PATHRET: BENCHMARKING WHOLE-SLIDE IMAGE RETRIEVAL BEYOND LABEL AGREEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide image (WSI) retrieval returns prior cases that pathologists can directly compare, but conventional label-based evaluation measures diagnostic agreement with the query rather than which slides are actually returned. Replacing a retrieved slide with another assigned the same diagnostic relevance at the same rank leaves the evaluation unchanged. Such metrics can therefore measure whether the right kind of case was retrieved, but cannot distinguish which instances were selected within that diagnostic group. We compare seven retrieval systems from different methodological families and analyze queries for which a pair of systems each returns five slides with the query's exact subtype. In these cases, label-based evaluation regards both systems as fully successful. Yet among 474 such queries for PRISM and Yottixel-K, 72.8% have no slide in common between the two returned top-five sets. Thus, retrievals that appear equally successful under diagnostic labels can present substantially different cases to the user. To address this limitation, we introduce PathRet, a benchmark of $4{,}249$ WSIs from 13 cancer types that evaluates both label-defined clinical relevance and visual relatedness of the retrieved slides under a fixed retrieval protocol. Visual relatedness is measured by embedding WSI patches with a fixed pathology foundation encoder and comparing the query and retrieved slides at the slide level; an auxiliary measure compares their visual-phenotype distributions. When applied to real retrieval systems, the two views can lead to different system comparisons. PRISM scores higher than Prov-GigaPath on subtype and stage relevance, whereas Prov-GigaPath scores higher on both slide-level representation similarity and phenotype-composition similarity. This shows that a system retrieving more diagnostically relevant cases does not necessarily select cases that are more visually related to the query. WSI retrieval should therefore distinguish retrieving diagnostically appropriate cases from selecting visually related cases within those diagnostic groups, rather than evaluating retrieval through a single view alone.

1 INTRODUCTION

When pathologists encounter a challenging or unfamiliar case, a diagnostic label alone may not be enough. They may need to open prior slides and compare them directly with the case at hand. Whole-slide image (WSI) retrieval provides such references from large archives for education, rare-case review, and diagnostic support. The output of retrieval is therefore not merely a diagnostic label, but a set of concrete slides that a pathologist can inspect and compare. Label-based evaluation is not inherently inappropriate. When retrieval is used as a probe of pathology foundation-model representations, the goal is often to ask whether slides with the same diagnosis are placed near one another in the representation space (Zhang et al., 2025; Alfasly et al., 2025). For that purpose, label-based metrics are well suited. Metrics such as Hit@K, majority vote, and mean average precision have been widely used in this setting (Kalra et al., 2020; Chen et al., 2022; Wang et al., 2023; Tizhoosh & Pantanowitz, 2024), and PathRet retains this perspective as one axis of evaluation.

The question changes when retrieval itself is the end task. What matters then is not only whether cases with the correct label were retrieved, but also which cases were selected among those sharing that label. Label-based evaluation can answer the first question, but it cannot distinguish the second. Label-based metrics are computed from the relevance values of the returned slides, not from the slides themselves, so any metric built this way is unchanged when one returned slide is replaced by

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.

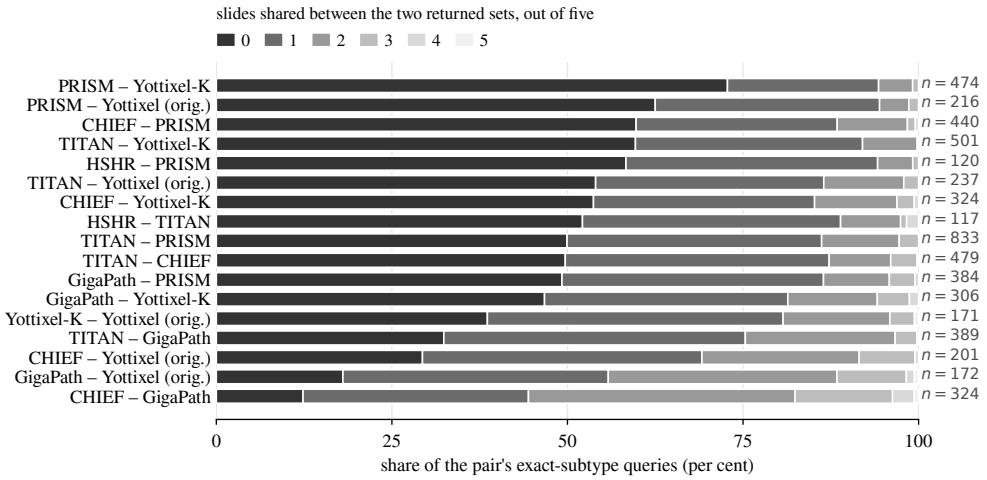


Figure 1: Each row is a system pair on the queries where both return five slides of the query's exact subtype, so both relevance sequences are $[1, 1, 1, 1, 1]$ and every label-defined metric here is identical. The bar is how many of the five they share. Rows are the 17 of 21 pairs with at least 100 such queries. The quantity is a set intersection over slide identifiers and uses no visual measure.

another of the same relevance at the same rank. We refer to these two questions as *class recovery* and *instance selection*, respectively.

We examine how this distinction appears in real WSI retrieval outputs. We compare seven systems from different methodological families and analyze queries for which two systems each return five slides with the query's exact subtype. Under label-based evaluation, both retrievals are fully successful. Yet among 474 such queries for PRISM and Yottixel-K, 72.8% have no slide in common between the two returned top-five sets (Figure 1). Thus, two retrievals that appear equally successful under diagnostic labels can present entirely different cases to the user. Figure 1 summarizes this phenomenon across the 17 system pairs with at least 100 such queries.

To address this limitation, we introduce PathRet. Layer 1 evaluates label-defined clinical relevance, while Layer 2 evaluates visual relatedness by embedding WSI patches with a fixed pathology encoder and comparing slide-level representations; an auxiliary measure compares phenotype distributions. In other words, Layer 1 asks whether diagnostically appropriate cases were retrieved, whereas Layer 2 asks how visually related the selected cases are to the query.

Applying both views to real retrieval systems leads to different system comparisons. PRISM scores higher than Prov-GigaPath on the subtype and stage axes, whereas Prov-GigaPath scores higher on the slide-level visual measure. This shows that retrieving more diagnostically appropriate cases and selecting more visually related cases within them can reflect different retrieval properties. WSI retrieval should therefore distinguish these two aspects rather than evaluate retrieval through a single view alone.

Our contributions are threefold.

1. **We formalize the distinction between class recovery and instance selection for WSI retrieval** and characterize the invariance of label-defined metrics to substitutions among equally relevant candidates. This makes explicit what label-defined evaluation cannot observe.
2. **We show this blind spot at scale in real retrieval systems:** among 474 queries where PRISM and Yottixel-K each return five exact-subtype matches, 72.8% of their top-five sets share no slide.
3. **We introduce PathRet**, a benchmark of $4{,}249$ WSIs across 13 cancer types that separately evaluates label-defined clinical relevance and representation-based visual relatedness under a fixed retrieval protocol, with all methods compared on the same cohort, candidate pool, and scoring rules.

2 Related Work

Whole-slide retrieval and evaluation. WSI retrieval has evolved from patch-selection and indexing pipelines such as Yottixel (Kalra et al., 2020), SISH (Chen et al., 2022), RetCCL (Wang et al., 2023) and HSHR (Li et al., 2023) to recent pathology foundation-model and vision-language approaches (Alfasly et al., 2025; Wang et al., 2025). Despite these methodological differences, evaluation has largely defined success through diagnostic labels, using metrics such as accuracy, majority vote, and mean average precision (Kalra et al., 2020; Chen et al., 2022; Wang et al., 2023). Recent studies have expanded the scale and reproducibility of this evaluation. Lei et al. (2026) compare ten pipelines on 9,387 TCGA slides with patient-level leave-one-patient-out retrieval, Alfasly et al. (2025) evaluate foundation models across 23 organs and 117 subtypes, and Patho-Bench standardises public tasks, splits, and evaluation code (Zhang et al., 2025). Across these settings, relevance remains primarily defined by diagnostic labels.

Reader-based assessment. Some studies instead assess the returned cases directly. HERE101 asks three pathologists to score the top-five results for 101 diagnostic queries (Zhong et al., 2025), while Shang et al. (2024) evaluate retrieval systems with pathologists and report cases where strong conventional scores coexist with clinically inadequate results. Such assessment directly evaluates the cases presented to users, but is difficult to reproduce at benchmark scale because collections may be private and reader panels must be reconvened.

Benchmark and metric design. Broader benchmark literature has shown that conclusions can depend strongly on evaluation protocol and metric choice. Prior work has revisited progress under matched conditions (Musgrave et al., 2020), re-annotated saturated retrieval benchmarks (Radenović et al., 2018), and argued that metrics should follow the problem rather than convention (Maier-Hein et al., 2024). Fixed learned representations have also been used as common comparison spaces in generative and perceptual evaluation (Heusel et al., 2017; Zhang et al., 2018).

Information retrieval has also developed evaluation beyond plain relevance, including novelty and diversity, intent-aware diversification, and nugget-based evaluation (Clarke et al., 2008; Agrawal et al., 2009; Lin & Demner-Fushman, 2006). These approaches show that retrieval evaluation need not collapse to binary or graded relevance alone. However, they do not address the setting PathRet targets, where diagnostic relevance can remain fixed while the concrete slides selected within the same relevance class still differ.

In WSI retrieval, automated evaluation has therefore remained largely label-based, whereas direct assessment of returned cases has required pathologist readers. PathRet retains label-based evaluation and adds fixed-representation visual relatedness as a separate automated evaluation axis.

3 The Structural Blind Spot of Label-Defined Evaluation

3.1 Class recovery and instance selection

A retrieval system can be evaluated through two distinct questions. The first is whether it retrieves cases from the correct diagnostic class. The second is which cases it selects among those belonging to that class. We refer to these as *class recovery* and *instance selection*, respectively.

The distinction depends on what retrieval is used for. When retrieval serves as a probe of representation quality, class recovery is often the target of interest. If the question is whether slides with the same diagnosis are placed near one another, label-defined evaluation is well suited to that purpose. When retrieval itself is the end task, however, the output is a set of concrete slides presented to the user. Evaluation must then distinguish not only whether the correct diagnostic class was recovered, but also which cases were selected within that class. Label-based evaluation measures the former but cannot distinguish the latter.

3.2 Formal setting and substitution invariance

Let $C(q)$ denote the candidate set for a query slide q, and let a retriever return the ranked list

$$R_k(q) = (c_1, \ldots, c_k), \qquad c_i \in C(q).$$

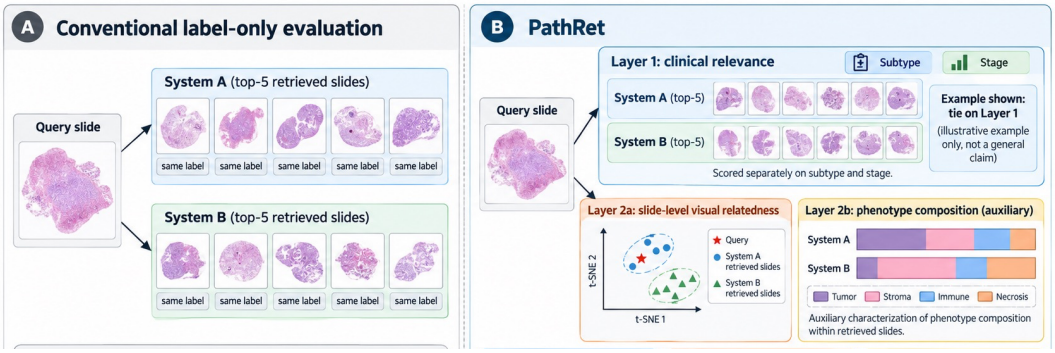


Figure 2: Conventional label-only retrieval evaluation (A) and the two targets PathRet reports (B).

Let $\ell(c)$ denote the label of candidate c on the axis being evaluated, and let $r(q, c)$ denote the relevance assigned by the evaluation to candidate c for query q. A metric is *relevance-determined* if, for a fixed query, it depends on the returned slides only through their relevance sequence,

$$M(q, R_k) \;=\; F_q\big(r(q, c_1), \ldots, r(q, c_k)\big),$$

where F_q may also depend on quantities fixed by the query, such as the ideal gain. A metric is *label-determined* if it depends on the returned slides only through their label sequence,

$$M(q, R_k) \;=\; G_q\big(\ell(c_1), \ldots, \ell(c_k)\big).$$

Hit@K, AP@K, MAP@R, and normalised discounted cumulative gain with label-derived relevance are relevance-determined. Majority vote at K is not relevance-determined because its value can depend on how non-relevant candidates are distributed across labels, but it is label-determined.

Proposition 1 (Substitution invariance)**.** *Let M be relevance-determined. Replacing a returned candidate c_i with another candidate c_i' satisfying $r(q, c_i') = r(q, c_i)$ leaves M unchanged. Likewise, if M is label-determined, replacing c_i with c_i' satisfying $\ell(c_i') = \ell(c_i)$ leaves M unchanged.*

We state this consequence explicitly because it fixes the information available to label-defined evaluation and makes the empirical analyses below possible. The proposition identifies exactly what these metrics cannot observe. If two retrievals produce the same relevance sequence, a relevance-determined metric receives the same input from both, regardless of which slides realise those relevance values. The two retrievals may therefore contain entirely different slides while receiving exactly the same score. The same argument applies to label-determined metrics when the label sequence is fixed.

3.3 Finer grading does not remove the blind spot

Finer relevance grading can reduce this blind spot, but it cannot remove it. Introducing multiple relevance levels allows an evaluation to distinguish different degrees of diagnostic match and reward their ordering. Once the relevance value at each rank is fixed, however, the metric still cannot distinguish which candidate realises that value.

Grading therefore makes the indistinguishable candidate sets smaller rather than eliminating them. Even with finer diagnostic distinctions, candidates assigned the same relevance remain interchangeable from the metric's perspective. Finer grading narrows the blind spot, but does not remove it. This motivates evaluating label-defined relevance and instance-level visual relatedness as separate retrieval targets.

4 The PathRet Benchmark

PathRet evaluates WSI retrieval along two separate targets under a single fixed retrieval protocol. Layer 1 measures label-defined clinical relevance, and layer 2 measures visual relatedness in a fixed pathology representation. The two targets answer different questions and are therefore reported separately rather than combined into a composite score. The cohort, candidate pool, relevance definitions, visual evaluator and statistical procedure are fixed across methods.

4.1 COHORT AND RETRIEVAL PROTOCOL

PathRet contains 4,249 diagnostic WSIs from 3,980 patients across 13 Genomic Data Commons projects. We include slides whose case records contain cancer type, subtype and stage and whose magnification can be established from the file. Clinical labels are taken from the diagnosis entry marked as the primary tumour. Detailed cohort construction and exclusions are given in Appendices A.1 and A.2.

Every slide can serve as a query, and its candidate pool contains the other slides in the cohort. The query itself and all slides from the same patient are excluded. In the primary setting, slides sharing the query's normalized GDC Source Site label are also excluded to prevent direct same-site retrieval; Section 7.1 audits the remaining Source-Site association. Cancer type is not used as a candidate filter. Slides from other cancer types remain in the pool and receive zero relevance in layer 1.

Rare classes are retained in the candidate pool. Instead of imposing a class-size threshold, eligibility is determined separately for each query and scoring axis from the query's eligible pool. This avoids simplifying the retrieval task by removing rare-class candidates globally. Query and candidate definitions are released with the benchmark, and we designate $k = 5$ as the primary cut-off and also score every method at $k \in \{1, 3, 5, 10, 20, 50, 100\}$; the remaining cut-offs are released with the per-method result artifacts.

Uncertainty in retrieval scores and paired differences between methods is estimated with a patient-cluster bootstrap. Patients are the resampling unit, so all slides from a resampled patient remain together, and method comparisons are paired within the same bootstrap draw. Point estimates are arithmetic means over eligible queries.

4.2 CLINICAL RELEVANCE

Retrieval returns prior cases, and a prior case is clinically relevant along more than one dimension. Layer 1 reads two of them from the case record. Subtype is the diagnostic identity of the tumour within its cancer type, and stage is the recorded extent of disease at diagnosis. Both are metadata-defined properties of the case a slide belongs to, they describe different aspects of clinical similarity between a query case and a returned one, and they are therefore graded and scored as two separate axes. Cancer type acts as a gate over both. A method is run once and the resulting ranked list is scored separately on the two axes.

For subtype relevance, a candidate from a different cancer type receives 0, a candidate from the same cancer type with a different subtype receives 0.5, and an exact subtype match receives 1. For stage relevance, a candidate from a different cancer type receives 0. Within the same cancer type, a different major stage receives 0, the same major stage with a different substage receives 0.5, and an exact stage match receives 1.

Each axis a is scored with normalised discounted cumulative gain at cut-off k,

$$\mathrm{NDCG}_a@k \;=\; \frac{\sum_{i=1}^{k} r_a(q, c_i)/\log_2(i+1)}{\mathrm{IDCG}_a@k(q)}.$$

Here, $r_a(q, c_i)$ is the relevance of the candidate at rank i, and $\mathrm{IDCG}_a@k(q)$ is the maximum discounted gain achievable from the query's full eligible candidate pool. This normalisation accounts for both rank and the amount of relevant material available to each query, allowing rare and common classes to be evaluated relative to what is achievable for that query.

A query contributes to an axis only when its ideal gain on that axis is positive. A slide that is the only member of its subtype is therefore excluded as a subtype query, while remaining available as a candidate and, when eligible, as a query on the stage axis.

4.3 VISUAL RELATEDNESS

Layer 2 measures visual relatedness in a fixed pathology representation shared across retrieval methods. Layer 2a is the primary measure, a slide-level representation similarity. Layer 2b is an auxiliary characterization of the same retrieval in terms of phenotype composition. It is read from the same

Table 1: The PathRet evaluation contract. Each row is a choice that prior evaluations make differently or leave unstated, and a comparison that departs from any of them is not comparable with the entries reported here. The relevance grades are ordinal, so a partially correct retrieval is scored as such, and we refer to them as *graded relevance*.

Component	PathRet specification
Retrieval pool	released per-axis query and candidate definitions; the query and the slides of its patient excluded; the same normalized GDC Source Site excluded in the primary setting
Cancer gate	a candidate of a different cancer type receives zero relevance; cancer type is a gate and not a candidate filter
Layer 1	subtype, the diagnostic identity within a cancer type, and stage, the recorded extent of disease, as two separate graded relevance axes defined in Section 4.2
Eligibility	an axis is scored only where the query's ideal gain on it is positive
Layer 2a	CONCH v1.5, up to 256 patches per slide, mean pooled, compared by cosine
Layer 2b	all-patch phenotype composition over $K = 128$, compared by $1 - \sqrt{\mathrm{JS}}$
Reporting	$k = 5$ primary with the remaining cut-offs released; patient-cluster bootstrap, differences paired inside the draw

patch representation as layer 2a and is reported as a description of what the returned slides contain, not as an independent confirmation of layer 2a.

A fixed scoring representation is necessary because similarities measured inside different retrievers' own spaces are not directly comparable. Scoring every method in one representation places them on a common measurement scale, whose behaviour is audited in Section 7.

Layer 2a, the primary measure. For each slide, up to 256 patches are encoded with the fixed pathology encoder, drawn without replacement under a seed derived from the slide identifier, and a slide with fewer eligible patches contributes all of them. Patch features are mean pooled into a single slide-level vector, and query-candidate similarity is computed with cosine similarity. No learned slide-level aggregation is used.

Layer 2b, an auxiliary measure. Layer 2b assigns patch-level representations to one of $K = 128$ fixed visual phenotypes and represents each slide by its phenotype composition p. The composition is formed over all eligible released patches of the slide rather than over the layer 2a budget, and patch features are centred by the slide's all-patch mean before assignment. The dictionary, PCA transform and centroids are fixed by the benchmark and used unchanged across methods. Query and candidate phenotype compositions are compared by $S_{\mathrm{phen}}(q, c) = 1 - \sqrt{\mathrm{JS}(p_q, p_c)}$, where JS is the Jensen-Shannon divergence. Larger values indicate more similar phenotype compositions.

The protocol. Table 1 summarises the full PathRet specification. Direct comparison with PathRet results requires the same cohort, candidate definitions, relevance rules, evaluator, preprocessing and statistical protocol. Encoder pretraining provenance and additional evaluator checks are documented separately in the appendices.

5 Real Retrieval Systems under PathRet

5.1 Seven systems on common support

Table 2 reports seven WSI retrieval systems spanning barcode retrieval, hypergraph hashing, foundation-model embeddings, and slide-level foundation models, all evaluated under the same PathRet protocol. All systems are scored on a common universe of 4,240 slides, excluding nine slides without the mosaic support required by the Yottixel variants. For the four methods also evaluable on all 4,249 slides, the largest change across the four evaluation axes is 0.001, so we use the 4,240-slide support throughout. This yields 4,240 subtype queries, 4,225 stage queries, and 4,240 queries for both layer 2 measures.

Table 2: Seven WSI retrieval systems evaluated under PathRet on the common $4{,}240$-slide support at the primary cut-off $k = 5$. Values are means with 95% patient-cluster bootstrap intervals. TCGA exposure indicates pipeline-level pretraining exposure, where *Unresolved* means that exposure of the exact checkpoint cannot be established from available provenance. Encoder-level provenance is given in Appendix A.6.

Method	TCGA exposure	NDCG subtype	NDCG stage	Layer 2a	Layer 2b
TITAN	Unresolved	0.7589 [0.7504,0.7670]	0.3355 [0.3266,0.3449]	0.9044 [0.9028,0.9061]	0.4811 [0.4787,0.4837]
PRISM	Unresolved	0.7220 [0.7125,0.7310]	0.3073 [0.2984,0.3162]	0.8866 [0.8849,0.8885]	0.4522 [0.4497,0.4548]
Yottixel-K	Yes	0.6520 [0.6418,0.6624]	0.2726 [0.2642,0.2809]	0.8626 [0.8600,0.8651]	0.3885 [0.3853,0.3915]
CHIEF	Yes	0.5646 [0.5538,0.5756]	0.2406 [0.2324,0.2490]	0.8664 [0.8642,0.8687]	0.3987 [0.3960,0.4015]
Prov-GigaPath	No	0.5479 [0.5375,0.5583]	0.2417 [0.2335,0.2497]	0.9004 [0.8984,0.9023]	0.4652 [0.4626,0.4679]
Yottixel (orig.)	No	0.4144 [0.4031,0.4254]	0.1828 [0.1757,0.1895]	0.8343 [0.8310,0.8375]	0.3703 [0.3671,0.3734]
HSHR	Not reported	0.3799 [0.3699,0.3895]	0.1642 [0.1575,0.1708]	0.8072 [0.8041,0.8103]	0.3379 [0.3352,0.3406]

The systems use two ranking regimes. Yottixel (Kalra et al., 2020), Yottixel-K, HSHR (Li et al., 2023) and TITAN (Ding et al., 2025) follow their published retrieval rules. PRISM (Shaikovski et al., 2024), CHIEF (Wang et al., 2024) and Prov-GigaPath (Xu et al., 2024) use the same adaptation: each model's released slide representation is candidate-pool centred, L2-normalised, and ranked by Euclidean distance. Appendix A.4 gives the representations and implementation details. No method is adapted using PathRet labels for this evaluation. Table 2 supports two analyses below: whether label-equivalent retrievals return the same cases, and whether diagnostic relevance and visual relatedness yield different system comparisons.

5.2 The same label success, different retrieved cases

We test whether retrievals judged equally successful by label-defined evaluation return the same cases by comparing returned slide identifiers directly, without using layer 2.

Cancer type. We first condition on queries where both systems attain maximum AP@5 under conventional cancer-type relevance, so both return the relevance sequence $[1, 1, 1, 1, 1]$ and receive the same maximal label-defined score. Across the 21 system pairs, their top-five sets share no slide on 28.4% to 77.6% of such queries, and no pair returns the same ranked top-five list on any query.

Exact subtype. We repeat the analysis under the finer exact-subtype condition, selecting queries where both systems return five slides carrying the query's exact subtype, so both carry the subtype relevance sequence $[1, 1, 1, 1, 1]$. Seventeen of the 21 pairs have at least 100 such queries, and across those 17 the proportion on which the two top-five sets share no slide ranges from 12.3% to 72.8% (Fig. 1). Exact ranked-list agreement is zero for all 21 pairs; exact unordered-set agreement occurs on only four queries across three pairs.

The largest disagreement occurs between PRISM and Yottixel-K. Among the 474 queries for which both systems return five exact-subtype matches, 72.8% $[69.0, 76.7]$ have no returned slide in common, where two independent uniform top-five draws from each query's eligible exact-subtype pool would be disjoint on 90.3%. The systems therefore agree more than independent selection, yet remain disjoint on most qualifying queries, and that identity variation is what saturated label metrics cannot express. Appendix B.1 gives the full baseline analysis.

These results show that the blind spot in Section 3 occurs frequently in real retrieval outputs: systems can achieve the same exact-subtype label success while returning substantially different cases.

5.3 The two targets can yield different system comparisons

A comparison between two systems can come out one way on diagnostic relevance and the other way on visual relatedness. PRISM against Prov-GigaPath is the case we report. At $k = 5$, PRISM exceeds Prov-GigaPath on subtype relevance by $+0.1741$ $[+0.1644, +0.1837]$ and on stage relevance by $+0.0655$ $[+0.0571, +0.0741]$, while Prov-GigaPath exceeds PRISM on layer 2a by $+0.0137$ $[+0.0122, +0.0153]$. All three intervals are 95% patient-cluster bootstrap intervals and none of them covers zero. The auxiliary phenotype-composition measure points in the same direction at this cut-off, by $+0.0130$ $[+0.0102, +0.0157]$, and is treated as a characterization of the returned slides rather than independent confirmation of layer 2a.

Which system leads therefore depends on which retrieval property is being measured, and a single combined score would report one of the two answers and suppress the other.

6 A Controlled Stress Test

Section 5.2 shows that real retrieval systems can return substantially different cases even when label-defined evaluation regards them as equally successful. We now ask the question directly: how visually different can the retrieved slides still be when their diagnostic relevance is identical?

For each query we first keep only the candidates with the highest possible label relevance, and require at least k of them. From that set we construct two top-k retrievals, a near arm holding the candidates most similar to the query in phenotype-composition space and a far arm holding the least similar. Because every selected slide carries the same maximum label relevance, the two retrievals receive identical label-defined scores by construction. We then compare the two returned sets with the primary layer 2a measure. The retrievals are selected using phenotype composition, whereas layer 2a scores them using mean-pooled slide representations and cosine similarity from the same underlying patch encoder, and we also rescore the same returned slides in UNI, Virchow2 and the Prov-GigaPath tile representation. The construction therefore holds the label-defined answer at its maximum and varies only which slides are selected inside it.

Under the conventional cancer-type definition both arms attain 1.0000 on Hit@5, majority vote at 5 and AP@5, and under the subtype and stage definitions both attain the maximum PathRet layer 1 NDCG on the corresponding axis. Both arms are drawn from the same set of maximally relevant candidates, so the near arm is not the best layer 2a score in the whole pool.

Despite identical label-defined scores, the two retrievals differ substantially under layer 2a, with gaps of $+0.2065$ for cancer type, $+0.1752$ for subtype and $+0.1503$ for stage (Table 3). Even with finer subtype or stage matching, slides with the same maximum relevance can still differ substantially in visual relatedness to the query.

The separation persists when the same returned slides are rescored in the other three representations. All twelve gaps, three relevance definitions by four representations, are positive and range from $+0.1383$ to $+0.2367$, and every interval excludes zero (Appendix B.3).

Section 5.2 shows that real systems return different cases while scoring the same; this section shows how visually different equally relevant retrievals can still be.

7 Robustness of the Reported Comparisons

Section 5.3 reports a reversal between diagnostic relevance and visual relatedness for PRISM and Prov-GigaPath. We test whether that comparison changes under two aspects of the evaluation protocol, the Source-Site exclusion and the fixed scoring representation.

7.1 The recorded Source Site

The scoring representation is associated with Source Site. Among pairs matched on the clinical label, same-Source-Site slides receive higher similarity than different-Source-Site slides on all three label conditions, with every interval above zero. The primary PathRet protocol therefore excludes candidates that share the query's normalized GDC Source Site, which removes direct same-site

Table 3: Controlled stress test at $k = 5$. Near and far arms are selected among maximum-relevance candidates by phenotype-composition similarity and scored with layer 2a. Values are means with 95% patient-cluster bootstrap intervals.

Relevance definition	Matched q	Near arm 2a	Far arm 2a	Gap, 95% interval
Cancer type	4,249	0.9106	0.7041	$+0.2065$ $[+0.2041, +0.2089]$
Subtype	4,099	0.9058	0.7306	$+0.1752$ $[+0.1725, +0.1778]$
Stage	4,054	0.8938	0.7435	$+0.1503$ $[+0.1476, +0.1530]$

retrieval without removing Source-Site-associated information from the representation itself. Appendix B.5 gives the three contrasts and their weightings.

Relaxing that exclusion does not change the comparison of Section 5.3. Read on the same rankings with only the query and its patient excluded, PRISM still exceeds Prov-GigaPath on both label axes and Prov-GigaPath still exceeds PRISM on both visual measures, with every interval excluding zero. Appendix B.5 gives the seven-system values under both settings together with how often each system returns a same-site slide.

7.2 The scoring representation

The same frozen returned identifiers of the seven systems were also rescored in UNI, in Virchow2 and in the Prov-GigaPath tile representation. The PRISM against Prov-GigaPath contrast keeps its direction in all four fixed representations and every paired interval excludes zero, so that comparison is not a property of CONCH v1.5. The complete seven-system visual ordering is a different quantity and is sensitive to the scoring representation, which Appendix B.4 reports in full. We therefore treat the cross-representation pairwise result as the robust finding, rather than the complete ranking of all seven systems.

8 Discussion and Limitations

Label-defined evaluation answers class recovery directly, and Proposition 1 fixes what it can see while doing so: once relevance has been assigned, the metric receives the relevance sequence and not the slides, so which case was selected inside a relevance-equivalent set is outside what it can express. Section 5.2 shows that real systems use that freedom: on queries where two systems each return five slides of the query's exact subtype, they still share no slide on most queries while every label-defined metric assigns both the same maximal score. Section 6 measures the other end of the same quantity, how far apart two retrievals can be placed once the label-defined answer is fixed at its maximum.

Reporting the two targets separately is what makes either of them readable, since a comparison can come out one way on diagnostic relevance and the other way on visual relatedness and a single combined number would suppress one of the two answers. Fixing the cohort, the candidate pool, the relevance rules, the cut-off and the visual evaluator, and releasing them, is what lets a difference between two systems be read as a difference in retrieval behaviour rather than in protocol.

Three things bound what these results establish. The cohort is drawn from a single public archive, so the numbers characterize methods and protocols on this cohort rather than across external institutions or deployment settings. Layer 2 is a fixed learned representation and its ordering has not been calibrated against pathologist judgement, so absolute values and complete visual rankings are properties of the specified scorer, and reader calibration is the external validation this benchmark does not yet carry. Layer 1 is bounded by the recorded clinical ontology, so a dimension of clinical similarity that cancer type, subtype and stage do not encode is not represented in it.

9 Conclusion

Whole-slide retrieval hands a pathologist cases to open, and automated evaluation maps those cases to diagnostic relevance before scoring them. That mapping is what makes the scores comparable, and it is also what makes instance identity invisible to them. Real systems reach into the space this leaves. Where two systems return five slides of the query's exact subtype, so that every label-defined metric here is saturated and identical, their returned sets can share no slide at all, and finer grading shrinks that blind spot without removing it. PathRet therefore reports label-defined clinical relevance and representation-based visual relatedness as two targets under one fixed protocol, never combines them, and audits the fixed evaluator instead of assuming it. The released cohort, query lists, coordinates and scoring code let a later method be measured on the same terms.

REPRODUCIBILITY STATEMENT

We release the benchmark resources rather than describing them. These are the cohort specification and its exclusion counts, the per-axis query and candidate definitions, the patch coordinates, the phenotype dictionary, the Source Site mapping, the scoring code, the stress-test retrievers of Section 6, and the corrected clinical label extraction of Appendix A.2 together with the labels it produces.

The analysis plan was fixed before the remaining baseline results were available. The exact artifact and SHA-256 supporting every reported analysis, together with its generation provenance, are indexed in the released canonical results manifest. Several result files were regenerated to correct artifact labels and field naming, and recursive numeric comparison confirmed that no reported value changed.

Layer 2a reads up to 256 patches per slide at $20\times$ and 512 pixels square, drawn without replacement under a seed derived from the slide identifier, so the draw is the same on every run, and a slide with fewer eligible patches contributes all of them and is not excluded. Layer 2b uses the fixed all-patch phenotype composition over $K = 128$ visual phenotypes rather than that budget. Intervals are patient-cluster bootstrap intervals, paired inside the draw when two methods are compared, as Appendix A.3 sets out. Every method is scored at $k \in \{1, 3, 5, 10, 20, 50, 100\}$ and the released per-method artifacts carry all seven; the main text reports $k = 5$.

AI USE STATEMENT

The authors used generative AI assistants to critique and refine the research methodology and the experimental design, to suggest robustness checks and alternative analyses, to assist with code implementation, refactoring and debugging, to audit experimental provenance and cross-check reported results against the code and data that produced them, to discuss statistical design and the interpretation of results, and to assist with drafting and editing this manuscript.

All AI-assisted analyses, code changes, experimental outputs and manuscript claims were reviewed and verified by the authors. The final scientific decisions, the reported results and the conclusions were determined by the authors, who take responsibility for the complete content of this submission.

REFERENCES

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 5–14. ACM, 2009. doi: 10.1145/1498759.1498766.

Saghir Alfasly, Ghazal Alabtah, Sobhan Hemati, Krishna Rani Kalari, Joaquin J. Garcia, and H. R. Tizhoosh. Validation of histopathology foundation models through whole slide image retrieval. *Scientific Reports*, 15:3990, 2025. doi: 10.1038/s41598-025-88545-9.

Chengkuan Chen, Ming Y. Lu, Drew F. K. Williamson, et al. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 2022.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 659–666. ACM, 2008. doi: 10.1145/1390334.1390446.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31:3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Shivam Kalra, Hamid R. Tizhoosh, et al. Yottixel. an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 2020.

Isaiah Lahr, Saghir Alfasly, Peyman Nejat, Jibran Khan, Luke Kottom, Vaishnavi Kumbhar, Areej Alsaafin, Abubakr Shafique, Sobhan Hemati, Ghazal Alabtah, Nneka Comfere, Dennis Murphree, Aaron Mangold, Saba Yasir, Chady Meroueh, Lisa Boardman, Vijay H. Shah, Joaquin J. Garcia, and H. R. Tizhoosh. Analysis and validation of image search engines in histopathology. *IEEE Reviews in Biomedical Engineering*, 18:350–367, 2025. doi: 10.1109/RBME.2024.3425769.

Tianhao Lei, Parsa Esmaeilkhani, Saghir Alfasly, Wataru Uegami, Judy C. Boughey, Matthew P. Goetz, Krishna R. Kalari, and H. R. Tizhoosh. Multi-institutional pan-cancer validation of pathology foundation models for whole-slide image retrieval using TCGA. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-03122-2.

Shengrui Li, Yubo Zhao, Jiacheng Zhang, Tuo Yu, Ji Zhang, and Yue Gao. High-order correlation-guided slide-level histology retrieval with self-supervised hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11008–11023, 2023. doi: 10.1109/TPAMI.2023.3269810.

Jimmy Lin and Dina Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 383–390. Association for Computational Linguistics, 2006. doi: 10.3115/1220835.1220884.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 2024.

Lena Maier-Hein, Annika Reinke, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 2024.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton JV Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H. R. Tizhoosh. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. PRISM: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

Helen H. Shang, Mohammad Sadegh Nasr, Jai Prakash Veerla, Jillur Rahman Saurav, Amir Hajighasemi, Parisa Boodaghi Malidarreh, Manfred Huber, Chace Moleta, Jitin Makker, and Jacob M. Luber. Histopathology slide indexing and search — are we there yet? *NEJM AI*, 1(5), 2024. doi: 10.1056/AIcs2300019.

H. R. Tizhoosh and Liron Pantanowitz. On image search in histopathology. *arXiv preprint arXiv:2401.08699*, 2024.

Hongyi Wang, Zhengjie Zhu, Junlin Hou, Jiabo Ma, Fang Wang, Yue Shi, Qiuyu Cai, Jili Wang, Bo Luo, Zhizhong Chai, Zhengyu Zhang, Li Liang, Xiuming Zhang, Yen-Wei Chen, Lanfen Lin, and Hao Chen. Accurate and scalable multimodal pathology retrieval via attentive vision-language alignment. *arXiv preprint arXiv:2510.23224*, 2025.

Xiyue Wang, Yuexi Du, Sen Yang, et al. RetCCL. clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 2023.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, and Hongping Tang. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634:970–978, 2024. doi: 10.1038/s41586-024-07894-z.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of AI models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zisha Zhong, Jinlin Huang, Xinjing Li, Lijuan Song, Lanqi Gong, Beibei Ru, Rohit Paul, Jason Levine, Yu Zhang, Kenneth Aldape, and Peng Jiang. Web engine for tumor pathology image retrievals on massive scales. *bioRxiv 2025.10.25.684566*, 2025. doi: 10.1101/2025.10.25.684566.

Shujin Zhu, Yuehua Li, Shivam Kalra, and H. R. Tizhoosh. Multiple disjoint dictionaries for representation of histopathology images. *Journal of Visual Communication and Image Representation*, 55:243–252, 2018. doi: 10.1016/j.jvcir.2018.06.001.

Xinyu Zhu, Zhiguo Jiang, Kun Wu, Jun Shi, and Yushan Zheng. Lifelong histopathology whole slide image retrieval via distance consistency rehearsal. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15004 of *Lecture Notes in Computer Science*, pp. 274–284, 2024. doi: 10.1007/978-3-031-72083-3_26.

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. *arXiv preprint arXiv:2408.00738*, 2024.

A Benchmark construction and reproducibility

A.1 Cohort construction in detail

Of the $4{,}315$ slides whose case records all three axes, $4{,}249$ form the released cohort. A slide is excluded when the file records no magnification and its level-zero image is a reduced copy capped at $10{,}000$ pixels, which leaves the patch size undefined, and the rest could not be opened. Of the $4{,}315$ metadata-eligible slides, $4{,}249$ were retained, and the remaining 66 either lacked a usable magnification definition or could not be opened.

The exclusion unit is not the tissue source site code. Within the released cohort of $4{,}249$ slides, 263 distinct tissue source site codes resolve to 113 recorded GDC Source Site labels and, after lexical normalization, to 100 normalized GDC Source Site labels. Of those normalized labels, 60 carry more than one source site code and together account for $3{,}771$ slides. The mapping is released with the cohort because the primary setting excludes the query's normalized Source Site label rather than its code. The normalization is lexical and is not a judgement about which sites are the same place.

Clinical labels come from the Genomic Data Commons case record. A case there carries one diagnosis entry per recorded tumour, in an order the interface does not guarantee, so reading the first entry mislabels a case whose earlier or metastatic tumour is listed first. Selecting the entry marked as the primary tumour instead changes the diagnosis of 5.5% of cases, the grade of 8.0% and the stage of 15.8%, and recovers stage for 646 cases that the first-entry rule leaves without one. Those cases would otherwise drop silently out of a cohort that requires all three axes.

Table 4: Cases whose label changes when the primary tumour is selected rather than the first listed entry, on the scored cohort of 4,039 cases. Grade sits inside the per-tumour array and is affected, which was not visible on a five-project cohort where grade is almost never recorded.

Field	Cases whose value changes
Diagnosis	5.5%
Grade	8.0%
Stage	15.8%
Cancer site and project	0%

Tumour grade is recorded in the same case entry and is not used as a PathRet relevance axis. It is present for 1,540 of the 4,249 released slides, and seven of the thirteen projects record it for none of their slides, so it does not define a relevance axis that every query in the cohort can be scored on.

A.2 Label quality

A benchmark rests on the reliability of its annotation, so the source records were audited before the cohort was fixed. A case record in this archive carries one diagnosis entry for every tumour ever recorded for the patient, including earlier cancers and metastases, and the interface does not guarantee their order. Taking the first entry assigns a patient-dependent fraction of cases the wrong diagnosis, the wrong stage and the wrong grade, and the results are plausible disease names, which is why the error survives inspection. Selecting the entry marked as the primary tumour fixes it.

Subtype labels are taken from the GDC `primary_diagnosis` of the primary-tumour diagnosis record. In TCGA-TGCT this categorical label can itself denote a mixed tumour, for instance *Mixed germ cell tumor*, ICD-O 9085/3, while component identities and proportions are not available in the metadata PathRet uses. Layer 1 therefore scores the recorded category rather than an inferred histologic composition. This illustrates the broader limitation Proposition 1 formalises, that label-defined relevance cannot resolve variation the recorded label ontology does not encode.

Table 4 reports the effect on the scored cohort.

Fields recorded outside the per-tumour array do not move at all, which identifies the cause rather than describing the symptom.

The more damaging consequence is not a wrong label. Under the first-entry rule 646 cases carry no stage at all, because the first tumour listed happens to have none. A cohort requiring all three axes would have dropped those cases without recording that it had done so, and a missing slide leaves no trace in a results table.

A.3 Statistical protocol

The cohort holds 4,249 slides from 3,980 patients, and some patients contribute several slides, so the resampling unit is the patient. Retrieval-score uncertainty and paired differences between methods are estimated by a patient-cluster bootstrap over 2000 replicates, resampling patients and carrying every slide of a resampled patient together. Method comparisons are paired inside the same draw, since both methods answer the same released query list and an unpaired interval discards that. Point estimates are the arithmetic mean over eligible query slides.

The Source-Site contrast of Section 7.1 is defined per anchor slide rather than per retrieval, so it uses an anchor-balanced point estimand with patient-cluster uncertainty. Pair-weighted and Source-Site-balanced weightings are reported beside it in Appendix B.5.

A difference is reported with its interval. A difference whose interval covers zero is reported as covering zero rather than as an ordering.

A.4 Implementation of the seven retrieval systems

Table 5 records what each system retrieves and how its ranking is formed. Three of the rows are published retrieval systems run by their own rules, one is a foundation model run by the retrieval

Table 5: What each of the seven systems retrieves and how it ranks. A published retrieval pipeline is run by the rule its own paper specifies. A common adaptation is the shared construction of Section 5.1, applied identically to the three foundation models by one script. No entry is fitted on the PathRet cohort.

Method	Representation ranked	Slide aggregation	Ranking rule	Implementation
TITAN (Ding et al., 2025)	TITAN slide embedding over CONCH v1.5 patch features	the TITAN slide model	pool-centred, L2, ascending L_2	published retrieval rule
PRISM (Shaikovski et al., 2024)	1280-d image embedding over Virchow tile features	the PRISM slide model	pool-centred, L2, ascending L_2	common adaptation
Yottixel-K (Kalra et al., 2020; Riasatian et al., 2021)	1024-bit mosaic barcodes over KimiaNet features	none, a bag of barcodes	median of minimum Hamming distance	published pipeline
CHIEF (Wang et al., 2024)	768-d WSI feature over CTransPath tile features	the CHIEF slide model	pool-centred, L2, ascending L_2	common adaptation
Prov-GigaPath (Xu et al., 2024)	768-d LongNet slide embedding over its tile encoder	the Prov-GigaPath slide encoder	pool-centred, L2, ascending L_2	common adaptation
Yottixel (Kalra et al., 2020)	1024-bit mosaic barcodes over ImageNet DenseNet-121 features	none, a bag of barcodes	median of minimum Hamming distance	published pipeline
HSHR (Li et al., 2023)	hypergraph over self-supervised slide codes	collection-wide incidence matrix	published hypergraph similarity	published pipeline, author ckpt

rule its own paper specifies, and three are foundation models turned into retrieval baselines by one shared construction.

The two Yottixel variants are the same published pipeline. Each slide is reduced to a mosaic of patches, each patch to a 1024-bit barcode, and two slides are compared by the published bag-of-barcodes distance, the median over the query's barcodes of the smallest Hamming distance to any barcode of the candidate. The variants differ in one component, the patch feature extractor that produces the barcodes, which is a DenseNet-121 pretrained on ImageNet for the original and KimiaNet (Riasatian et al., 2021) for Yottixel-K. The mosaics, the barcode length and the distance are identical, so the pair isolates the effect of that feature extractor inside a fixed retrieval rule, which is why both are reported.

The three foundation-model rows share one ranking script and therefore one retrieval construction. What differs between them is the released slide representation being ranked, and in each case that representation is the model's own output rather than a pooling introduced here. PRISM contributes its 1280-dimensional image embedding over Virchow tile features, CHIEF its 768-dimensional WSI feature over CTransPath tile features, and Prov-GigaPath its 768-dimensional LongNet slide embedding over its own tile encoder. The centring step reads the mean of the candidate pool, so the transform applied to a query depends on which slides are in the pool, and no benchmark label enters it.

A.5 The protocol audit in full

Each paper selects its own cohort, its own metric and its own normalisation, and Table 6 records those choices for the published work we could read in full. The last of its eight rows is not a retrieval system but a study that reproduced four of them, and it is recorded here because it makes the same choices any evaluation must make. No two rows agree on enough columns for their numbers to be placed side by side. The two most standardised entries settle different halves of the problem. Lei et al. (2026) fix a public cohort and a candidate pool and grade by majority voting and macro F1, so what is standardised is everything around a metric that Proposition 1 covers. Zhong et al. (2025) put the returned slides in front of three pathologists, which is the question labels cannot answer, on a collection that is not public. A later method can reproduce the first pool without the second judgement and obtain the second judgement only by convening its own readers.

Three properties of the table are worth stating separately.

The metric families do not agree. Accuracy at one, majority vote at k, mean average precision and macro F1 all appear, and the first two are reported under the same phrase. Majority vote asks whether most of the top k carry the query label and accuracy at k asks whether any of them does,

Table 6: What each published evaluation of whole-slide retrieval chose. The columns are the properties that must agree before two reported numbers describe the same task. A cell marked *not reported* was not stated in the source paper, which is itself the finding for the candidate-pool column. Rows are restricted to work read in full.

Work	Cohort	Patch encoder	Metrics	Candidate pool	Source control
Yottixel (Kalra et al., 2020)	300 in-house + 2,020 public	frozen ImageNet CNN	accuracy, majority vote	*not reported*	none
SISH (Chen et al., 2022)	public archive	self-supervised on the same archive	mAP@5, majority vote	*not reported*	none
RetCCL (Wang et al., 2023)	public archive	contrastive on the same archive	majority vote	*not reported*	none
Foundation-model validation (Alfasly et al., 2025)	11,444 slides, 23 organs, 117 subtypes	four foundation models	macro F1 at top-1, 3, 5	*not reported*	none
PathSearch (Wang et al., 2025)	one public archive + 4 public + 3 in-house	vision-language	Acc@1, MV@3, MV@5	*not reported*	none
Slide-FM validation (Lei et al., 2026)	9,387 slides, 17 organs, 60 diagnoses, one public archive	four slide FMs, ABMIL, patches	macro F1, MV@n	within-organ; leave-one-patient-out	none
HERE (Zhong et al., 2025)	101 queries against a private collection	self-supervised	three pathologists score the top 5	private	none
Search-engine analysis (Lahr et al., 2025)	4 internal (1,269 patients) + 3 public (1,216 patients)	autoencoder, KimiaNet, DenseNet, CCL	F1 at top-1, MV@3, MV@5	leave-one-patient-out; five-fold for one engine	none
This benchmark	4,249 slides, 13 projects, one public archive	fixed by the specification	graded NDCG per axis, and layer 2	released query lists	three settings

which are different questions with the same name in the same tables (Tizhoosh & Pantanowitz, 2024).

Specifying a protocol is not the same as being comparable. Two of the eight rows define the retrieval pool. Lei et al. (2026) restricts candidates to the query's organ and excludes every slide of the query's patient. Lahr et al. (2025) evaluates three of its four engines leave-one-patient-out and the fourth, which fits an autoencoder on each dataset it is applied to, by five-fold cross-validation, and reports all four in one table. The remaining six do not fully specify the candidate pool. Specification is what makes a protocol reproducible, and it is not what makes two numbers comparable. A pool restricted to the query's organ and a pool over a whole archive pose different retrieval tasks, and a table whose rows were validated under two schemes places two tasks beside each other. Where the pool is not specified at all, neither question can be put to the paper. This archive holds several slides per patient, and a second section of the same block is a near copy, so leaving them in makes the task substantially easier in a way that cannot be reconstructed.

Cohorts include data that cannot be obtained. Yottixel is evaluated on three hundred slides from one hospital together with a public subset, and the most recent method we read adds three in-house cohorts to four public ones. Those numbers cannot be reproduced, and they are cited as baselines.

The field's own survey does not compare protocols. Tizhoosh & Pantanowitz (2024) review these systems and compare them by their architecture, dividing each into how it selects patches, how it encodes them and how it indexes them. There is no table of what each one measured. The survey also notes that the separation between training and evaluation data could not be verified for two of the systems it reviews, and Table 6 records that several of the encoders in use were pretrained on the archive they are evaluated on.

Retrieval quality is defined as label agreement. The most recent method we read marks each returned slide in its qualitative figures with a check when the label matches the query and a cross when it does not, and reports a case in which all three compared systems return five correct slides out of five. Section 3.2 is a statement about what those marks can and cannot distinguish.

Table 7: Pretraining provenance and archive exposure for the encoders whose exposure is material to the reported analyses, separated by what the primary sources establish. The exposure column uses the four statuses of Table 2. For the evaluator the exact checkpoint's exclusion of this archive is not directly documented in the provenance available to us, so it is recorded as unresolved rather than as verified or as violated.

Encoder and role	Pretraining corpus	TCGA pretraining exposure
CONCH v1.5, primary evaluator	biomedical literature figures, restored from a UNI checkpoint	Unresolved for the exact checkpoint
UNI, rescoring only	two hospitals and an expression atlas	No
Virchow2, robustness only	3.1 million slides in an internal MSKCC dataset	Not reported, not stated in the sources reviewed
Prov-GigaPath tile, robustness only	Prov-Path, 171,189 slides from the Providence network	No, TCGA is described as downstream evaluation
TITAN, published baseline	Mass-340K, 335,645 slides, read over CONCH v1.5 features	Unresolved, stated for the slide model but not for CONCH
KimiaNet, retriever	the archive, 7,126 slides	Yes
RetCCL, retriever	the archive in full and one more	Yes

A.6 Pretraining provenance and archive exposure

If the fixed encoder was pretrained on the benchmark cohort, independence between the evaluator and the data it grades is compromised. Because archive exposure can change how a result should be read, we document the available pretraining provenance and archive-exposure status of the evaluator, of the robustness encoders and of the retrieval encoders for which this issue is material to the reported analyses.

The designated evaluator encoder is a vision-language model pretrained on biomedical literature figures rather than on whole-slide collections (Lu et al., 2024). Whether the exact checkpoint the specification fixes excludes this archive is not directly documented in the provenance available to us, so we state the share of the cohort provably held out from it as not directly documented rather than as none or as all. The condition above is therefore one this benchmark states and cannot verify for its own evaluator from the sources available. Two further representations appear only as the robustness rescoring of Appendix B.3. The Virchow2 model card gives its pretraining dataset as 3.1 million whole-slide images in an internal Memorial Sloan Kettering dataset, and the paper describes the originating institutions of those slides as diverse (Zimmermann et al., 2024), so whether this archive is among them is not stated in the sources reviewed. Prov-GigaPath is pretrained on Prov-Path, 171,189 slides from about 30,000 patients in the Providence health network, and that work uses this archive as a downstream evaluation set rather than as pretraining data (Xu et al., 2024). Neither is recorded as excluded from this archive, because neither source states an exclusion. TITAN appears as a published baseline in Table 2. Its slide model is pretrained on Mass-340K, 335,645 whole-slide images from private collections and the public GTEx consortium, and the authors state that this archive is not among them (Ding et al., 2025). TITAN reads CONCH v1.5 patch features, however, and the exclusion of this archive from that exact checkpoint is the unresolved item above, so the slide model and the patch encoder are recorded separately rather than as one clearance.

The retrievers are a different matter. Two of the patch encoders in common use for whole-slide retrieval were pretrained on the same public archive this benchmark draws on. KimiaNet was trained on more than 240,000 patches drawn from 7,126 slides of that archive (Riasatian et al., 2021), and RetCCL was pretrained on all 32 thousand of its slides together with a second public collection (Wang et al., 2023). For RetCCL the cohort is therefore contained in the pretraining corpus by the authors' own description, and no part of it is held out. KimiaNet releases a test-set manifest but not the list of slides it trained on, so the overlap with our cohort bounds from one side only and cannot be computed. Table 7 separates what the primary sources document from what follows only from lineage and from what is unresolved.

Declaring pretraining. A submission declares the pretraining corpus of every encoder it uses and states what share of the benchmark cohort is provably absent from it. Methods are reported together

in one table with their pipeline-level archive exposure stated as a column, so a reader can see which entries were exposed to the archive without the entries being separated from one another.

B EVIDENCE BEYOND LABEL AGREEMENT

B.1 IDENTITY OVERLAP UNDER LABEL-PERFECT RETRIEVAL

Section 5.2 conditions on the queries where both systems return five slides of the query's exact subtype. On those queries each system still chooses five slides out of a pool of candidates that all carry that subtype, and the size of that pool sets how much two selectors would share by opportunity alone. Across the $4{,}090$ queries with at least five such candidates the pool has a median of 216 slides and an interquartile range of 62 to 343. On the 474 queries that qualify for the PRISM and Yottixel-K pair it is larger, with a median of 311.5 and an interquartile range of 181 to 355.5, because requiring both systems to be exact-subtype perfect favours queries from larger subtypes.

The baseline is exact rather than simulated. Two independent uniform draws of five slides from a pool of N share j of them with probability $\binom{5}{j}\binom{N-5}{5-j}/\binom{N}{5}$, and averaging over the queries of a pair gives the opportunity baseline for that pair. For PRISM against Yottixel-K the baseline puts zero shared slides at 0.9032 and the mean number of shared slides at 0.1015. The two systems reach 0.7278 $[0.6900, 0.7665]$ and 0.3376 $[0.2851, 0.3937]$ respectively, so the observed zero-overlap share sits -0.1754 $[-0.2120, -0.1366]$ below the baseline and the observed mean overlap sits $+0.2361$ $[+0.1833, +0.2913]$ above it. The two systems are therefore far more alike than independent selection, and the identity variation that remains is what the saturated label metrics cannot express.

B.2 THE CONSTRUCTIVE COUNTEREXAMPLE IN FULL

Table 3 in Section 6 reports all three relevance definitions with their matched-query denominators. The near-composition and far-composition arms are the nearest and the furthest candidates of maximum relevance in all-patch phenotype-composition space. A query enters a row only when the relevance definition in force leaves at least k eligible candidates of maximum relevance, which is why the three rows have different denominators. The cancer-type row uses the relevance definition conventional evaluations adopt, which PathRet itself uses as a gate rather than as a reported axis, and the subtype and stage rows are the two graded layer 1 axes on which both arms additionally attain the maximum NDCG.

B.3 RESCORING THE SAME RETRIEVALS ACROSS FIXED REPRESENTATIONS

Selection happened once. The near-composition and far-composition arms of Section 6 were chosen in all-patch phenotype-composition space, their returned slide identifiers were frozen, and those identifiers were then rescored through four fixed representations. No retrieval was rerun and no candidate pool changed, so every row of Table 8 reads the same returned slides and only the scoring representation varies.

UNI shares model lineage with the evaluator, whose vision tower is restored from a UNI checkpoint (Appendix A.6), so it tests whether the separation is specific to the evaluator's own parameters. Virchow2 and the Prov-GigaPath tile encoder were pretrained outside that lineage, and they are what carries the result beyond it. Agreement among the four is agreement among learned representations and is not a reader judgement.

Virchow2 is read at $2{,}560$ dimensions, the class token of $1{,}280$ concatenated with the mean of the patch tokens, with the four register tokens excluded. The Prov-GigaPath tile encoder is read at $1{,}536$ dimensions. Both read the same released $20\times$ 512-pixel source tiles as layer 2a and then apply their own deterministic preprocessing, which for Prov-GigaPath is a resize to 256 followed by a centre crop to 224. The source tile geometry is shared and the effective field of view is therefore not identical.

The auxiliary layer 2b measure also separates the same arms, and it is construction-aligned, since the arms are selected in the phenotype-composition space that layer 2b scores. It is therefore not independent evidence and the separation it reports is not comparable with the layer 2a values above.

Table 8: The same frozen retrievals read through four fixed representations. The returned slide identifiers are identical in every row and only the scoring space varies. CONCH v1.5 is the primary evaluator and repeats Table 3, UNI shares model lineage with it, and Virchow2 and the Prov-GigaPath tile encoder were pretrained outside that lineage. All twelve gaps are positive and no interval covers zero. Intervals are patient-cluster bootstrap, paired inside the draw. None of the four is a clinical reference.

Representation	Axis	Near arm	Far arm	Gap, 95% interval
CONCH v1.5, primary evaluator	Cancer type	0.9106	0.7041	$+0.2065$ $[+0.2041, +0.2089]$
	Subtype	0.9058	0.7306	$+0.1752$ $[+0.1725, +0.1778]$
	Stage	0.8938	0.7435	$+0.1503$ $[+0.1476, +0.1530]$
UNI, shared lineage	Cancer type	0.6273	0.4390	$+0.1883$ $[+0.1850, +0.1918]$
	Subtype	0.6208	0.4556	$+0.1652$ $[+0.1618, +0.1687]$
	Stage	0.6077	0.4671	$+0.1406$ $[+0.1371, +0.1440]$
Virchow2, distinct lineage	Cancer type	0.7623	0.5255	$+0.2367$ $[+0.2335, +0.2401]$
	Subtype	0.7547	0.5464	$+0.2082$ $[+0.2047, +0.2118]$
	Stage	0.7368	0.5592	$+0.1776$ $[+0.1740, +0.1811]$
Prov-GigaPath tile, distinct lineage	Cancer type	0.6506	0.4683	$+0.1822$ $[+0.1789, +0.1856]$
	Subtype	0.6439	0.4836	$+0.1603$ $[+0.1567, +0.1638]$
	Stage	0.6347	0.4964	$+0.1383$ $[+0.1347, +0.1419]$

B.4 The real systems read in four fixed representations

The returned slide identifiers of the seven systems in Table 2 were rescored in four fixed representations at $k = 5$ over the same $4{,}240$-query common support, under the same exclusions and the same patient-cluster bootstrap. No retrieval was rerun and no candidate pool changed, so only the space the score is read in varies. Table 9 reports the result.

Two things follow, and they point in different directions. The pairwise reversals that Section 5.3 reports are preserved in every one of the four representations. PRISM scores below Prov-GigaPath by -0.0137 $[-0.0153, -0.0122]$ under CONCH v1.5, -0.0881 $[-0.0900, -0.0863]$ under UNI, -0.0467 $[-0.0488, -0.0447]$ under Virchow2 and -0.1149 $[-0.1171, -0.1130]$ under the Prov-GigaPath tile representation, and Yottixel-K scores below Prov-GigaPath by -0.0378 $[-0.0395, -0.0362]$, -0.0604 $[-0.0623, -0.0586]$, -0.0592 $[-0.0612, -0.0572]$ and -0.0910 $[-0.0928, -0.0893]$ respectively. All eight intervals exclude zero.

The full seven-system ordering is not stable. Spearman correlation between the CONCH v1.5 ordering and the others is 0.9286 for Virchow2, 0.3214 for UNI and 0.2857 for the Prov-GigaPath tile representation, and TITAN moves from first under CONCH v1.5 to fifth under UNI and under Prov-GigaPath. A system's position on this layer is therefore a property of the pair of representations being compared and not of the system alone, which is why PathRet fixes one scorer and reports it as such.

Prov-GigaPath is first under UNI, under Virchow2 and under the Prov-GigaPath tile representation, and second to TITAN under CONCH v1.5. One of the four representations is its own tile encoder, and we do not read its position under that scorer as evidence of intrinsic retrieval superiority, because a retriever and a scorer that share a representation are aligned by construction. None of the three alternative representations is a clinical reference, and this analysis measures agreement between fixed instruments rather than correctness.

B.5 Source-Site sensitivity and system robustness

The evaluator's own Source-Site association. Table 10 reports the anchor-balanced contrast between same-Source-Site and different-Source-Site pairs at a fixed clinical label, which is what motivates the primary exclusion. Two other weightings are available and we report them rather than choose among them by size. Weighting pairs equally gives $+0.0272$ on cancer type, $+0.0564$ on subtype and $+0.1184$ on stage. Weighting Source Sites equally gives $+0.0405$, $+0.0502$ and $+0.1226$. The pair-weighted estimate is sensitive to the concentrated pair distribution. We tested

Table 9: The same returned identifiers of the seven systems read in four fixed representations, at $k = 5$ on the 4,240-query common support. Only the scoring space varies between columns. CONCH v1.5 is the specified evaluator, UNI shares model lineage with it, and Virchow2 and the Prov-GigaPath tile encoder lie outside that lineage. None of the three alternatives is a clinical reference. Intervals are 95% patient-cluster bootstrap intervals.

Method	CONCH v1.5	UNI	Virchow2	Prov-GigaPath tile
TITAN	0.9044 [0.9028,0.9061]	0.6526 [0.6496,0.6557]	0.7622 [0.7593,0.7649]	0.6730 [0.6693,0.6764]
PRISM	0.8866 [0.8849,0.8885]	0.6275 [0.6243,0.6305]	0.7419 [0.7390,0.7446]	0.6510 [0.6475,0.6543]
Yottixel-K	0.8626 [0.8600,0.8651]	0.6552 [0.6518,0.6585]	0.7294 [0.7260,0.7327]	0.6749 [0.6711,0.6787]
CHIEF	0.8664 [0.8642,0.8687]	0.6679 [0.6651,0.6706]	0.7473 [0.7443,0.7503]	0.7049 [0.7017,0.7082]
Prov-GigaPath	0.9004 [0.8984,0.9023]	0.7156 [0.7129,0.7183]	0.7887 [0.7859,0.7914]	0.7659 [0.7627,0.7690]
Yottixel (orig.)	0.8343 [0.8310,0.8375]	0.6534 [0.6502,0.6564]	0.7251 [0.7215,0.7286]	0.6757 [0.6714,0.6797]
HSHR	0.8072 [0.8041,0.8103]	0.6206 [0.6171,0.6240]	0.6844 [0.6811,0.6877]	0.6508 [0.6468,0.6546]

Table 10: Association between the evaluator's similarity and the recorded GDC Source Site, holding the clinical label fixed. The estimand is anchor-balanced and intervals are patient-cluster bootstrap. Leave-one-Source-Site-out produces no sign change on any axis. This is an association with a recorded label and not a measured effect of a physical preparation or scanning laboratory.

Axis	Same-Source-Site contrast	95% interval	Anchors
Cancer type	$+0.0416$	$[+0.0392, +0.0434]$	4,164
Subtype	$+0.0505$	$[+0.0482, +0.0534]$	4,000
Stage	$+0.1162$	$[+0.1109, +0.1202]$	4,026

whether the high-mass Source Sites carry a lower contrast and did not find that they do, so we do not offer that as the explanation.

A submitted ranking is a full ordering, so reading it under a looser exclusion is what that ranking would have returned. The seven systems were therefore read twice on the same rankings, once under the primary exclusion of self, same patient and same normalized GDC Source Site, and once under an unrestricted setting that excludes only self and the same patient. No retrieval was rerun.

The comparison of Section 5.3 holds under both settings. PRISM exceeds Prov-GigaPath on subtype by $+0.1741$ $[+0.1644, +0.1837]$ under the primary setting and $+0.0565$ $[+0.0489, +0.0645]$ unrestricted, and on stage by $+0.0655$ $[+0.0571, +0.0741]$ and $+0.0090$ $[+0.0005, +0.0172]$. Prov-GigaPath exceeds PRISM on layer 2a by 0.0137 $[0.0122, 0.0153]$ and 0.0246 $[0.0234, 0.0257]$, and on layer 2b by 0.0130 $[0.0102, 0.0157]$ and 0.0274 $[0.0248, 0.0298]$. Every interval excludes zero and every sign is preserved. Yottixel-K against Prov-GigaPath behaves the same way on all four axes.

How often a system returns a same-Source-Site slide varies widely across the seven. Under the unrestricted setting the share of returned slides drawn from the query's own Source Site runs from 0.4057 $[0.3939, 0.4170]$ for PRISM to 0.7895 $[0.7791, 0.8004]$ for Yottixel-K. A same-site slide is often also the clinically closest one, because laboratories differ in which diseases they contribute, so this share is reported as a property of each system under a looser pool rather than as a fault.

C Sensitivity of benchmark choices

C.1 Sensitivity to the partial relevance gain

Layer 1 gives a candidate of the same cancer type but a different subtype, and a candidate of the same major stage but a different substage, a partial gain of $\alpha = 0.5$. That value is a choice, so layer 1 was recomputed from the same frozen rankings at $\alpha \in \{0.25, 0.5, 0.75\}$. The canonical benchmark keeps $\alpha = 0.5$.

The subtype ordering of the seven systems is identical at all three values. The stage ordering changes in one place, at $\alpha = 0.25$, where CHIEF and Prov-GigaPath exchange fourth and fifth position. That pair is already unresolved under the canonical setting, where their stage difference is -0.0011 $[-0.0081, +0.0058]$, and it remains unresolved at $\alpha = 0.25$, where it is $+0.0008$

Table 11: The two layer 1 comparisons of Section 5.3 recomputed at three partial relevance gains from the same frozen rankings. The canonical benchmark uses $\alpha = 0.5$. GigaPath abbreviates Prov-GigaPath. Every entry is positive and every interval excludes zero. Intervals are 95% patient-cluster bootstrap intervals.

Axis	Pair	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$
Subtype	PRISM − GigaPath	$+0.1717$ $[+0.1614, +0.1816]$	$+0.1741$ $[+0.1644, +0.1837]$	$+0.1764$ $[+0.1666, +0.1860]$
	Yottixel-K − GigaPath	$+0.0956$ $[+0.0859, +0.1056]$	$+0.1041$ $[+0.0945, +0.1139]$	$+0.1127$ $[+0.1026, +0.1228]$
Stage	PRISM − GigaPath	$+0.0592$ $[+0.0510, +0.0679]$	$+0.0655$ $[+0.0571, +0.0741]$	$+0.0718$ $[+0.0628, +0.0810]$
	Yottixel-K − GigaPath	$+0.0254$ $[+0.0173, +0.0338]$	$+0.0308$ $[+0.0227, +0.0394]$	$+0.0362$ $[+0.0275, +0.0454]$

$[-0.0060, +0.0076]$. The swap is a change in point estimates that the intervals do not separate at either value.

The two comparisons Section 5.3 reports keep their direction and exclude zero on both axes at all three values, as Table 11 shows. The conclusions drawn in Section 5 therefore do not rest on the canonical α.

C.2 Stability of the evaluator's design choices

Patches read per slide. Layer 2a reads a fixed upper budget of patches per slide rather than all of them, whereas layer 2b is formed over all eligible patches and this budget does not apply to it. The budget below is therefore a layer 2a choice, and it must not be a free parameter. We compare the sampling error it introduces against the spread the evaluator has to resolve. The median cosine between two different slides is 0.6772, with a 5 to 95% range of 0.4798 to 0.8921, and that spread is what the evaluator exists to measure. Against it, two independent draws of 256 patches from the same slide agree at a median cosine of 0.99921, a sampling error of 4.8% of the between-slide spread. A budget of 512 lowers that to 3.2%, and 4.0% of slides have fewer than 512 eligible patches against 0.6% that have fewer than 256. The specification fixes the budget at 256 on that trade-off between sampling precision and coverage, without reference to downstream retrieval performance. Patches are drawn without replacement under a seed derived from the slide identifier, and a slide with fewer eligible patches than the budget contributes all of them rather than being excluded.

An earlier decision under a different patch geometry fixed a count of 512. That setting used 256-pixel patches, so one patch covered a quarter of the tissue area one covers here, and the two numbers do not refer to the same quantity. The budget above is the current one.

C.3 Preprocessing

The segmentation that supplies the evaluator's patches offers two removal options and no evaluation in Table 6 states which it used, so we measured them rather than inheriting a default.

Artifact removal is not used. Over 200 slides from 28 Source Sites it removes a median 14% of the tissue area, takes more than half of it from one slide in eight, and empties 23 slides completely. How much it removes is explained far better by the recorded Source Site than by the cancer type, at $\eta^2 = 0.744$ against 0.308, and the worst-affected Source Site keeps 11.9% of its tissue. A Source-Site-dependent transform applied before scoring would inject the very signal Section 7.1 measures and Section 4.1 excludes.

Pen-mark removal is used, and for the opposite reason. A pen mark is added by a person after the slide is made and is associated with where the slide was prepared, so removing it takes provenance-associated signal out rather than putting it in. Leaving the marks is not a neutral choice either. The two options therefore act on the confound in opposite directions, which is why one is kept and the other is not.

The canonical setting used throughout this paper removes pen marks and does not remove artifacts, at a patch budget of 256. We use this analysis only to characterize the effect of the preprocessing choices and do not infer downstream retrieval-performance differences from it, and it is recorded as a scope statement rather than as a requirement on submissions. Table 12 reports what each setting does to the tissue.

Table 12: What each segmentation setting does to the tissue. This analysis is descriptive and is not used to draw downstream retrieval-performance conclusions.

	No removal	Pen marks removed
Tissue area retained, median	1.000	0.954
Contour fragments, median factor	1.00	1.67
Slides altered	—	96.8%

Declaring preprocessing. A submission declares the segmentation and patching it used, with the tool version and the checkpoint, and the benchmark releases the patch coordinates so that the step need not be reproduced at all. The released coordinates are byte-identical across independent runs on different machines.

C.4 Phenotype mass weighting in layer 2b

Layer 2b also depends on how mass is assigned across the fixed phenotype vocabulary. PathRet fixes abundance weighting as part of the benchmark specification. Alternative weighting schemes change the degree of overlap with the slide-level measure and are therefore treated as alternative metric definitions rather than as interchangeable settings of the same one.

D Optional training regime

D.1 The development split

The primary regime of this paper is the full $4{,}249$-slide evaluation, in which no benchmark-specific fitting is performed using the PathRet cohort. That is separate from prior pretraining on this archive, recorded encoder by encoder in Appendix A.6, and from an inference procedure that reads the collection it searches, which belongs to a method's inference regime and is not training. Trainable methods need a reproducible protocol as well, so we also release a development split.

We constructed a patient-disjoint, metadata-balanced split of $70/15/15$ under seed 3, preserving cancer-type, subtype and stage distributions while using the normalized GDC Source Site as a soft balancing variable. Training holds $2{,}790$ patients and $2{,}967$ slides, validation 593 patients and 640 slides, and test 597 patients and 642 slides, with no patient appearing in two parts. Training is for fitting, validation for hyperparameter and checkpoint selection, and test for final retrieval evaluation. Test queries search candidates drawn from the test split alone, so no training or validation slide enters the test candidate pool, and the candidate exclusions of Section 4.1 apply inside it. This development split defines a separate training regime and is not directly comparable to the full-cohort evaluation.

Table 13 reports that regime.

LWSR. Among the development-split methods reported in Table 13, LWSR (Zhu et al., 2024) is the one trained on the PathRet training split. It was trained on the official training split of $2{,}967$ slides as six sequential tasks, fixed before any scoring, which partition the thirteen projects as COAD with READ, STAD with ESCA, LIHC with CHOL and PAAD, HNSC with THCA, BRCA with TGCT, and MESO with UVM. That partition is one this benchmark fixes for this cohort and is not claimed to be clinically or anatomically optimal. The published model, its losses and its replay mechanism are otherwise retained, with a reservoir replay buffer of 100 and a fixed 70-epoch schedule, and LWSR retrieves by ascending Euclidean distance between the final model's 512-dimensional class-token slide embeddings, with no centring and no normalisation. The scored checkpoint is the final one, task five at epoch 69, and validation was not used to select it. A leakage audit found no training or validation slide in the test pool, and the ranking reproduces position for position from that checkpoint for all 642 test queries.

BoVW. Three things carry this name and they are kept apart here. The original framework is Zhu et al. (2018), which represents a slide by multiple disjoint dictionaries over local binary pattern features. Lahr et al. (2025) analyses a modified configuration of it, in which the local binary patterns are replaced by an autoencoder so that the comparison is made at the deep-learning level, and

Table 13: The development-split regime. Every entry is evaluated on the official frozen test split of 642 slides from 597 patients, where test queries search test candidates only and no training or validation slide enters the pool. Methods that fit on the training split and methods that fit nothing both appear here, so the test pool poses the same task for all of them. These numbers are not comparable with Table 2, which uses a different query population and a different candidate pool, and the same method appearing in both is two results rather than one measured twice. Intervals are patient-cluster bootstrap. This table reports only the methods included in the present development-split evaluation.

(a) Protocol and provenance

Method	PathRet fitting	Checkpoint	Inference
HSHR (author)	No	Author	Transductive
LWSR	Yes	Final	Inductive
TITAN	No	Released	Transductive
BoVW (PathRet adaptation)	Yes	Final	Transductive

(b) Scores on the test split

Method	NDCG subtype	NDCG stage	Layer 2a
HSHR (author)	0.3094 [0.2876, 0.3315]	0.1360 [0.1211, 0.1517]	0.7706 [0.7621, 0.7786]
LWSR	0.6155 [0.5876, 0.6431]	0.2781 [0.2567, 0.3010]	0.8457 [0.8386, 0.8525]
TITAN	0.7669 [0.7449, 0.7888]	0.3339 [0.3105, 0.3586]	0.8810 [0.8757, 0.8860]
BoVW (PathRet adaptation)	0.1269 [0.1133, 0.1402]	0.0515 [0.0431, 0.0605]	0.7090 [0.7007, 0.7171]

that modified configuration is what the row below reproduces. The PathRet run is that configuration adapted to this benchmark. BoVW, in the configuration analysed by Lahr et al. (2025), is fitted on the PathRet development split. Its autoencoder was trained on the training split alone, validation words monitored the loss, no test word was seen during training, and the scored checkpoint is the last of ten epochs with no checkpoint selection. The tracked files of the implementation this run started from, at commit `f41be59` of the public repository, were not modified, but the pipeline that produced the row is not the published entry point: scripts written for this benchmark import those files and replace the patch and word sourcing, the tissue threshold, the words per patch and the clustering implementation. Twenty patches per slide are read from the released canonical $20\times$ 512-pixel coordinates rather than sampled from a thumbnail mask. The 2,048-cluster visual dictionary is fitted on the word embeddings of the 642 test slides, which is the published pipeline's own collection-dependent construction rather than a change made here, and it is why the inference regime is transductive. Retrieval is the published chi-square distance between raw-count visual-word histograms.