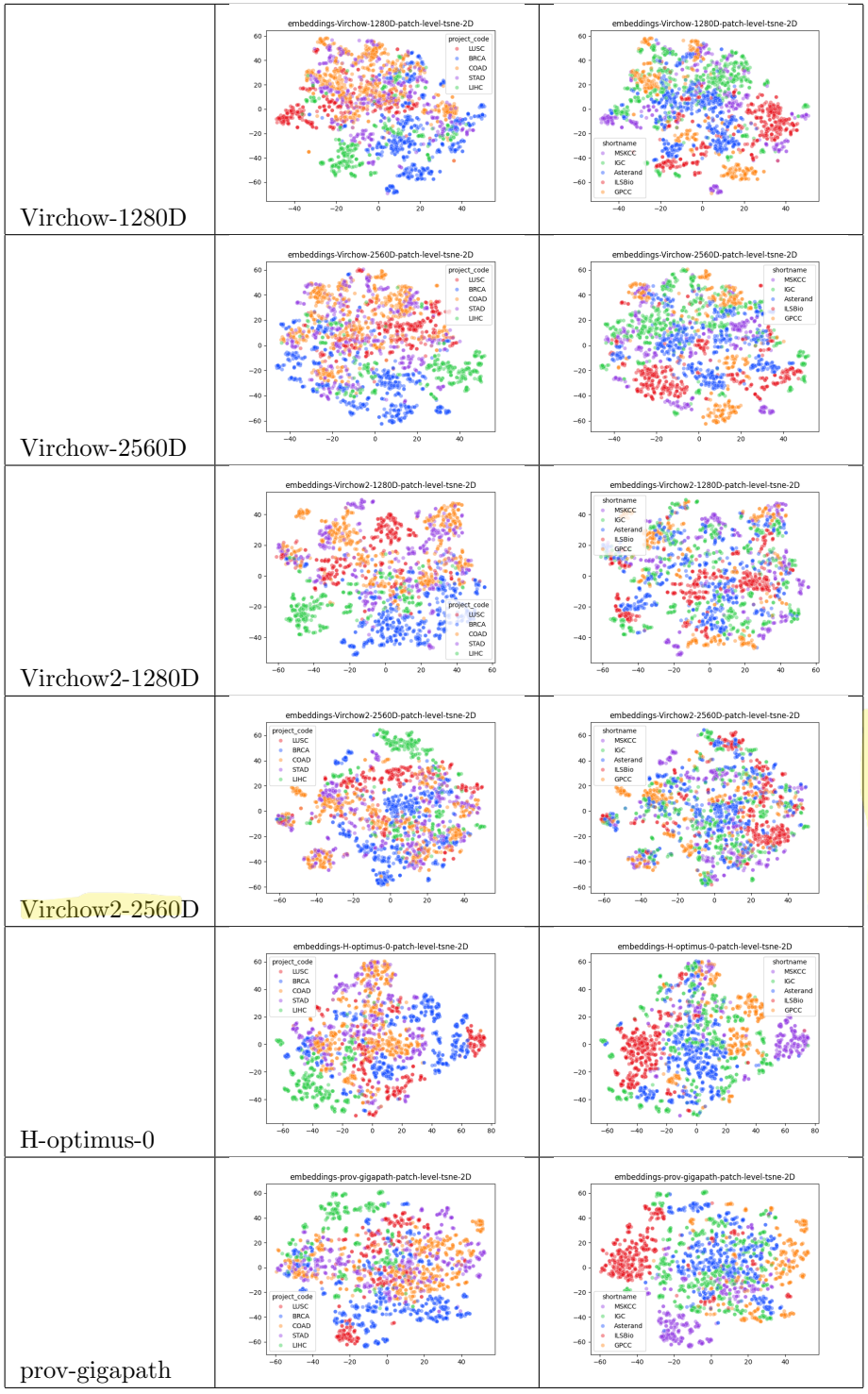


Figure 6: Colorings of the t-SNE embeddings of all patches by cancer type (left) and medical center (right)

5.4 Relation between Prediction Performance and Robustness

Ideally, a model should in our view demonstrate high prediction performance on relevant tasks, and at the same time show high robustness to irrelevant and confounding differences such as medical center differences. To evaluate what trade-off models achieve, we plot prediction performance on the cancer type classification task versus the prediction accuracy of the medical center, which relates inversely to robustness.

Figure 7 shows the results for prediction of cancer type and medical center from embeddings. The top row shows prediction using *knn* with k=3 and logistic regression. For logistic regression (top right), we see that all models except SRA predict the medical center to a very high degree of accuracy: EXAONEPath, Phikon and Phikon-v2 have cross-validated accuracies of 0.987, 0.987 and 0.993 respectively, the latter approaching perfect center prediction. See Table 2 for numerical results.

The prediction performance for cancer type appears to be correlated with that for medical center; this raises the question whether high cancer type prediction accuracy is based on confounding medical center features. It is therefore questionable whether this prediction performance will generalize to unseen, new medical centers (Out Of Distribution evaluation).

For *knn* on the full embeddings (top left), the accuracy of cancer type and medical center are reduced compared to logistic regression. There is a larger spread between the various results; and using *knn*, there is one model, Virchow2, that performs better on cancer type prediction than on medical center prediction, indicating a better relation between biologically relevant prediction performance and robustness. The bottom two graphs show analogous results, but based on using 2D t-SNE coordinates as input rather than the full embeddings.

5.5 Effect of Medical Center Influences on Regression

It could be argued that the strong influence of medical centers on prediction performance observed above is restricted to downstream models that use all dimensions of the embedding space; and that models using regression can select those dimensions that code for biologically relevant features such as cancer type while ignoring dimensions encoding confounding information such as medical centers.

To test whether medical center influences affect logistic regression, the following analysis is performed. For each sample wrongly predicted by a logistic regression model, the fraction of knn runs making a center-related prediction error is calculated. A knn prediction error is considered to be center-related if the majority of its neighbors has:

- an incorrect class label prediction for the sample, and
- the same medical center

Model	Mean Accuracy Cancer Type	Std Dev Cancer Type	Mean Accuracy Medical Center	Std Dev Med. Center
SRA_MoCo_v3	0.486	0.036	0.692	0.027
phikon	0.829	0.037	0.987	0.012
phikon-v2	0.83	0.038	0.993	0.007
UNI	0.713	0.034	0.956	0.013
UNI2-h	0.754	0.027	0.96	0.014
hibou-b	0.689	0.03	0.933	0.017
Virchow-1280D	0.727	0.038	0.932	0.022
Virchow2-1280D	0.786	0.03	0.957	0.016
H-optimus-0	0.767	0.038	0.948	0.019
prov-gigapath	0.711	0.031	0.934	0.019
EXAONEPath	0.808	0.037	0.987	0.011

Table 2: Mean and standard deviation for the accuracy of cancer type prediction and medical center prediction using the full embedding vectors as input and logistic regression as the learning method.

Results are shown in Figure 8.

6 Methods

6.1 Patch Extraction

For patch extraction, WSITools [33] was used. We submitted PR #11 to enable extracting a random selection of patches. Using this tool, patches of size 512x512 are extracted at the highest available resolution and downscaled by a factor 2 to 256x256. Given that the highest available resolution is typically 40X, this will typically result in a 20X resolution patch.

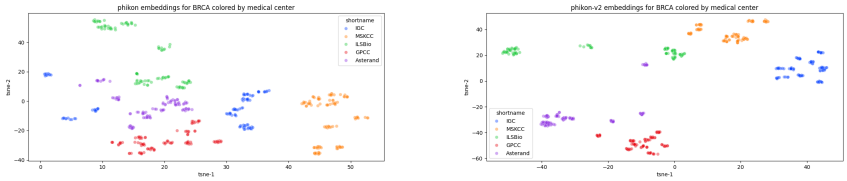


Figure 9: Embeddings for breast cancer colored by medical center for Phikon (left) and Phikon-v2 (right).

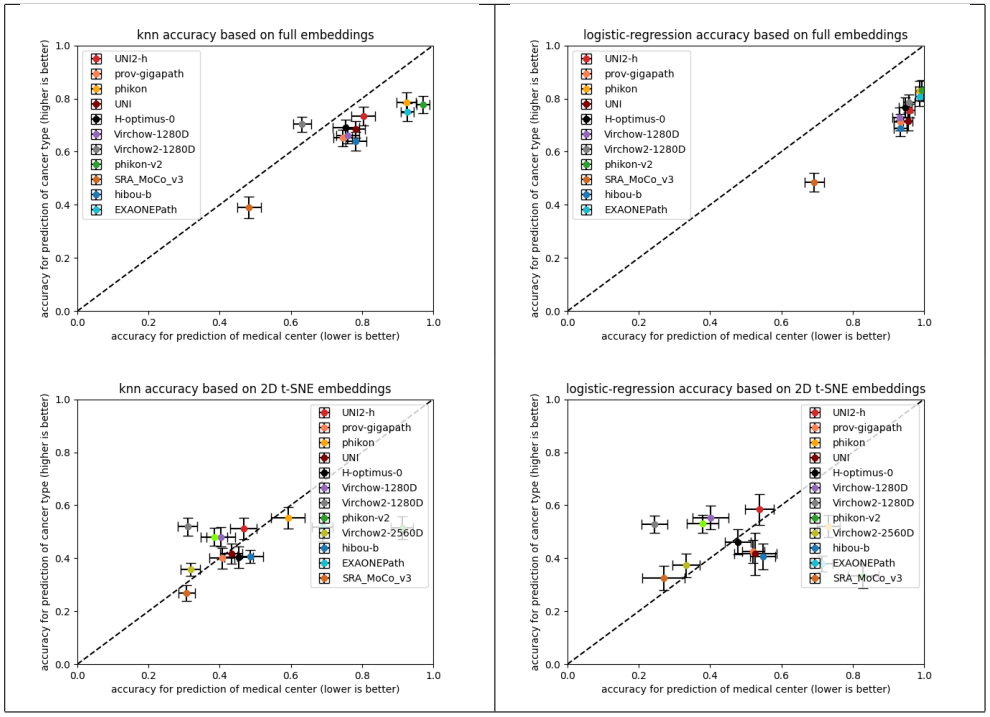


Figure 7: Top row: Accuracy of cancer type prediction vs. center prediction when using the full embedding vectors as input using knn (left) and logistic regression (right). Bottom row: Accuracy of cancer type prediction vs. center prediction when using the 2D t-SNE embedding vectors as input using knn (left) and logistic regression (right).

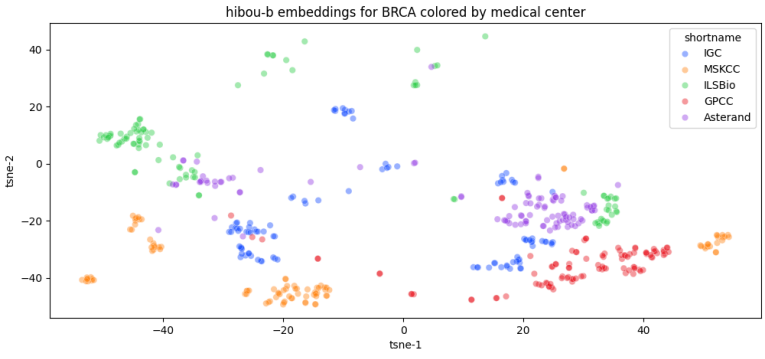


Figure 10: Embeddings for breast cancer colored by medical center for Hibou-B.

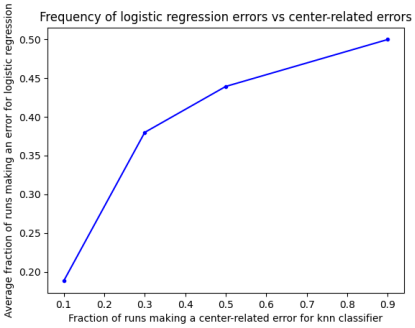


Figure 8: Relation between logistic regression errors and center-related knn errors. Samples that are more frequently misclassified by knn based on medical center are also more frequently misclassified by logistic regression, suggesting center similarities also affect logistic regression predictions.

7 Discussion

7.1 Patch-level vs WSI-level Prediction

Some patches may not contain sufficient information to determine the tissue of origin type / cancer type; thus, perfect classification may not be achievable at patch level, and higher levels of prediction accuracy may be achieved for an analogous WSI-level prediction task. The goal here however is not to maximize prediction accuracy, but rather to analyze the embedding space, and evaluate to what extent confounding center-related information influences classification decisions. A patch-level analysis provides the most direct way to link foundation model embeddings to medical centers; a WSI-level approach would introduce an extra level of indirection (e.g. a MIL layer) between the foundation model and the downstream model output that would influence this relation and thus potentially obfuscate the analysis.

7.2 Is Representation of Medical Center Information a Problem?

It may be argued that SSL algorithms are designed to capture any differences between images, that differences between medical centers result in real differences between the images, and that it is therefore to be expected, or even desirable that pathology FMs learn to recognize, distinguish and represent medical centers. And one may attempt to reduce the influence of medical centers in post-hoc adaptations of the FM, or in the downstream model.

Our belief however is that removing this influence is unlikely to be possible in an unbiased way; instead, it seems likely that the dimensions representing the medical center are not exactly orthogonal to dimensions representing biological information, and that it is therefore difficult or impossible to completely remove

the influence of medical centers post hoc. In other words, medically relevant properties such as cancer risk are likely to be correlated with medical centers, as patient cohorts differ between medical centers.

Furthermore, it was seen that the prediction performance for cancer type appears to be correlated with that for medical center; this raises the question whether high cancer type prediction accuracy is based on confounding medical center features. It is therefore questionable whether this prediction performance will generalize to unseen, new medical centers (Out Of Distribution evaluation).

The application of AI in the medical domain brings with it a high degree of responsibility; if biases related to medical centers affect model predictions, and thereby influence patient diagnosis, treatment options, and outcomes, then it is the responsibility of practitioners in the medical AI domain to measure and reduce these influences to the maximal feasible extent.

8 Conclusion

In this work, robustness is viewed as insensitivity to confounding features. The Robustness Index, a novel metric to evaluate the degree to which biological information dominates confounding information such as the medical center, was introduced. Foundation models were seen to differ significantly in robustness according to this metric. Uni2-h and Virchow2 were found to be most robust, and Virchow2 was the only model so far with a robustness index above one, meaning biological information (cancer type) dominates confounding information (medical center) across the $k = 50$ nearest neighbors.

It was seen that distance in embedding space strongly correlates with both the probability of encountering same-cancer-type neighbors and same-medical-center neighbors. This influence is not just local, but was seen to extend across the entire embedding space.

Using the notion of *same-center confounders*, the impact of medical centers on prediction was evaluated, and it was found that all pathology foundation models evaluated here represent medical centers to a large extent.

A 2D projection of the embedding space was visualized. The resulting images show visually that the organization of the embedding space shows a clustering by medical center; more strongly so than a clustering by tissue or cancer type.

The robustness index and the other analysis techniques described in this work are intended as tools that may enable the development of more robust pathology foundation models.

9 Appendix

9.1 Model Selection

Ten publicly available pathology foundation models were selected for evaluation, focusing on patch-level models. In addition, SRA_MoCo_v3 [34] was evaluated; while this model has been trained on a small single-tissue dataset, and can thus

not be viewed as a foundation model, it is aimed at providing a more robust model. The selection consists of the following models:

- Phikon [5]
- Phikon-v2 [6]
- EXAONEPath [18]
- Prov_gigapath [13]
- SRA_MoCo_v3 [34]
- UNI [10]
- UNI2-h [10]
- Hibou [7]
- H-Optimus-0 [12]
- Virchow [8]
- Virchow2 [9]

9.2 Embedding Generation

Embeddings are generated using the default approach for each model. For Virchow and Virchow2, this means the average of the patch tokens is concatenated to the class token, resulting in a 2560-dimensional embedding, indicated with '-2560D'. To check whether this expansion of the embedding space affects performance, results with just the class token, indicated with '-1280D', are included for Virchow and Virchow2 as well. For all remaining models, the class token is the standard output, and is used here.

9.3 Fraction of Same-Center Confounders: Full Results

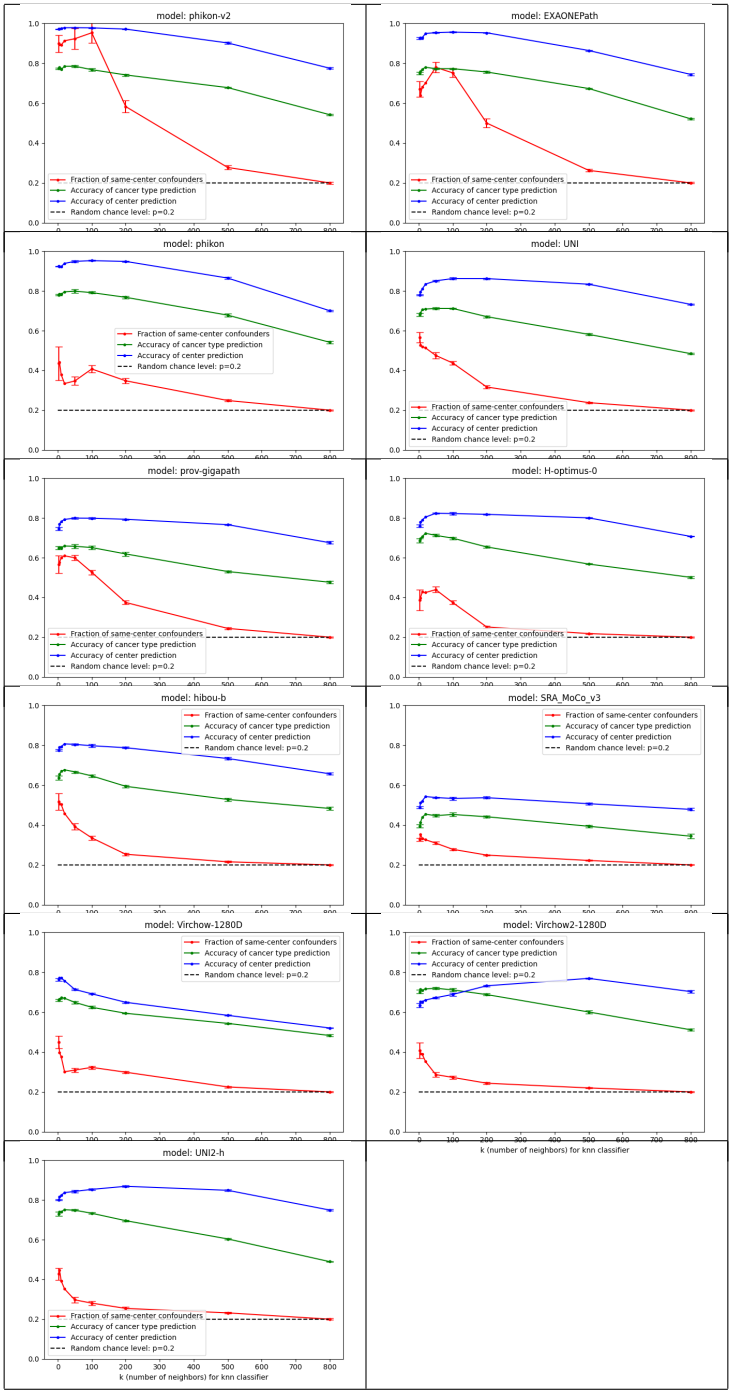


Figure 11: Fraction of same-center confounders: (i) accuracy of tissue of origin / cancer type prediction (green), (ii) the accuracy of medical center prediction (blue), and (iii) fraction of same-center confounders. All models show a substantial and significant influence of same-center confounders.

9.4 Frequency Same Cancer Type / Medical Center: Full Results

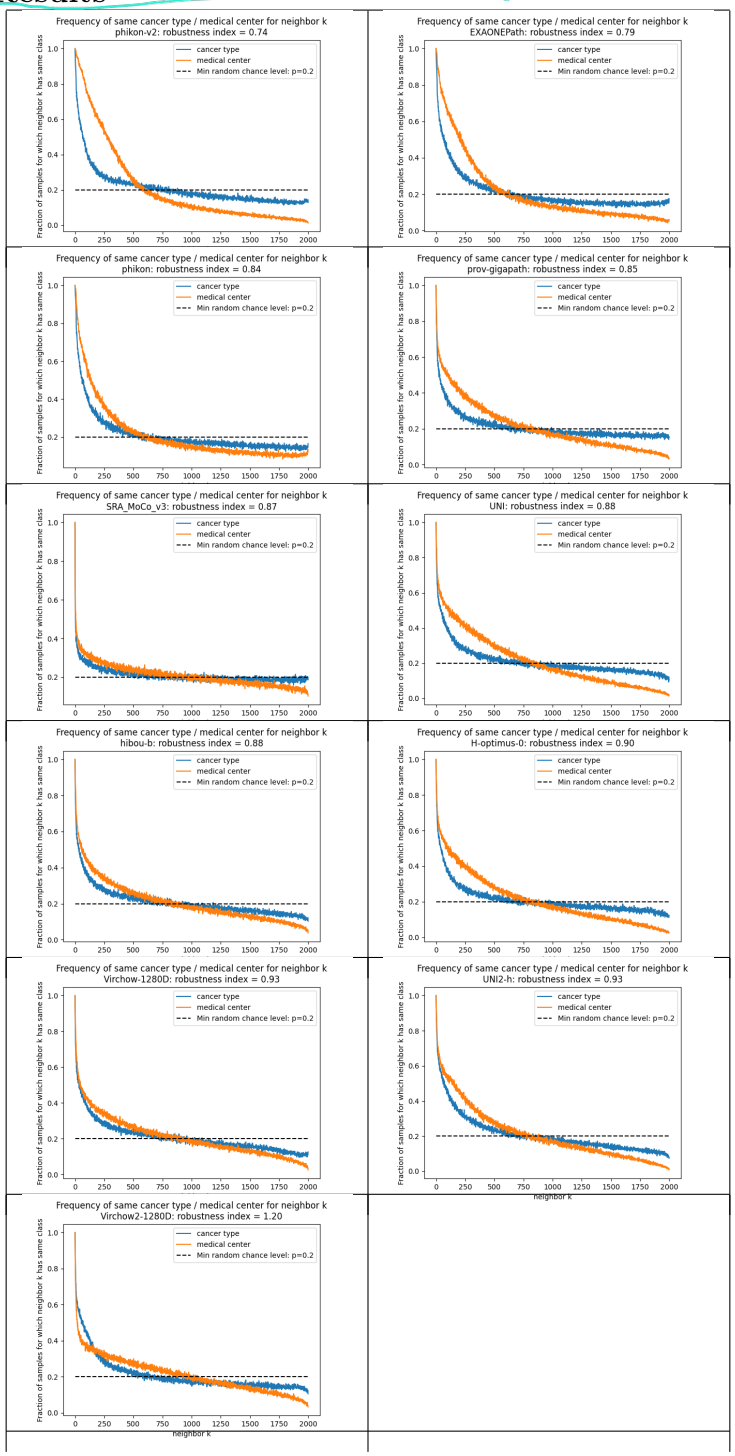


Figure 12: Fraction of samples for which the k-th neighbor has the same cancer type (blue) or medical center (orange), in order of increasing robustness

10 Online resources

We intend to make the patch dataset constructed and used in this work available online. An extended version of this work, combined with related simultaneous research from the TU Berlin BIFOLD group and Aignostics, is in preparation.

11 Acknowledgements

The authors would like to thank Hans Pinckaers, Jonas Dippel, Alexander Möllers, Maximilian Alber, other colleagues at Aignostics, and the TU Berlin BIFOLD group for valuable suggestions that improved the article. The results presented here are based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov. We thank Bodong Zhang for kindly providing the SRA_MoCo_v3 model for inclusion in this evaluation.

References

[1] 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.

[2] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations, 2020.

[3] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv preprint arXiv:2012.03583*, 2020.

[4] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G 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*, pages 16144–16155, 2022.

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

[6] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv e-prints*, 2024.

[7] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology, 2024.

[8] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A million-slide digital pathology foundation model. *arXiv:2309.07778v5*, 2024.

[9] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024.

[10] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[11] Hamid Manoochehri, Bodong Zhang, Beatrice S. Knudsen, and Tolga Tasdizen. Sra: A novel method to improve feature embedding in self-supervised learning for histopathological images, 2024.

[12] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024.

[13] Hongyi Xu, Naoto Usuyama, Jitendra Bagga, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024.

[14] Khaled Saab et al. Capabilities of gemini models in medicine, 2024.

[15] Google. Path foundation. `https://github.com/Google-Health/imaging-research/tree/master/path-foundation`, 2024. Accessed: 2024-11-24.

[16] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, Georg Gerber, Anil V Parwani, and Faisal Mahmood. Towards a visual-language foundation model for computational pathology, 2023.

[17] Shekoofeh Azizi et al. Robust and efficient medical imaging with self-supervision, 2022.

[18] Juseung Yun, Yi Hu, Jinhyung Kim, Jongseong Jang, and Soonyoung Lee. Exaonepath 1.0 patch-level foundation model for pathology. *arXiv preprint arXiv:2408.00380*, 2024.

[19] Gavino Faa, Ferdinando Coghe, Andrea Pretta, Massimo Castagnola, Peter Van Eyken, Luca Saba, Mario Scartozzi, and Matteo Fraschini. Artificial intelligence models for the detection of microsatellite instability from whole-slide imaging of colorectal cancer. *Diagnostics*, 14(15), 2024.

[20] Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464, July 2022.

[21] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, December 2019.

[22] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton JV Campbell, Liron Pantanowitz, HR Tizhoosh, and Shahryar Rahnamayan. Biased data, biased ai: deep networks predict the acquisition site of tcga images. *Diagnostic pathology*, 18(1):67, 2023.

[23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[24] 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.

[25] Jonah Kömen, Hannah Marienwald, Jonas Dippel, and Julius Hense. Do histopathological foundation models eliminate batch effects? a comparative study. *arXiv preprint arXiv:2411.05489*, 2024.

[26] Matouš Elphick, Samra Turajlic, and Guang Yang. Are the latent representations of foundation models for pathology invariant to rotation? *arXiv preprint arXiv:2412.11938*, 2024.

[27] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva de Souza, Alexi Baidoshvili, Geert Litjens, Bram van Ginneken, Iris Nagtegaal, and Jeroen van der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 160–163, 2017.

[28] Milad Sikaroudi, Shahryar Rahnamayan, and Hamid R Tizhoosh. Hospital-agnostic image representation learning in digital pathology. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3055–3058. IEEE, 2022.

[29] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 201–209. Springer, 2018.

[30] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton JV Campbell, Liron Pantanowitz, HR Tizhoosh, and Shahryar Rahnamayan. Biased data, biased ai: deep networks predict the acquisition site of tcga images. *Diagnostic pathology*, 18(1):67, 2023.

[31] 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.

[32] Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology, 2025.

[33] Jun Jiang. Whole slide image pre-processing tools for deep learning tasks. `https://github.com/smujiang/WSITools`, 2019.

[34] Hamid Manoochehri, Bodong Zhang, Beatrice S. Knudsen, and Tolga Tasdizen. SRA: A Novel Method to Improve Feature Embedding in Self-supervised Learning for Histopathological Images. *arXiv e-prints*, page arXiv:2410.17514, October 2024.