A CONCEPT-BASED INTERPRETABILITY BENCHMARK FOR PATHOLOGY FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology foundation models are usually compared by downstream accuracy, which can hide differences in what their representations encode and what downstream predictors use. Existing concept-based methods often use concepts or explanations that depend on the model being analysed, which makes direct cross-model comparison difficult. We introduce a model-independent framework that compares foundation models using the same predefined pathology concepts. We derive 23 morphology factors from structured pathology reporting standards and ground them in $9{,}672$ caption-asserted factor–panel examples from $4{,}586$ open-access articles. Using factor-specific probes on 16 frozen encoders, we measure factor accessibility, or how well each factor can be linearly recovered from a representation. Encoders differ in both overall accessibility and their factor-level profiles, and models with similar overall accessibility can differ substantially in which factors are most accessible. We then remove factor-associated information using the corresponding factor directions and measure factor sensitivity, or how much downstream performance changes after removal. In melanoma relapse prediction, accessibility is only weakly related to sensitivity. Across three melanoma prediction tasks, sensitivity patterns are consistent across retraining but differ across encoders and tasks. Our framework therefore provides a common concept space for comparing both what foundation models represent and what factor-associated information downstream predictors are sensitive to, turning concept-based interpretability into a tool for comparative foundation-model analysis.

1 INTRODUCTION

Pathology foundation models are increasingly used as general-purpose representations for downstream tasks and are compared by training a predictor on top of each representation and measuring task performance (Neidlinger et al., 2026). Downstream accuracy does not reveal which pathology-relevant information is accessible in each representation: two encoders can reach similar accuracy while making different information accessible, and similar downstream performance can rely on different information. We therefore compare encoders using the same predefined morphology factors.

We address this problem with a model-independent concept space defined before any foundation model is analysed. Structured pathology reporting standards from the International Collaboration on Cancer Reporting (ICCR) and the Royal College of Pathologists (RCPath) (International Collaboration on Cancer Reporting, 2026; Royal College of Pathologists, 2026) provide a natural source of such concepts, specifying morphological findings such as mitotic activity, nuclear pleomorphism, tumour budding and lymphovascular invasion. Because these factors are defined by pathology practice rather than by a particular model, the same concepts can be used to compare encoders with different architectures, feature dimensions and pretraining objectives. Reporting standards define the concepts but supply no image examples, so we ground these factors in open-access scientific figures whose captions describe the corresponding morphology, producing visual supervision for 23 factors (Figure 1).

This benchmark lets us measure **factor accessibility**: how well each morphology factor can be linearly recovered from a frozen encoder. Applying the same factor-specific probes to 16 pathology and general-purpose vision encoders reveals substantial differences not only in overall morphology accessibility, but also in which factors individual encoders make accessible. Models with nearly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

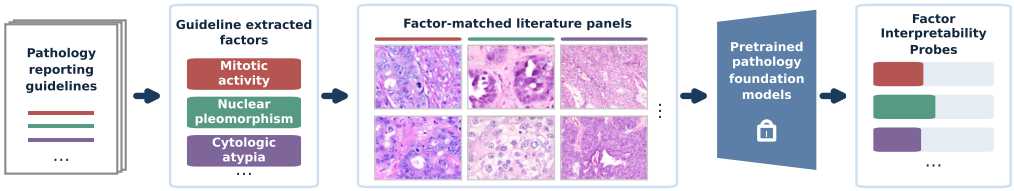

Figure 1: The benchmark pipeline.

identical aggregate accessibility can have different factor-level profiles. The benchmark reveals factor-level differences between encoders that are not visible from downstream performance alone.

Accessibility does not tell us whether a downstream predictor uses the accessible information. We therefore use the corresponding factor directions to intervene on trained downstream models: for each factor, we remove information along its learned linear direction and measure the resulting change in downstream performance. We call this **factor sensitivity**. In melanoma, accessibility is only weakly related to sensitivity, and sensitivity profiles change with both the encoder and the downstream task while remaining reproducible across retraining. What a foundation model represents and what a downstream predictor uses are therefore distinct properties. Using the same externally defined factor space for both probing and intervention lets us measure both within a single framework.

Our contributions are:

1. **A model-independent semantic benchmark for pathology foundation models.** We derive 23 morphology factors from structured pathology reporting standards and ground them in panel-level examples from open-access scientific literature, providing the same externally defined concept space for heterogeneous encoders.
2. **A factor-level characterization of foundation-model representations.** Using the same probe protocol across 16 frozen encoders, we show that models differ not only in overall morphology accessibility but also in their factor-specific representation profiles, including models with similar aggregate performance.
3. **A shared concept space for connecting representation content to downstream use.** We use the corresponding factor directions for downstream interventions and show that accessibility and sensitivity are distinct: which factor-associated information a predictor is sensitive to varies with both the encoder and the downstream task.

2 Related Work

Concept-based interpretability. Concept-based methods explain model representations using human-interpretable variables. Existing approaches define concepts from manually collected examples (Kim et al., 2018), supervised concept bottlenecks (Koh et al., 2020), language-aligned representations (Yuksekgonul et al., 2023), or concepts discovered automatically from model activations (Ghorbani et al., 2019). These approaches may use concepts that depend on the model, the supervision source, or the representation type: manually defined concept sets are difficult to scale, language-based approaches require compatible multimodal representations, and discovered concepts can differ from one model to another. We instead use a fixed concept factor set extracted from pathology reporting standards and ground those concepts with visual examples from the scientific literature. This gives us the same externally defined factors for every encoder, independent of model architecture or pretraining objective.

Interpreting pathology models. Prior work in computational pathology has explained models through pathologist-defined morphological features (Gamble et al., 2021), language-aligned concepts (Kapse et al., 2025; Lazard et al., 2026), and latent directions discovered within pathology representations (Kim et al., 2026a). These approaches can reveal meaningful morphology in indi-

vidual models, but the interpretation is often tied to a particular model, task, or source of supervision. That makes cross-model comparison difficult, because differences in the explanations may come from differences in the concept space itself. Our approach fixes the morphology factors before analysing any model and learns the corresponding factor directions separately within each encoder. This lets us compare vision-only pathology models with different architectures, feature dimensions, and pretraining strategies using the same semantic reference points.

Evaluating pathology foundation models. Pathology foundation models are compared through downstream task performance (Neidlinger et al., 2026) or global properties of their representations, such as robustness (de Jong et al., 2025) or representational similarity (Mishra & Lotter, 2026). These evaluations tell us whether one model performs better than another, or whether two representations are globally similar, but they do not tell us what pathology-related information is accessible in each representation. A task-level score can therefore hide meaningful representational differences between models with similar performance. We address this by evaluating encoders along the same set of morphology factors, producing an interpretable profile of which factors are more or less accessible. We then use the same factors to intervene on downstream representations, allowing us to connect differences in representation content to differences in downstream model behavior.

3 Constructing the morphology benchmark

We construct the benchmark by grounding morphology factors from structured pathology reporting standards in examples from the scientific literature. We first derive a model-independent set of histologically assessable factors, then retrieve open-access figures that describe those factors, split compound figures into panels, and pair each panel with its corresponding subcaption. This yields 11,564 matched panel–caption pairs across 35 candidate factors from 5,932 articles. After applying the benchmark inclusion criteria, we retain 23 factors with 8,875 present and 797 absent factor–panel examples from 4,586 articles.

3.1 Deriving the morphology factors

We collect 75 ICCR cancer dataset documents (International Collaboration on Cancer Reporting, 2026) and 91 RCPath reporting proformas (Royal College of Pathologists, 2026). We normalize their reporting elements into histologically assessable factors by merging synonymous entries and removing administrative, non-histologic, and non-visual elements.

The resulting factor set spans cellular morphology, invasion and spread, architecture, tissue state, and host or inflammatory response. Examples include mitotic activity, nuclear pleomorphism, tumour budding, lymphovascular invasion, tumour necrosis, and inflammatory infiltrate. Additional construction and exclusion rules are provided in Appendix A.1.

3.2 Grounding factors in scientific figures

Terminology in reporting standards does not always match the language used in scientific figure captions. We therefore define factor-specific search terms and use them to retrieve candidate figures from open-access articles. Because figures frequently contain multiple panels showing different findings, we split compound figures into individual panels and pair each panel with its corresponding subcaption.

For each candidate factor–panel pair, we classify the subcaption as asserting that the morphology is **present**, **absent**, or not usable for that factor. We retain high-confidence present and absent examples for the benchmark. Full retrieval and assertion-labelling details are provided in Appendix A.

3.3 Benchmark composition and validation

After literature matching and assertion labelling, we retain 23 factors with sufficient supervision for evaluation (Table 1). The benchmark contains 8,875 present and 797 absent factor–panel examples from 4,586 articles. Eight factors contain sufficient explicitly absent examples to support direct present-versus-absent evaluation.

Table 1: Retained supervision per benchmark factor; labelled = present + absent.

Group	Factor	Present	Absent	Labelled examples
Cellular morphology	cytologic atypia	1,849	272	2,121
	mitotic activity	1,085	149	1,234
	nuclear pleomorphism	753	35	788
	sarcomatoid features	146	5	151
Invasion and spread	invasive component	765	33	798
	dysplasia	316	45	361
	in-situ component	228	1	229
	lymphovascular invasion	185	31	216
	perineural invasion	110	19	129
	nodal metastasis	103	3	106
Architecture	solid	367	0	367
	papillary	333	4	337
	cribriform	203	3	206
	trabecular	184	0	184
	infiltrative growth	179	3	182
	tumour budding	106	10	116
Tissue state	tumour necrosis	594	60	654
	histologic calcification	324	10	334
	ulceration	264	13	277
Host and inflammatory	vasculitis	268	78	346
	inflammatory infiltrate	196	9	205
	tumour-infiltrating lymphocytes	166	9	175
	interstitial inflammation	151	5	156

We audit 1,000 automatically assigned factor–caption labels using a different model family. Caption–label agreement is 99.1% after reweighting to the benchmark composition, including 99.3% for present assertions and 97.0% for absent assertions. Most disagreements arise from ambiguous captions rather than incorrect factor assignment. Additional validation is provided in Appendix A.4.

4 Characterizing morphology representations across foundation models

We use the benchmark to measure **factor accessibility**: how well each morphology factor can be linearly recovered from a frozen encoder. We evaluate 16 frozen encoders (Chen et al., 2024a; Vorontsov et al., 2024; Zimmermann et al., 2024; Xu et al., 2024; Saillard et al., 2024; Karasikov et al., 2025; Filiot et al., 2023; 2024; kaiko.ai et al., 2024; Kang et al., 2023; Siméoni et al., 2025; He et al., 2016) using the same probe protocol across models. We fit a separate logistic-regression probe for each encoder–factor pair on frozen panel embeddings using article-grouped five-fold cross-validation. Feature standardization, regularization and evaluation are identical across encoders; each encoder keeps its own documented input preprocessing. Full probe settings are given in Appendix B.

We evaluate accessibility in two complementary settings. The **broad benchmark** uses all 23 factors and measures how well factor-positive panels can be distinguished from the broader literature corpus. The **explicit-polarity benchmark** uses the eight factors with sufficient present and absent examples and measures direct present-versus-absent discrimination. We use the broad benchmark primarily for encoder-level comparison and explicit polarity for factor-level analysis.

Figure 2 summarizes the broad 23-factor benchmark and its context-matched controls. Table 2 reports the full encoder comparison under both broad and explicit-polarity evaluation.

4.1 Foundation models differ in overall morphology accessibility

Foundation models differ substantially in morphology accessibility from their frozen features. On the broad 23-factor benchmark, macro AUC ranges from 0.641 for ImageNet-pretrained ResNet-50 to 0.786 for UNI2 (Figure 2a). The strongest pathology encoders remain ahead as comparison

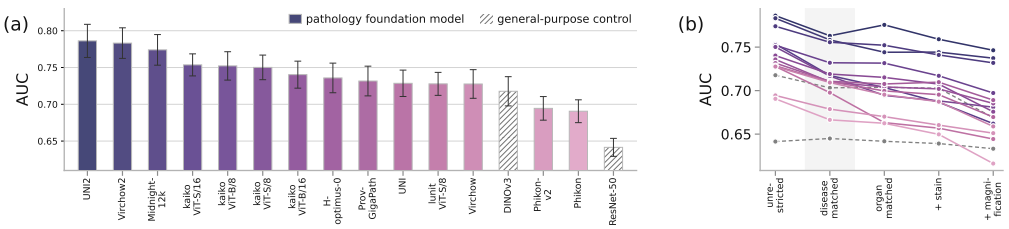

Figure 2: Guideline-factor accessibility: **(a)** unrestricted, **(b)** context-matched.

panels are progressively matched on disease, organ, stain, and magnification (Figure 2b). Under the strictest matching condition, the mean decrease is 0.055 AUC, while the relative encoder ordering remains largely preserved.

The same encoder-level structure appears under explicit-polarity evaluation. UNI2 reaches 0.702 macro AUC across the eight eligible factors, followed by Virchow2 at 0.694 and Midnight-12k at 0.688. The general-purpose DINOv3 encoder reaches 0.639 and ResNet-50 0.579. Across all 16 encoders, the ranking obtained from the broad 23-factor benchmark correlates strongly with the explicit-polarity ranking (Spearman $\rho = 0.897$).

The differences are also visible within model families. UNI2 improves over UNI by 0.074 polarity AUC and is higher on 6 of 8 factors; Virchow2 improves over Virchow by 0.084 and is higher on 7 of 8. UNI2 exceeds DINOv3 by 0.063 and ResNet-50 by 0.123.

Table 2: Morphology accessibility across 16 frozen encoders. Values are macro AUC over factors; † marks the two non-pathology encoders.

	Broad, 23 factors					**Polarity**
Encoder	Unrestr.	Disease	Organ	+Stain	+Mag.	8 factors
UNI2	.786	.763	.776	.759	.746	.702
Virchow2	.783	.758	.744	.744	.737	.694
Midnight-12k	.774	.756	.752	.741	.732	.688
kaiko ViT-S/16	.754	.717	.704	.687	.662	.661
kaiko ViT-B/8	.752	.732	.732	.717	.697	.664
kaiko ViT-S/8	.750	.717	.694	.688	.681	.639
kaiko ViT-B/16	.740	.719	.715	.707	.676	.630
H-optimus-0	.736	.710	.704	.702	.685	.643
Prov-GigaPath	.732	.711	.708	.710	.689	.637
UNI	.729	.710	.700	.695	.670	.628
Lunit ViT-S/8	.728	.698	.663	.657	.644	.643
Virchow	.728	.709	.695	.687	.659	.610
DINOv3†	.718	.703	.704	.703	.669	.639
Phikon-v2	.694	.679	.670	.660	.651	.565
Phikon	.691	.666	.662	.650	.616	.605
ResNet-50†	.641	.645	.642	.639	.633	.579

4.2 Some morphology factors are consistently more accessible than others

Section 4.1 compares encoders after averaging over morphology factors. We now ask the complementary question: are some morphology factors consistently more accessible than others across encoders? To isolate factor-level differences, we use the explicit-polarity benchmark, which directly compares present and absent examples of the same morphology.

Averaged across the 16 encoders, accessibility ranges from 0.733 AUC for mitotic activity to 0.532 for lymphovascular invasion (Table 3). Tumour necrosis and cytologic atypia are also relatively accessible, at 0.686 and 0.660, whereas nuclear pleomorphism, dysplasia, and vasculitis fall in the intermediate 0.616–0.627 range.

These differences indicate that factor accessibility varies substantially across morphology. Mitotic activity is recoverable from most encoders, whereas lymphovascular invasion remains difficult across nearly all of them.

Factor-level comparisons are sensitive to the evaluation setting: factor rankings differ substantially between broad and explicit-polarity evaluation ($\rho = -0.357$). We therefore use explicit polarity for all factor-level analyses.

Table 3: Explicit-polarity accessibility by morphology factor. Values are the mean over the 16 encoders $\pm$ its standard error.

Factor	Present	Absent	AUC	AP
Mitotic activity	1,085	149	0.733 ± 0.013	0.947 ± 0.003
Tumour necrosis	594	60	0.686 ± 0.013	0.949 ± 0.003
Cytologic atypia	1,849	272	0.660 ± 0.009	0.924 ± 0.003
Invasive component	765	33	0.640 ± 0.026	0.972 ± 0.003
Nuclear pleomorphism	753	35	0.627 ± 0.017	0.972 ± 0.002
Dysplasia	316	45	0.620 ± 0.020	0.907 ± 0.008
Vasculitis	268	78	0.616 ± 0.010	0.841 ± 0.005
Lymphovascular invasion	185	31	0.532 ± 0.012	0.869 ± 0.005

4.3 Encoders have morphology-specific representation profiles

The results so far reveal two global effects: some encoders have higher morphology accessibility overall, and some factors are more accessible across nearly all encoders. We next ask whether these two effects fully explain the accessibility matrix, or whether individual encoders have additional strengths and weaknesses for particular morphology factors.

Let $A_{m,f}$ denote the accessibility of factor f for encoder m. We decompose the matrix into an overall mean, an encoder effect, a factor effect, and an encoder–factor interaction:

$$A_{m,f} = \bar{A} + (\bar{A}_{m,\cdot} - \bar{A}) + (\bar{A}_{\cdot,f} - \bar{A}) + I_{m,f}.$$

The interaction term $I_{m,f}$ measures the part that cannot be explained by an encoder being globally stronger or a factor being globally easier. A large interaction therefore indicates morphology-specific differences between encoder representations.

Under explicit polarity, factor identity explains $\mathbf{44.3}\%$ of the total variation and encoder identity $\mathbf{20.2}\%$, while the encoder–factor interaction explains $\mathbf{35.5}\%$ (Table 4). Thus, more than one third of accessibility variation depends on the specific encoder–factor combination.

Table 4: Variance decomposition of the accessibility matrix.

Evaluation	Factor	Encoder	Interaction
Broad, 23 factors	76.9%	14.5%	8.6%
Broad, same 8 factors	76.9%	14.7%	8.4%
Explicit polarity, 8 factors	44.3%	20.2%	35.5%

UNI2 and Virchow2 have similar macro polarity AUC, 0.702 and 0.694, but differ in which factors they make accessible: UNI2 is higher on invasive component (0.785 versus 0.651), Virchow2 on mitotic activity (0.820 versus 0.761). Their aggregate scores conceal opposing factor-level differences. Likewise, UNI2 has a clear aggregate advantage over DINOv3 and UNI, but exceeds DINOv3 on only five of eight factors and UNI on six of eight.

Aggregate accessibility therefore does not capture the factor-level differences between encoders: they differ not only in overall morphology accessibility, but in which specific factors are most accessible. In Section 5 we ask whether these accessibility differences translate into differences in downstream sensitivity.

5 WHAT MORPHOLOGY INFORMATION IS A DOWNSTREAM MODEL SENSITIVE TO?

Section 4 showed that foundation models differ in their factor-accessibility profiles. We now ask whether those differences affect what a downstream predictor is sensitive to. We study this on melanoma relapse prediction with the same 16 encoders: for each encoder we train the same downstream model, and then selectively remove the representation component associated with one morphology factor at a time. If removing a factor changes downstream performance, we call the predictor **sensitive** to that factor, and we quantify **factor sensitivity** by the signed AUC drop $\mathrm{AUC}_{\mathrm{original}} - \mathrm{AUC}_{\mathrm{erased}}$.

We run these interventions for the eight factors with enough explicitly present and explicitly absent examples to fit a polarity direction, removing that direction (Belrose et al., 2023) from every patch representation and rerunning the trained classifier without retraining either the encoder or the downstream model. All downstream analyses use the same $1{,}741$ VisioMel melanoma slides (Société Française de Pathologie et al., 2025) for which relapse, ulceration and Breslow thickness are recorded; Breslow is the binary endpoint thickness $\geq$ 2mm; the downstream model is a common ABMIL architecture (Ilse et al., 2018) on fixed data splits, and matched random-direction removal serves as a control. Training and intervention settings are given in Appendix C.

Figure 3 gives a complementary, retrieval-based view of these factors on two melanoma slides. The relapse model attends to localized regions of each slide, while the factor maps highlight regions scored highly for mitotic activity, nuclear pleomorphism, and vasculitis. A patch's score for a factor is the fraction of its nearest literature panels that are labelled present for that factor, so each high-scoring patch can be shown beside the factor-positive panels it retrieves. This is a visual check that tissue receiving a high factor score resembles the published morphology the factor was defined from.

Targeted erasure reduces mean polarity-probe AUC from 0.639 to 0.482, while matched random-direction removal leaves it at 0.637. Downstream no-op and random perturbations have negligible effects; full validation is given in Appendix C.2.

5.1 FACTOR ACCESSIBILITY DOES NOT IMPLY DOWNSTREAM SENSITIVITY

Factor accessibility is weakly related to downstream sensitivity in relapse prediction (Table 5). Tumour necrosis is among the more accessible factors (mean polarity AUC 0.686), yet removing its associated information changes relapse AUC by 0.16 percentage points. Lymphovascular invasion shows the opposite pattern: it has the weakest polarity accessibility among the eight factors (0.532), but its removal changes relapse AUC by 0.43 percentage points on average.

Across individual encoder–factor combinations, the pooled rank correlation between accessibility and relapse sensitivity is $\rho = 0.114$. Because factors differ in their overall accessibility, we also compute the correlation separately within each factor, comparing encoders only against other encoders evaluated on the same morphology. The median within-factor correlation is -0.003, with

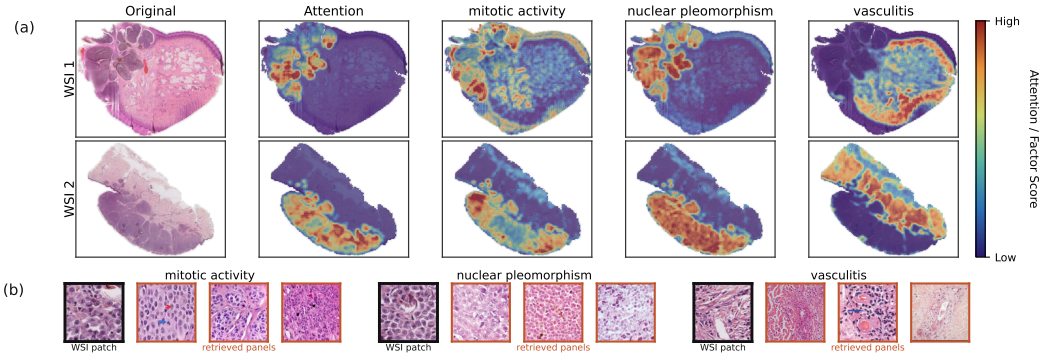

Figure 3: A relapse model's attention and retrieval-based factor scores on two melanoma slides.

six of eight factors between -0.132 and $+0.285$. A factor can be highly accessible on the literature benchmark yet have little downstream sensitivity, while comparatively weak accessibility can still correspond to a measurable downstream effect.

Table 5: Factor accessibility and relapse sensitivity. Polarity AUC is the mean over the 16 encoders $\pm$ its standard error; the downstream effect is the mean over encoders of each encoder's seed-averaged signed ΔAUC. *Above random* and *top factor* are defined in Appendix C.2.

Factor	Polarity AUC	Downstream effect (AUC pp)	Above random	Top factor
Mitotic activity	0.733 ± 0.013	0.85	15/16	2/16
Cytologic atypia	0.660 ± 0.009	0.57	15/16	5/16
Nuclear pleomorphism	0.627 ± 0.017	0.46	10/16	3/16
Lymphovascular invasion	0.532 ± 0.012	0.43	13/16	3/16
Invasive component	0.640 ± 0.026	0.26	10/16	1/16
Vasculitis	0.616 ± 0.010	0.22	9/16	0/16
Tumour necrosis	0.686 ± 0.013	0.16	6/16	2/16
Dysplasia	0.620 ± 0.020	0.09	5/16	0/16

5.2 Factor sensitivity differs across encoder representations

Encoders with nearly identical relapse performance can have different factor sensitivities. UNI2 and Virchow2 reach relapse AUC 0.798 and 0.797, yet erasing cytologic-atypia information costs 1.51 and 0.19 percentage points respectively. With the downstream task, architecture, cohort, data splits, and training procedure fixed, factor sensitivity varies across encoder–factor pairs. Across the full experiment of 16 encoders $\times$ 8 factors $\times$ 3 training seeds, the encoder–factor interaction explains 63.1% of the variation in sensitivity, compared with 16.8% for factor identity and 14.6% for the encoder main effect. Downstream-training seed explains 0.2% of the variance, indicating that these encoder-specific sensitivity patterns are reproducible across retraining.

These profiles are reproducible across independent downstream training runs: the eight-factor sensitivity profiles obtained from the same encoder correlate at Spearman $\rho = 0.861 \pm 0.013$, whereas profiles from different encoders correlate at only $\rho = 0.186 \pm 0.018$. Changing the foundation-model representation therefore alters the sensitivity profile substantially more than retraining the same downstream predictor.

This reproducibility also holds at the case level. For each factor, the relapse cases most affected by factor removal are similar across independent retraining runs, overlapping with Jaccard similarity 0.491 at the top decile, compared with 0.122 between different factors and a chance value of 0.053 (Figure 4(a)).

These results show that encoder choice changes the factor-sensitivity profile even under an otherwise fixed downstream pipeline.

5.3 The same representation gives different factor sensitivities across tasks

We next hold the encoder representation fixed and compare factor sensitivity across downstream tasks. Using the same cohort and the same encoder representations, we repeat the eight interventions for relapse, Breslow thickness and ulceration prediction. For each encoder and task, the eight intervention effects form a factor-sensitivity profile; both the encoder representation and the factor-removal directions are held fixed, and only the downstream predictor changes. For a fixed encoder, factor-sensitivity profiles are highly reproducible across independent runs of the same task (mean Spearman $\rho = 0.865$), but are much less similar across different tasks ($\rho = 0.459$). This pattern holds for all 16 encoders, each of which lies below the $y = x$ line in Figure 4(b). The downstream task therefore changes the factor-sensitivity profile, even when the representation and factor directions are fixed.

The extent of this task dependence varies across task pairs. Relapse and Breslow profiles are most similar on average ($\rho = 0.679$), followed by relapse–ulceration ($\rho = 0.414$) and Breslow–ulceration ($\rho = 0.284$); Appendix Figure 5 gives the encoder-level breakdown.

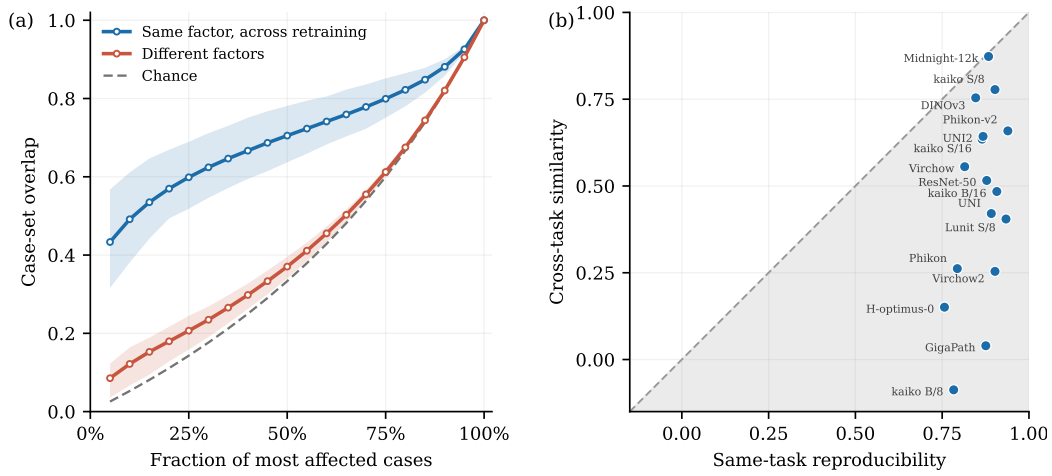

Figure 4: Factor effects reproduce across retraining but change with the task.

Section 5 shows that factor accessibility and downstream sensitivity capture distinct properties. A highly accessible factor is not necessarily one to which the downstream predictor is highly sensitive. Moreover, which factors matter varies with both the encoder and the prediction task: changing either one changes the sensitivity profile. Factor accessibility therefore describes what factor information is linearly recoverable from a representation, whereas factor sensitivity describes how sensitive a particular encoder–task model is to removing that information.

6 Discussion and Conclusion

Downstream accuracy does not show what factor information is accessible in a representation or how a downstream predictor responds to that information. We introduce a common morphology concept space that makes both measurable across heterogeneous pathology encoders.

Foundation models differ not only in overall morphology accessibility, but also in which specific factors are most accessible, so models with nearly identical aggregate scores can make different morphological information available. A factor-level benchmark makes these differences visible even when aggregate scores are similar.

Representation content and downstream sensitivity are also distinct. Factor accessibility is only weakly related to factor sensitivity: morphology that is easy to recover is not necessarily morphology a predictor is sensitive to. Sensitivity also changes when either the encoder or the downstream task changes, while remaining reproducible across independent retraining. Accessibility and sensitivity therefore provide complementary views of a representation and of its downstream behaviour.

The methodological contribution is a model-independent semantic framework for foundation-model interpretation. The morphology factors are defined externally from pathology reporting standards, before analysing any encoder, and are then mapped separately into each representation. This makes heterogeneous foundation models directly comparable in the same semantic space. Using the same factor set for downstream interventions further connects what each representation contains to how removing factor-associated information affects downstream performance.

The benchmark labels are grounded in caption assertions rather than independent image-level pathologist annotation, so the validation establishes caption–label agreement rather than visual validity. Factor removal tests sensitivity to linearly encoded information, not the causal effect of changing the underlying morphology; the factor direction is learned from literature panels, while the erasure operator is estimated from downstream training-fold patches. Finally, the downstream experiments use one melanoma cohort and three prediction targets, so generalization across diseases and institutions is not evaluated.

AI USE STATEMENT

Generative AI tools were used to assist with code generation, ontology extraction and factor-specific terminology drafting, assertion-polarity labelling as an explicit component of the method, brainstorming and feedback on experimental design and analysis, interpretation and presentation of results, and drafting and editing text. Ontology extraction from guideline PDFs was followed by page-by-page human verification. AI-assisted methodological suggestions and interpretations were reviewed by the authors against the underlying experiments, and AI-generated code used in the reported analyses was reviewed and validated through reproducible loading, alignment, and permutation tests.

ETHICS STATEMENT

The corpus derives from open-access literature. Article licences are resolved individually; panels under NoDerivatives or bespoke terms will be released as source coordinates rather than crops, and the same rule governs what the paper reproduces: every literature panel shown in Figure 1 is CC0, and every panel shown in Figure 3 is CC BY, reproduced as a crop with attribution to its source article — Tan et al. (2020), Xu et al. (2025) and Dickens et al. (2021) (mitotic activity); Nakamura et al. (2019), De Leo et al. (2024) and Shimazaki et al. (2023) (nuclear pleomorphism); Guo et al. (2017), Cho et al. (2022) and Fraticelli et al. (2021) (vasculitis). Whole-slide data are public and de-identified, released by the French Society of Pathology under CC BY 3.0 and Licence Ouverte 2.0 (Société Française de Pathologie et al., 2025). The dataset is a research instrument for representation analysis and is neither intended nor validated for clinical use.

REPRODUCIBILITY STATEMENT

We will release the relevant code and data needed to reproduce our benchmark and evaluation upon acceptance. That release will comprise the factor ontology, terminology and matching rules, article-disjoint splits, licence metadata, embeddings for all 16 encoders, model-loading verification records, and scripts for every table and figure. The split is article-disjoint; the probing experiments cross-validate over all panels with article-grouped folds rather than holding the test partition back, and the retrieval experiments query test panels against training figures.

REFERENCES

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

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, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Illeon Cho, Jwa-Kyung Kim, and Sung Gyun Kim. Iga vasculitis presenting as nephrotic syndrome following covid-19 vaccination: a case report. *BMC nephrology*, 23(1):403, 2022.

Edwin D de Jong, Eric Marcus, and Jonas Teuwen. Current pathology foundation models are unrobust to medical center differences. *arXiv preprint arXiv:2501.18055*, 2025.

Antonio De Leo, Martina Ruscelli, Thais Maloberti, Sara Coluccelli, Andrea Repaci, Dario De Biase, and Giovanni Tallini. Molecular pathology of endocrine gland tumors: genetic alterations and clinicopathologic relevance. *Virchows Archiv*, 484(2):289–319, 2024.

Tracey-Anne Dickens, Maria Franchina, Adam Gajdatsy, and Nima Mesbah Ardakani. Primary orbital melanoma arising in an atypical diffuse (plaque-like) blue naevus/melanocytosis: a case report and review of literature. *BMC ophthalmology*, 21(1):425, 2021.

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

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

Paolo Fraticelli, Devis Benfaremo, and Armando Gabrielli. Diagnosis and management of leukocytoclastic vasculitis. *Internal and emergency medicine*, 16(4):831–841, 2021.

Paul Gamble, Ronnachai Jaroensri, Hongwu Wang, Fraser Tan, Melissa Moran, Trissia Brown, Isabelle Flament-Auvigne, Emad A Rakha, Michael Toss, David J Dabbs, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communications medicine*, 1(1):14, 2021.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Yanhua Guo, Shouyang Ding, Hui Guo, Shenqiu Li, Xia Lu, Zhishui Chen, Zhonghua Klaus Chen, Changsheng Ming, and Nianqiao Gong. Graft-versus-host-disease after kidney transplantation: a case report and literature review. *Medicine*, 96(26):e7333, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

International Collaboration on Cancer Reporting. Published datasets, 2026. URL `https://www.iccr-cancer.org/datasets/published-datasets/`. Accessed 21 September 2026.

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

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3344–3354. IEEE, 2023.

Saarthak Kapse, Pushpak Pati, Srikar Yellapragada, Srijan Das, Rajarsi R Gupta, Joel Saltz, Dimitris Samaras, and Prateek Prasanna. GECKO: Gigapixel vision-concept contrastive pretraining in histopathology. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20020–20030. IEEE, 2025.

Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 573–583. Springer, 2025.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Chanwoo Kim, Jakub Kaczmarzyk, Deepika Savant, Zhen Zhao, Peter K Koo, and Su-In Lee. Dissecting and directing pathology foundation models. *bioRxiv*, 2026a.

Hyunjae Kim, Dain Kim, Pan Xiao, Serina S Applebaum, Younjoon Chung, Xuguang Ai, Yu Yin, Roy Jiang, Yuexi Du, Yawen Wei, et al. MedPMC: A systematic framework for scaling high-fidelity medical multimodal data for foundation models. *arXiv preprint arXiv:2607.07673*, 2026b.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Tristan Lazard, Kenza Bouzid, Julius Hense, Shruthi Bannur, Daniel Coelho de Castro, Daniel Shao, Rajesh Jena, Drew Williamson, and Stephanie Hyland. Sparse concept attribution for histomorphological hypothesis generation from whole-slide classifiers. *arXiv preprint arXiv:2609.02985*, 2026.

Vaibhav Mishra and William Lotter. Comparing computational pathology foundation models using representational similarity analysis. *Proceedings of machine learning research*, 297:1007–1022, 2026.

Yoshifumi Nakamura, Reiko Yamada, Maki Kaneko, Hiroaki Naota, Yu Fujimura, Masami Tabata, Kazuhiko Kobayashi, and Kyosuke Tanaka. Isolated pancreatic metastasis from malignant melanoma: a case report and literature review. *Clinical Journal of Gastroenterology*, 12(6):626–636, 2019.

Peter Neidlinger, Omar SM El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko Van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, et al. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature biomedical engineering*, 10(6):1113–1123, 2026.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Royal College of Pathologists. Cancer datasets and tissue pathways, 2026. URL https://www.rcpath.org/profession/guidelines/cancer-datasets-and-tissue-pathways.html. Accessed 21 September 2026.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. URL https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

Ryotaro Shimazaki, Masahiro Hagiwara, Chikayoshi Tani, Hiroyoshi Iwata, Hiroyuki Takahashi, Marika Fukuyama, Taisuke Matsuya, Koji Imai, Sayaka Yuzawa, Mishie Tanino, and Hideki Yokoo. Successful multidisciplinary treatment with laparoscopic hepatectomy and adjuvant therapy for metachronous solitary hepatic metastasis after excision of a primary anorectal malignant melanoma: A case report. *Current Oncology*, 31(1):203–210, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Société Française de Pathologie, Société Française de Dermatologie, Groupe de Cancérologie Cutanée, and Conseil National Professionnel des Pathologistes. Visiomel database: Predicting cutaneous melanoma relapse, 2025.

Laura Ling Ying Tan, Jiancheng Hong, Wei Lin Goh, Esther Wei Yin Chang, Valerie Shiwen Yang, Eileen Poon, Nagavalli Somasundaram, Mohamad Farid, Anita Sook Yee Chan, and Jason Yongsheng Chan. Clinical features and survival outcomes of ocular melanoma in a multi-ethnic asian cohort. *Scientific reports*, 10(1):16367, 2020.

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(10):2924–2935, 2024.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. *Advances in neural information processing systems*, 37:107984–108011, 2024a.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024b.

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):181–188, 2024.

Yitong Xu, Alejandro A Gru, Thomas Brenn, and Katharina Wiedemeyer. Brca1-associated-protein-1 inactivated melanocytic tumours: characterisation of the clinicopathological spectrum and immunohistochemical expression pattern of preferentially expressed antigen in melanoma. *Histopathology*, 86(2):294–301, 2025.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A BENCHMARK CONSTRUCTION AND VALIDATION

A.1 MORPHOLOGY FACTORS AND SEARCH TERMS

We extract reporting elements from current ICCR and RCPath cancer-reporting documents and normalize them to histologically assessable morphology factors: synonymous entries are merged, and administrative, non-histologic and non-visual elements are excluded. Removed entries include ISBNs, section headings, form prompts, and permissible values parsed as element names.

We exclude two terminology classes that produce systematically incorrect retrievals: anatomical layer names, which primarily retrieve normal structure, and entity names, which are often dominated by non-diagnostic experimental contexts. Exclusion decisions and their corresponding rules are retained as benchmark metadata.

Factor-specific search terms combine LLM-assisted term drafting, a manually curated caption-idiom lexicon, and corpus-derived terms. Retrieved captions are then manually reviewed to remove ambiguous or systematically incorrect matches. After normalization and review, 35 factors carry a reviewed term list and are taken forward for literature retrieval.

A.2 LITERATURE RETRIEVAL AND PANEL–CAPTION EXTRACTION

Candidate figures are retrieved from the PubMed Central Open Access corpus by matching the factor-specific lexicon against locally extracted figure captions, and are restricted to histopathology content before panel-level processing. Compound figures are split with the MedPMC multi-panel figure separation model (Kim et al., 2026b), a YOLOv10 detector (Wang et al., 2024a) fine-tuned for subfigure detection, and panel-specific captions are produced with the MedPMC caption-separation model, a fine-tune of InternVL2.5-4B-MPO (Chen et al., 2024b; Wang et al., 2024b). Each detected panel is paired with its panel-specific subcaption.

A.3 FACTOR MATCHING, ASSERTION LABELLING AND BENCHMARK FILTERING

Factor matching is lexical and polarity-blind: a caption is retrieved whenever it matches the factor lexicon, regardless of whether the finding is asserted or negated. We classify reporting-element–panel pairs as *present*, *absent*, *uncertain* or *not about the panel* using Qwen3.6-27B-FP8 (Qwen Team, 2026), with fixed few-shot examples and greedy decoding. For the 22 broad reporting elements carried forward from guideline extraction, we apply five-sample self-consistency to $16{,}040$ candidate pairs and retain only unanimous present or absent assignments.

We then remove pairs supported only by withdrawn search terms or lacking panel bounding boxes, and merge duplicates. The matched lexicon terms map each retained pair to one or more of the finer morphology factors of Table 1, producing a separate factor–panel example for each; a pair retrieved through a broad element can therefore yield more than one example. This mapping assigns labels rather than localizing regions within panels. Examples outside the released panel corpus are excluded.

Of the 35 candidate factors, we exclude histologic type, histologic grade and tumour extent of invasion as higher-order reporting tasks, and require at least 100 present examples of each remaining factor; nine fall below that threshold. The resulting benchmark contains $9{,}672$ factor–panel examples across 23 factors (Table 1). Eight factors additionally meet the explicit-polarity criterion of at least 30 explicitly absent examples; the eligible set is unchanged for thresholds from 20 to 31.

A.4 INDEPENDENT LABEL AUDIT

We independently audit the retained labels using Gemma-4-31B-it (Gemma Team, 2026), a different model family from the Qwen model used for dataset construction, given the pipeline's prompt and exemplars verbatim and shown only the factor, the matched terms and the subcaption. The sample of 1,000 factor–caption pairs, 800 present and 200 absent, was fixed by file hash before labelling. Caption–label agreement is 99.3% for present assertions and 97.0% for absent assertions, corresponding to 99.1% after reweighting to the benchmark composition. Of the twelve disagreements, ten arise from ambiguous captions and two from labelling errors. Table 6 summarises the validation.

Table 6: Label agreement and internal consistency checks.

Validation	N	Agreement
Independent audit, present assertions	800	99.3%
Independent audit, absent assertions	200	97.0%
Independent audit, reweighted to the dataset	1,000	99.1%
Same-model relabelling	300	98.7%
Greedy versus five-sample majority	16,040	97.0%

Table 6 also reports two internal consistency checks. *Same-model relabelling* re-labels 300 retained pairs with the same model family under greedy decoding and recovers the shipped label in 98.7% of cases; all four disagreements are the greedy pass assigning *not about the panel* where the shipped label is present or absent. *Greedy versus five-sample majority* compares the single-sample greedy label with the five-sample self-consistency label over all $16{,}040$ pairs that received both, agreeing on 97.0%.

The audit measures caption–label agreement; visual verification of the corresponding morphology is outside the scope of this audit.

B Additional factor-accessibility results

B.1 Experimental setup

For each encoder–factor pair, logistic regression ($C = 1$) is fit on frozen panel embeddings with five-fold cross-validation grouped by article. Features are standardized using training-fold statistics. Broad accessibility uses caption-asserted present panels against the remaining literature panels; explicit polarity uses explicitly present and absent panels with class balancing within each training fold. Each encoder uses its documented normalization, resolution and pooling rule.

B.2 Robustness to context matching

We progressively match comparison panels on disease, organ, stain and magnification. Table 2 reports per-encoder accessibility under each matching condition. Under the strictest condition, mean accessibility decreases by 0.055 AUC relative to the unrestricted setting.

B.3 Full explicit-polarity accessibility

Table 7 reports the full encoder-by-factor accessibility matrix for the eight factors with sufficient explicitly absent examples.

B.4 Accessibility decomposition control

Restricting the broad benchmark to the same eight factors leaves the decomposition nearly unchanged: 76.9% factor, 14.7% encoder and 8.4% interaction variation, against 76.9%, 14.5% and 8.6% over all 23 factors (Table 4). The larger interaction under explicit polarity is therefore associated with the evaluation setting rather than with the number of factors.

C Additional factor-sensitivity results

C.1 Experimental setup

The VisioMel cohort (Société Française de Pathologie et al., 2025) contains $1{,}741$ slides: 309 relapse-positive (17.7%), 371 ulceration-positive (21.3%), and 540 with Breslow thickness $\geq$ 2mm (31.0%). All three endpoints are treated as binary classification tasks using the same stratified five-fold partition. Relapse is the recorded relapse outcome, ulceration is the recorded presence of ulceration, and Breslow is defined as thickness $\geq$ 2mm.

Table 7: Explicit-polarity factor accessibility, every encoder and factor.

Encoder	Cyt.	Dysp.	Inv.	LVI	Mit.	Pleo.	Necr.	Vasc.	**Macro**
UNI2	.721	.712	.785	.485	.761	.720	.801	.633	.702
Virchow2	.737	.728	.651	.482	.820	.708	.736	.690	.694
Midnight-12k	.701	.749	.666	.537	.777	.670	.744	.660	.688
kaiko ViT-B/8	.676	.635	.727	.571	.740	.741	.625	.599	.664
kaiko ViT-S/16	.683	.672	.679	.607	.722	.667	.678	.579	.661
H-optimus-0	.670	.605	.690	.546	.733	.642	.643	.615	.643
Lunit ViT-S/8	.642	.535	.751	.548	.757	.627	.658	.628	.643
kaiko ViT-S/8	.661	.562	.636	.566	.756	.694	.647	.589	.639
DINOv3	.619	.519	.750	.570	.785	.553	.652	.662	.639
Prov-GigaPath	.652	.720	.567	.548	.733	.570	.710	.596	.637
kaiko ViT-B/16	.665	.658	.622	.437	.750	.591	.731	.587	.630
UNI	.629	.567	.687	.497	.706	.594	.679	.664	.628
Virchow	.647	.563	.626	.526	.688	.508	.713	.607	.610
Phikon	.596	.649	.485	.560	.717	.565	.705	.562	.605
ResNet-50	.614	.524	.449	.586	.606	.590	.634	.632	.579
Phikon-v2	.648	.524	.467	.441	.672	.600	.613	.559	.565
Mean	**.660**	**.620**	**.640**	**.532**	**.733**	**.627**	**.686**	**.616**	**.639**

Every task uses the same architecture and hyperparameters: gated-attention ABMIL with hidden width 256, attention dimension 128 and dropout 0.25, trained for 30 epochs with AdamW (10^{-4} learning rate, 10^{-5} weight decay) and weighted binary cross-entropy with $\mathrm{pos_weight} = n_-/n_+$. Models are evaluated out of fold by ROC AUC using the same stratified five-fold partition for every encoder and task, and each model is trained with three random seeds that vary only initialization and batch order.

For each encoder and polarity-eligible factor, we fit the factor direction w on literature panels: it is the present-versus-absent logistic probe direction of Section 4, held fixed thereafter. Writing $z = w^\top x$ for the factor score of a patch embedding x, we remove factor-associated linear information with LEACE (Belrose et al., 2023),

$$r(x) = x - W^+ P_{W\Sigma_{xz}} W(x - \mu), \qquad W = \Sigma_{xx}^{-1/2},$$

where $P_{W\Sigma_{xz}}$ is the orthogonal projection onto the column space of $W\Sigma_{xz}$ and W^+ is the pseudoinverse of W, estimating μ, Σ_{xx} and Σ_{xz} from VisioMel patch representations of that fold's training slides only, sampling 128 patches per slide, and applying the resulting transformation to the held-out patches of the fold. Because z is linear in x we have $\Sigma_{xz} = \Sigma_{xx} w$, so the eraser is rank one and, at full strength, satisfies $\mathrm{Cov}(r(x), z) = 0$: the linear factor score z is removed from the transformed representation. This guarantee concerns the prespecified score z; how much linearly recoverable information about the factor *label* survives is an empirical question, measured by re-fitting a probe after erasure in Appendix C.2.

The encoder and every downstream model parameter remain fixed during intervention, and no patch is removed from the bag. Attention is recomputed on the erased representations, so the reported effect includes the resulting change in attention weights.

We report factor sensitivity as the signed change $\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{original}} - \mathrm{AUC}_{\mathrm{erased}}$, so a positive value means performance decreases once the factor is removed and a negative value means it increases. Every sensitivity reported in this paper is signed, including the *above random* and *top factor* columns of Table 5; absolute values enter only in the random-direction band of Appendix C.2.

C.2 Intervention validation

Each intervention is matched by a random-direction control. For every fold we draw five directions i.i.d. from a standard normal in the whitened coordinate system of the factor directions, normalize them to unit length, and erase each with the same rank-one operator at the same strength on the same slides. Because the projection is scale invariant, the control needs a random direction rather than a matched norm. For an encoder and seed, the random band is the largest absolute downstream AUC change over those five directions.

Table 5 uses this band in its last two columns. *Above random* counts the encoders whose signed factor-removal effect exceeds that encoder's random band in all three training seeds; *top factor* counts the encoders for which that factor has the largest seed-averaged signed sensitivity among the eight.

Targeted factor removal reduces linearly recoverable factor information to approximately chance, while matched random removal leaves it nearly unchanged. Across the 128 encoder–factor combinations, 119 targeted interventions reduce probe AUC to at most 0.55, compared with 18 matched random controls. Matched random removal changes downstream AUC by at most 0.0021. Probe performance is re-estimated after erasure to measure residual linearly recoverable factor information.

C.3 Full factor-sensitivity results

Table 8 reports seed-averaged changes in relapse AUC, in percentage points, for every encoder–factor pair.

Table 8: Factor sensitivity in AUC percentage points, every encoder and factor.

Encoder	**AUC**	Cyt.	Dysp.	Inv.	LVI	Mit.	Pleo.	Necr.	Vasc.
Midnight-12k	0.804	0.44	0.01	0.05	1.15	0.46	0.46	0.30	0.32
UNI2	0.798	1.51	0.12	0.13	0.22	0.11	0.15	-0.09	-0.05
Virchow	0.797	0.29	-0.06	-0.07	0.24	0.06	0.10	-0.16	0.10
Virchow2	0.797	0.19	0.07	-0.24	0.02	0.15	0.71	-0.17	0.34
Lunit ViT-S/8	0.796	0.46	0.91	0.36	1.77	1.19	0.81	0.24	1.08
H-optimus-0	0.795	0.07	0.03	0.57	0.21	0.09	-0.03	-0.01	-0.01
Phikon-v2	0.795	0.20	0.04	0.15	0.08	2.87	0.40	0.06	0.28
DINOv3	0.791	0.40	0.28	0.28	0.16	0.42	0.05	0.72	0.19
Prov-GigaPath	0.788	0.04	0.17	0.58	0.74	1.86	-0.00	0.14	0.16
kaiko ViT-B/8	0.788	0.39	-0.00	-0.11	0.81	0.58	-0.04	-0.05	0.17
kaiko ViT-S/8	0.787	0.41	-0.06	0.38	0.23	2.08	2.27	-0.19	0.01
kaiko ViT-B/16	0.786	0.80	0.02	0.34	0.10	1.31	2.41	0.20	0.26
ResNet-50	0.785	0.09	0.10	-0.02	0.56	0.58	0.03	0.61	0.33
kaiko ViT-S/16	0.781	1.71	0.26	-0.09	0.31	0.95	0.32	1.04	0.16
Phikon	0.781	0.92	-0.31	0.86	-0.02	0.39	-0.16	-0.07	0.11
UNI	0.777	1.16	-0.19	0.97	0.30	0.43	-0.06	-0.03	0.01
Mean	**0.790**	**0.57**	**0.09**	**0.26**	**0.43**	**0.85**	**0.46**	**0.16**	**0.22**

C.4 Sensitivity-profile reproducibility

Across training seeds, sensitivity profiles from the same encoder correlate at Spearman $\rho = 0.861 \pm 0.013$, compared with $\rho = 0.186 \pm 0.018$ across encoders at fixed seed. Encoder–factor interaction accounts for 63.1% of sensitivity variance, while downstream-training seed accounts for 0.2% (Table 9).

The case-level analysis uses a per-slide effect rather than a cohort-level one. For a given encoder, factor and seed, the effect on slide i is the standardized absolute change in its relapse logit, $|\ell_i^{\text{base}} - \ell_i^{\text{erased}}|/\text{sd}(\ell^{\text{base}})$, with the standard deviation taken over all $1{,}741$ slides of that run. The most affected cases are the top decile under this quantity ($k = 174$ slides). We report the Jaccard similarity between two such sets: within a factor across seed pairs, and between factor pairs at fixed seed. For two independent sets of size k drawn from n slides, $(k^2/n)/(2k - k^2/n) = 0.053$ is an approximation to the expected Jaccard, taking the ratio of the expected intersection to the expected union; it is the chance value quoted in Section 5.2.

Table 9: Reproducibility and variance decomposition of factor sensitivity.

Quantity	**Value**
Same encoder, different seeds	0.861 ± 0.013
Different encoders, same seed	0.186 ± 0.018
Encoder $\times$ factor	63.1%
Factor	16.8%
Encoder	14.6%
Downstream-training seed	0.2%
Residual	5.3%

C.5 Cross-task factor-profile similarity

Figure 5 gives the encoder-level breakdown of the cross-task comparisons summarized in Section 5.3. Relapse and Breslow profiles are the most similar pair on average, and also the most consistent: they are the closest of the three pairs for 9 of the 16 encoders, and their per-encoder similarities span 0.024 to 0.929 (SD 0.219). Breslow and ulceration are both the least similar on average and the most variable, spanning -0.571 to 0.857 (SD 0.463), so the degree of task dependence is itself encoder-specific.

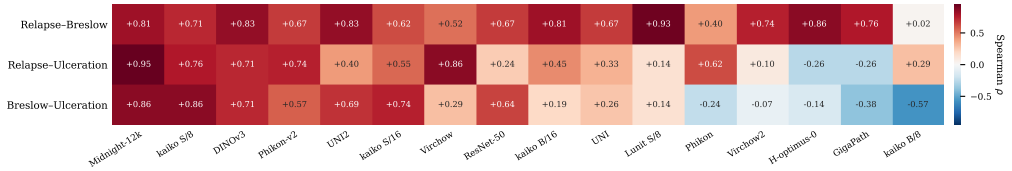

Figure 5: Cross-task factor-profile similarity for each encoder and task pair.

D Additional qualitative visualization

The paper uses two distinct per-patch factor scores. Figure 3 shows **retrieval-based factor scores**: the fraction of a patch's nearest literature panels labelled present for the factor. Figure 6 instead shows **factor-direction projections**: the projection of each patch embedding onto that encoder's corresponding factor direction, the same direction the intervention erases. The two are not the same quantity and are not directly comparable.

Figure 6 covers all 16 encoders and all eight polarity factors on the same slide. The same tissue region is shown for every encoder. Heatmaps are normalized independently; colour intensity should therefore be interpreted within, rather than across, cells.

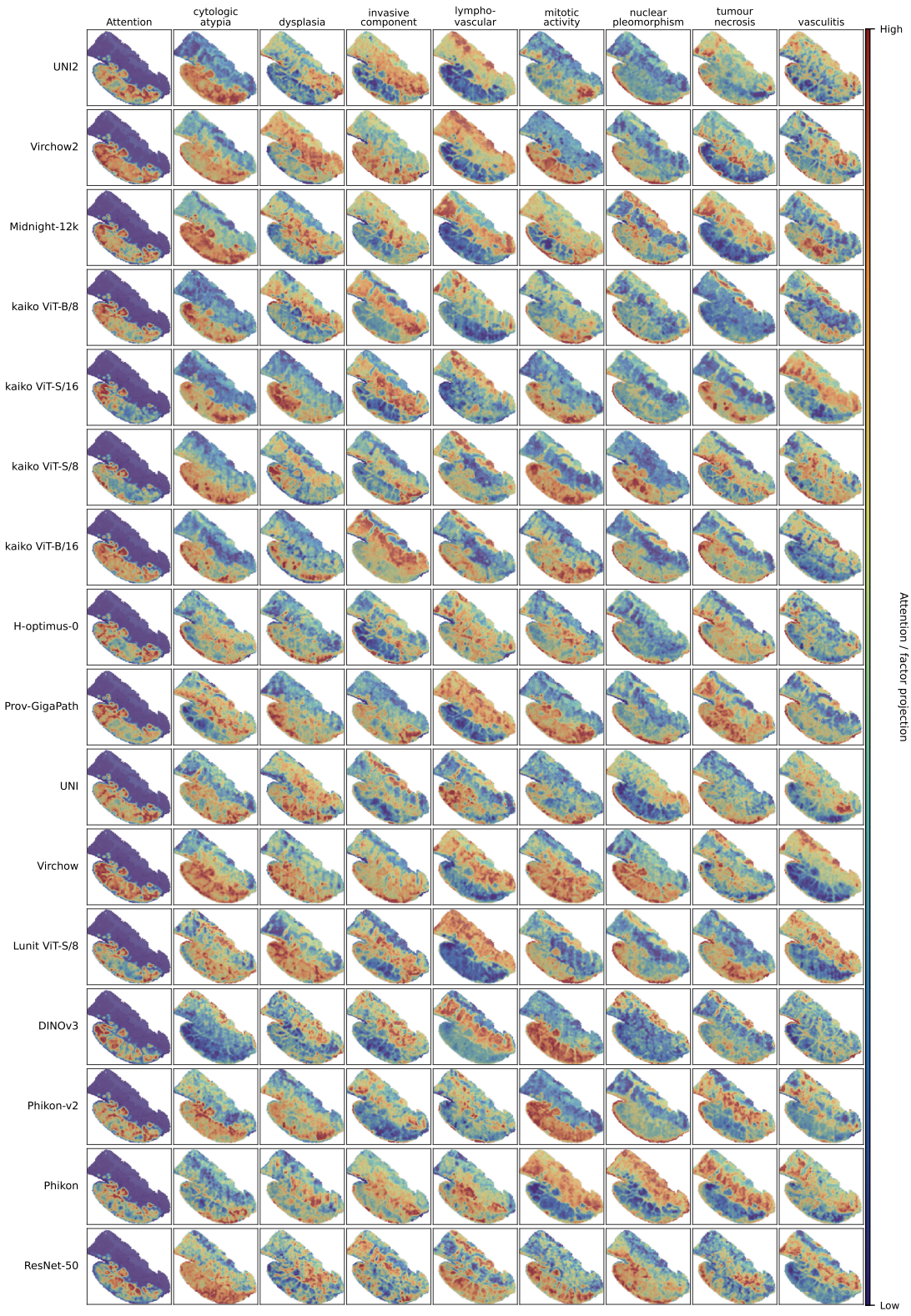

Figure 6: Attention and eight polarity-factor projections, all 16 encoders.