# Foundation Model Representations for Scalable Content-Based Retrieval in Whole-Slide Histopathology*

Sofia Matveeva[1][0009−0001−3307−1927], Andrey Krylov[1][0000−0001−9910−4501], and Alexander Khvostikov[1][0000−0002−4217−7141]

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia
matveevasv@my.msu.ru, kryl@cs.msu.ru, khvostikov@cs.msu.ru

**Abstract.** Content-based image retrieval offers a direct way to work with large whole-slide histopathology archives. Instead of relying only on manual annotations, a pathologist or researcher can search for visually similar tissue patterns or related whole-slide cases. Scalable retrieval systems in digital pathology, however, still often use visual descriptors inherited from general-purpose image models. Such descriptors are compact and convenient for indexing, but they are not necessarily well matched to H&E-stained tissue morphology.

We study the use of histopathology foundation model representations in scalable content-based retrieval for whole-slide images. CONCH and Virchow descriptors are integrated into a SISH-like retrieval pipeline and compared with the original DenseNet-based representation. A central practical issue is the mismatch between the input data and the models: whole-slide images, $1024 \times 1024$ tissue fragments, $224 \times 224$ patches, and variable-size patches require different preprocessing strategies. We therefore adapt multi-crop feature extraction, embedding aggregation, and binary descriptor construction to the input requirements of each foundation model.

The system is evaluated on four datasets: TCGA-300-WSI, TCGA-300-Patch, Kather100k, and WSSS4LUAD. The TCGA-300-WSI and TCGA-300-Patch datasets were curated in this study from TCGA whole-slide images. Together, the four datasets cover slide-to-slide and patch-to-patch retrieval scenarios. The results show that pathology-specific foundation models consistently improve retrieval quality over the original DenseNet121-based SISH descriptor across several top-$K$ settings, with Virchow giving the strongest overall performance.

**Keywords:** Whole-slide histopathology · Content-based image retrieval · Image mining · Foundation models · Digital pathology · Scalable retrieval · Visual search.

* Supported by the Russian Science Foundation, project no. 26-11-00118.

## 1 Introduction

The rapid adoption of whole-slide imaging has turned pathology into a digital discipline. A single whole-slide image (WSI) may contain billions of pixels and captures heterogeneous tissue architecture at several spatial scales. As hospitals and research centers accumulate large WSI repositories, retrieval by patient identifiers, textual reports, or manually assigned labels becomes insufficient for many clinical and research tasks. Content-based image retrieval (CBIR) addresses this limitation by retrieving morphologically similar cases directly from visual tissue content.

Content-based retrieval in digital pathology remains challenging because WSIs are gigapixel images, expert annotations are expensive, and histology images vary across staining protocols, scanners, tissue preparation procedures, and institutions. Existing scalable retrieval systems have therefore relied on compact descriptors and efficient indexing mechanisms. However, many such systems still use visual descriptors derived from general-purpose natural-image encoders, such as DenseNet121 pretrained on ImageNet. These descriptors are convenient for indexing, but they are not specifically optimized for nuclei, glands, stroma, necrosis, and other pathology-specific tissue structures.

Pathology foundation models trained directly on large histopathology datasets have recently become powerful feature extractors for computational pathology. CONCH learns visual-language representations from large-scale histopathology image-caption pairs [12], whereas Virchow is a large self-supervised pathology foundation model trained on approximately 1.5 million H&E WSIs [15]. Although such models have shown strong performance in classification, slide representation learning, and Yottixel-based retrieval, their effectiveness within the SISH retrieval framework has not yet been systematically evaluated.

In this work, we examine whether pathology foundation model representations can improve scalable histopathology retrieval while preserving the main search logic of SISH. We distinguish the original SISH baseline from the SISH-like variants considered in this study. The original baseline uses the DenseNet121-based binary descriptor, whereas the proposed SISH-like variants replace this descriptor with CONCH or Virchow representations while retaining the scalable candidate search, binary comparison, and WSI-level ranking mechanisms. Because these foundation models have different input requirements, we also adapt preprocessing, multi-crop feature extraction, embedding aggregation, and Hamming-distance threshold selection.

The main contributions are threefold. First, we define SISH-like retrieval variants that preserve the scalable indexing and ranking components of SISH while replacing the original DenseNet121 descriptor with CONCH or Virchow. Second, we introduce model-specific preprocessing strategies for converting $1024 \times 1024$ histopathology patches into foundation-model embeddings. Third, we evaluate the original SISH baseline and the SISH-like CONCH and Virchow variants on four retrieval benchmarks: TCGA-300-WSI, TCGA-300-Patch, Kather100k, and WSSS4LUAD. The TCGA-300-WSI and TCGA-300-Patch datasets were curated

in this study from TCGA whole-slide images to support controlled slide-level and patch-level retrieval evaluation.

## 2 Related work

**Histopathology image search and scalable retrieval.** Content-based image retrieval in histopathology aims to retrieve cases or local tissue regions that are visually similar to a query image. In large WSI archives, such systems can support case comparison, education, quality assurance, and diagnostic consensus [14, 11]. Retrieval of visually similar archival cases has also been proposed as a way to reduce inter-observer variability and support decision making in difficult diagnostic scenarios [9].

Early histopathology retrieval systems relied on handcrafted color, texture, and morphology descriptors. These descriptors are computationally efficient, but they have limited capacity to represent complex tissue architecture. Deep learning improved descriptor quality by learning visual representations directly from histology images. SMILY showed that convolutional neural network features can be used for reverse image search in histopathology and can retrieve visually similar tissue regions from large image collections [7]. Direct nearest-neighbor search over dense deep features, however, remains computationally challenging in large WSI archives.

Yottixel addressed this scalability problem by representing each WSI as a mosaic of representative tissue patches, extracting DenseNet121 features, converting them into compact binary barcodes, and ranking candidates by Hamming distance [8]. The method showed that WSI retrieval can be made practical by combining tissue sampling, compact descriptors, and efficient comparison. SISH further advanced scalable histopathology retrieval by introducing self-supervised discrete representations, a van Emde Boas tree for fast candidate search, and an uncertainty-based ranking strategy for WSI-level retrieval [3]. Unlike simple descriptor matching, SISH separates discrete candidate retrieval from descriptor-based filtering and ranking. This structure makes it possible to modify the visual descriptor while preserving the efficient indexing mechanism.

Compact binary representations and approximate search remain central to scalable pathology retrieval, especially when databases contain millions of patches or slides. Our work follows this direction. Rather than training a new hashing model or redesigning the indexing procedure, we test whether stronger pretrained pathology descriptors can improve a SISH-like retrieval pipeline while retaining its binary comparison and candidate search logic.

**Foundation models for histopathology retrieval.** Large-scale pathology foundation models have become general-purpose feature extractors for computational pathology. UNI demonstrated strong transfer performance across a wide range of pathology tasks [4], and GigaPath extended this direction to whole-slide representation learning using large-scale real-world pathology data [16]. More

recent multimodal WSI foundation models combine slide-level visual representations with additional sources of supervision or multimodal information [5]. Benchmarking studies also show that pretrained pathology foundation models can provide strong feature representations for weakly supervised computational pathology and other downstream analysis tasks [2, 13].

Using pathology foundation models for image retrieval is a natural extension of this trend. A recent study validated histopathology foundation models through WSI retrieval by replacing the original Yottixel descriptor with zero-shot foundation model embeddings [1]. The results showed that foundation model descriptors can improve retrieval quality compared with conventional CNN descriptors, confirming that representation choice is critical for scalable histology search. This finding is directly relevant to our work because it supports the hypothesis that modern pathology encoders can improve retrieval without task-specific supervised training.

Recent foundation-model retrieval studies have mainly focused on Yottixel-like WSI retrieval or direct comparison of slide-level representations. Our study instead focuses on a SISH-like retrieval setting. The CONCH- and Virchow-based variants retain the SISH components responsible for scalable candidate retrieval and ranking: VQ-VAE-based discrete keys, van Emde Boas tree search, binary Hamming-distance comparison, and uncertainty-based WSI aggregation. The modified part is visual descriptor extraction, together with the model-specific preprocessing and threshold selection required by the new descriptors. In this way, our study complements previous Yottixel-based validation work by evaluating pathology foundation model representations within a different scalable retrieval framework.

## 3 Datasets and Evaluation Protocol

Four datasets are used to evaluate slide-level and patch-level retrieval. **TCGA-300-WSI** is a subset of The Cancer Genome Atlas diagnostic WSI data. It contains 300 slides, with 50 slides sampled from each of six anatomical sites: brain, bronchus and lung, colon, heart/mediastinum/pleura, kidney, and stomach. The task is to retrieve slides from the same anatomical site.

**TCGA-300-Patch** is derived from the same TCGA-300 slides. Tissue regions are segmented automatically in HSV color space, empty glass background is removed, and non-overlapping $1024 \times 1024$ tissue fragments at $20\times$ magnification are extracted. Each patch inherits the anatomical label of its parent slide. The dataset consists of 35,335 patches distributed across the same six anatomical labels.

**Kather100k** (NCT-CRC-HE-100K) contains 100,000 $224 \times 224$ patches from colorectal cancer tissue and is labeled into nine tissue categories: adipose tissue, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium [10]. This benchmark tests whether retrieval preserves local tissue morphology rather than anatomical organ identity.

**WSSS4LUAD** contains 4,693 lung adenocarcinoma patches of variable size, assigned to three classes: normal tissue, stroma, and tumor [6]. The original patch sizes vary approximately from 150 to 300 pixels, so images are normalized before descriptor extraction. This benchmark provides a complementary patch-level setting for evaluating whether retrieval can identify diagnostically relevant tissue compartments within a single disease context.

Retrieval quality is evaluated using five top-$K$ metrics: majority vote (mMV@$K$), mean average precision (MAP@$K$), mean reciprocal rank (MRR@$K$), Recall@$K$, and Precision@$K$. Recall@$K$ is defined as 1 if at least one retrieved item with the query label appears among the top-$K$ results and 0 otherwise. Precision@$K$ is computed as the fraction of retrieved items in the top-$K$ list that share the query label. Metrics are reported for $K = 1, 3, 5, 10$. For all datasets, macro-averaged values are computed by first averaging query-level scores within each class and then averaging the resulting class-level scores over all classes.

### 3.1 Experimental Setup

A separate retrieval database was constructed for each method. The DenseNet121-based configuration was used as the original SISH baseline. The CONCH- and Virchow-based configurations were treated as SISH-like variants because they retained the original VQ-VAE-based indexing pipeline, integer latent-code representation, van Emde Boas tree candidate search, binary descriptor comparison, and WSI-level ranking logic, but replaced the DenseNet121 visual descriptor with foundation model representations. During retrieval, whole-slide images were represented by mosaic patches, candidate patches were retrieved through the unchanged index structure, and final filtering and ranking were performed using the Hamming distance between binarized feature representations.

To account for differences in descriptor dimensionality and empirical distance distributions, the Hamming-distance filtering threshold was defined separately for each descriptor type. Since the compared encoders produce binary descriptors of different lengths, the Hamming distance was normalized by descriptor length before thresholding. The original SISH threshold was retained for DenseNet121 descriptors, while model-specific thresholds were used for CONCH and Virchow. These thresholds were selected in preliminary experiments and then kept fixed across all evaluations for the corresponding descriptor type: 0.25 for DenseNet121, 0.125 for CONCH, and 0.33 for Virchow.

## 4 Method

### 4.1 Original SISH Components Retained in the SISH-like Pipeline

The retrieval framework follows the main design principles of SISH, but the proposed CONCH- and Virchow-based configurations should be understood as SISH-like rather than identical to the original method. The retained SISH components are mosaic-based WSI representation, VQ-VAE-based discrete key

generation, van Emde Boas tree candidate search, binary descriptor comparison using Hamming distance, and uncertainty-based aggregation for WSI-level retrieval. The modified component is visual descriptor extraction.

A WSI is first converted into a compact set of representative tissue patches. Background regions are removed by tissue segmentation, and informative tissue-containing regions are extracted as $1024 \times 1024$ patches. For WSI-level retrieval, these patches form a mosaic representation of the slide. For patch-level retrieval, the input image is treated as a single-patch mosaic.

Each patch is mapped into two complementary representations. The first is a discrete search key produced by a VQ-VAE encoder. The latent representation is quantized using a learned codebook and converted into an integer key, which is inserted into a van Emde Boas tree for efficient predecessor and successor search. The second representation is a compact binary visual descriptor. In the original SISH pipeline, this descriptor is obtained from DenseNet121 pretrained on ImageNet and then binarized.

At query time, the VQ-VAE key of each query patch is expanded by a fixed set of offsets. The tree returns candidate patches associated with nearby discrete keys, and these candidates are filtered and ranked using Hamming distance between binary descriptors. For patch-level retrieval, candidates are ranked directly by Hamming distance. For WSI-level retrieval, candidate lists from multiple query patches are aggregated using the uncertainty-based SISH ranking strategy. Candidate lists with lower retrieval uncertainty are ranked earlier, while lists with inconsistent retrieved label distributions or larger descriptor distances are filtered before the final slide ranking is produced.

This separation between discrete candidate search and binary descriptor comparison makes it possible to construct SISH-like variants. The indexing and candidate retrieval mechanism remain unchanged, while the DenseNet121 descriptor of the original SISH baseline is replaced by CONCH or Virchow.

### 4.2 Foundation Model Descriptors

In the SISH-like variants, the DenseNet121 descriptor used in the original SISH baseline is replaced by descriptors extracted from CONCH or Virchow. Both foundation models are used in a zero-shot regime, without fine-tuning on the target retrieval datasets. Their continuous embeddings are converted into binary descriptors using the same min–max binarization procedure as in the original SISH implementation, without modifying the binarization rule. This preserves compatibility with Hamming-distance filtering and keeps the descriptor comparison stage consistent with the baseline. For readability, we denote these two variants as SISH+CONCH and SISH+Virchow.

The compared encoders operate at different native input resolutions. CONCH uses $448 \times 448$ inputs, whereas Virchow uses $224 \times 224$ tiles. For this reason, $1024 \times 1024$ patches are processed using model-specific multi-crop strategies.

For CONCH, each $1024 \times 1024$ patch is decomposed into four $448 \times 448$ crops arranged in a $2 \times 2$ grid. The crop origins are $(0, 0)$, $(576, 0)$, $(0, 576)$, and $(576, 576)$, leaving a 128-pixel gap between crops along each axis. Each crop is

normalized using the CONCH preprocessing pipeline and encoded independently. The four crop-level embeddings are then averaged to obtain one descriptor for the original patch.

For Virchow, each $1024 \times 1024$ patch is decomposed into sixteen $224 \times 224$ crops arranged in a $4 \times 4$ grid. The stride is computed as

$$s_{\text{Virchow}} = \left\lfloor \frac{1024 - 224}{3} \right\rfloor = 266,$$

so crop origins along each axis are 0, 266, 532, and 798 pixels. Each crop is encoded independently. The class token is concatenated with the mean of patch tokens, producing a 2560-dimensional crop descriptor. Descriptors from all sixteen crops are averaged into a single patch-level representation.

For small-image datasets, the foundation-model variants do not use artificial upsampling to $1024 \times 1024$. Kather100k images are kept at their original $224 \times 224$ resolution, while WSSS4LUAD images are resized to $224 \times 224$. This avoids interpolation artifacts, reduces feature extraction cost, and keeps each foundation model close to its native input scale.

## 5 Experiments and Results

### 5.1 Quantitative Results

Table 1 reports macro-averaged retrieval metrics for Top-1, Top-3, Top-5, and Top-10. The results show a consistent advantage of pathology-specific foundation model representations over the original DenseNet121-based SISH descriptor. CONCH improves retrieval quality on all datasets, and Virchow achieves the strongest overall performance in nearly all evaluated settings.

In addition to the aggregate metrics in Table 1, Figures 1 and 2 show class-wise MAP@5. Figure 1 reports MAP@5 separately for the anatomical classes of TCGA-300-WSI, and Figure 2 summarizes MAP@5 for the classes of TCGA-300-Patch, Kather100k, and WSSS4LUAD. These plots show both the overall differences between SISH, SISH+CONCH, and SISH+Virchow and the classes for which retrieval remains more difficult.

**Table 1.** Macro-averaged retrieval metrics for Top-1, Top-3, Top-5, and Top-10. Higher is better for all metrics. Values are shown in percent. Bold indicates the best result among the three methods, and underlining indicates the second-best result.

| Dataset | Method | $K$ | mMV@$K$ | MAP@$K$ | MRR@$K$ | Recall@$K$ | Prec.@$K$ |
|---|---|---|---|---|---|---|---|
| TCGA-300 WSI | SISH | 1 | 61.0% | 61.0% | 61.0% | 61.0% | 61.0% |
| | | 3 | 61.7% | 55.9% | 65.7% | 72.0% | 59.1% |
| | | 5 | 61.7% | 49.4% | 66.9% | 77.3% | 54.4% |
| | | 10 | 61.3% | 39.0% | 68.2% | 87.7% | 47.9% |
| | SISH+CONCH | 1 | 65.7% | 65.7% | 65.7% | 65.7% | 65.7% |
| | | 3 | 65.0% | 60.3% | 69.4% | 74.3% | 62.6% |
| | | 5 | 66.3% | 57.2% | 70.7% | 80.0% | 61.3% |
| | | 10 | 67.0% | 50.3% | 72.4% | **92.0%** | 57.5% |
| | SISH+Virchow | 1 | **74.3%** | **74.3%** | **74.3%** | **74.3%** | **74.3%** |
| | | 3 | **73.7%** | **69.4%** | **76.5%** | **79.3%** | **70.8%** |
| | | 5 | **75.0%** | **64.6%** | **77.3%** | **82.7%** | **67.6%** |
| | | 10 | **72.3%** | **53.2%** | **78.3%** | 91.0% | **58.7%** |
| TCGA-300 Patch | SISH | 1 | 94.8% | 94.8% | 94.8% | 94.8% | 94.8% |
| | | 3 | 94.7% | 90.8% | 96.0% | 97.3% | 92.6% |
| | | 5 | 94.3% | 87.8% | 96.1% | 97.9% | 91.0% |
| | | 10 | 93.7% | 82.2% | 96.2% | 98.3% | 88.2% |
| | SISH+CONCH | 1 | 96.2% | 96.2% | 96.2% | 96.2% | 96.2% |
| | | 3 | 96.1% | 93.6% | 97.3% | 98.6% | 94.4% |
| | | 5 | 95.8% | 91.7% | 97.4% | 99.2% | 93.0% |
| | | 10 | 95.2% | 88.4% | 97.5% | 99.7% | 90.6% |
| | SISH+Virchow | 1 | **98.7%** | **98.7%** | **98.7%** | **98.7%** | **98.7%** |
| | | 3 | **98.4%** | **97.3%** | **99.0%** | **99.5%** | **97.5%** |
| | | 5 | **98.0%** | **95.9%** | **99.1%** | **99.7%** | **96.5%** |
| | | 10 | **97.2%** | **93.2%** | **99.1%** | **99.8%** | **94.2%** |
| Kather100k | SISH | 1 | 97.6% | 97.6% | 97.6% | 97.6% | 97.6% |
| | | 3 | 97.8% | 96.4% | 98.3% | 99.1% | 96.9% |
| | | 5 | 97.8% | 95.5% | 98.4% | 99.4% | 96.4% |
| | | 10 | 97.6% | 94.0% | 98.4% | 99.6% | 95.5% |
| | SISH+CONCH | 1 | 98.8% | 98.8% | 98.8% | 98.8% | 98.8% |
| | | 3 | 98.9% | 98.3% | 99.1% | 99.5% | 98.6% |
| | | 5 | 98.8% | 98.0% | 99.2% | 99.6% | 98.4% |
| | | 10 | 98.7% | 97.5% | 99.2% | 99.8% | 98.1% |
| | SISH+Virchow | 1 | **99.8%** | **99.8%** | **99.8%** | **99.8%** | **99.8%** |
| | | 3 | **99.7%** | **99.6%** | **99.8%** | **99.9%** | **99.6%** |
| | | 5 | **99.6%** | **99.4%** | **99.8%** | **99.9%** | **99.5%** |
| | | 10 | **99.5%** | **99.2%** | **99.8%** | **99.9%** | **99.3%** |
| WSSS4LUAD | SISH | 1 | 94.2% | 94.2% | 94.2% | 94.2% | 94.2% |
| | | 3 | 94.7% | 92.0% | 95.8% | 97.7% | 93.2% |
| | | 5 | 94.6% | 90.5% | 95.9% | 98.4% | 92.1% |
| | | 10 | 93.7% | 87.9% | 96.0% | 99.1% | 90.4% |
| | SISH+CONCH | 1 | 96.1% | 96.1% | 96.1% | 96.1% | 96.1% |
| | | 3 | 96.3% | 94.6% | 97.1% | 98.3% | 95.6% |
| | | 5 | 96.1% | 93.6% | 97.2% | 98.5% | 95.3% |
| | | 10 | 96.1% | 91.8% | 97.2% | 98.9% | 94.9% |
| | SISH+Virchow | 1 | **97.0%** | **97.0%** | **97.0%** | **97.0%** | **97.0%** |
| | | 3 | **97.1%** | **96.1%** | **97.7%** | **98.5%** | **96.5%** |
| | | 5 | **97.1%** | **95.4%** | **97.8%** | **99.0%** | **96.1%** |
| | | 10 | **97.0%** | **94.2%** | **97.8%** | **99.3%** | **95.3%** |

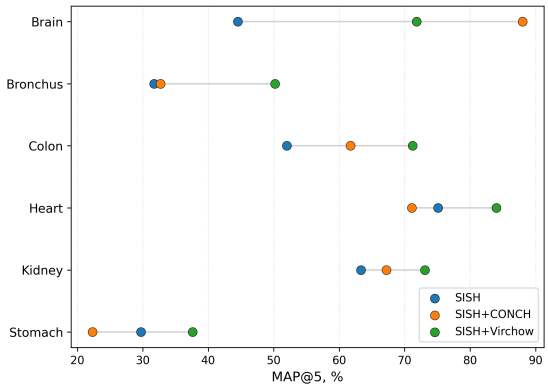


**Fig. 1.** Class-wise MAP@5 for whole-slide retrieval on the TCGA-300-WSI dataset.

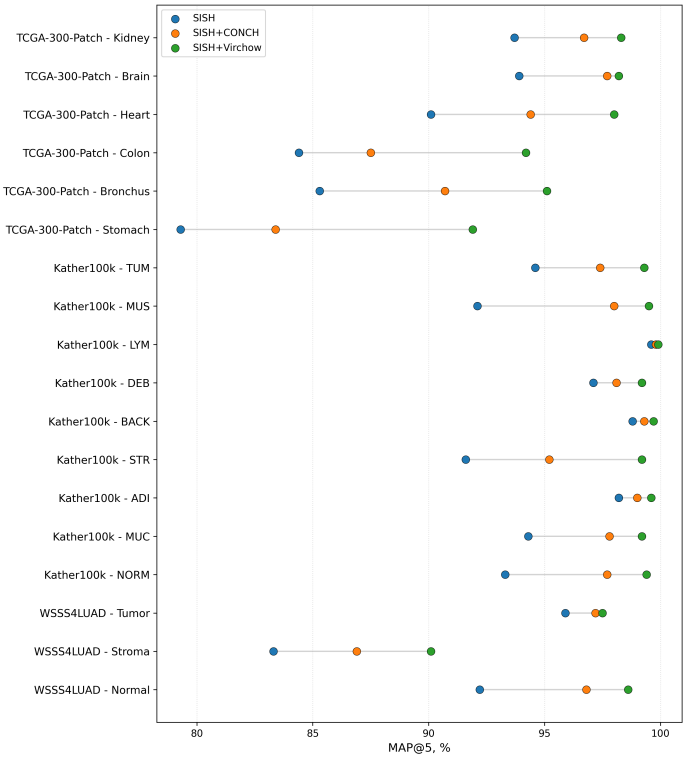


**Fig. 2.** Class-wise MAP@5 for patch-level retrieval on the TCGA-300-Patch, Kather100k, and WSSS4LUAD datasets.

We next analyze dataset-specific trends using the aggregate metrics in Table 1 and the class-wise MAP@5 results in Figures 1 and 2.

On TCGA-300-WSI, SISH+Virchow increases mMV@5 from 61.7% to 75.0% and MAP@5 from 49.4% to 64.6%. This is the most difficult setting because the system must retrieve slides from the same anatomical site using heterogeneous tissue regions from each WSI. The class-wise MAP@5 results show that Virchow improves all anatomical classes compared with the original SISH descriptor. The largest gains are observed for brain, bronchus/lung, colon, and stomach. Heart/mediastinum/pleura remains the easiest class for all methods. CONCH also improves macro-average MAP@5 from 49.4% to 57.2%, mainly through strong gains for brain, colon, and kidney, although its improvements are less uniform across anatomical classes.

On TCGA-300-Patch, all methods achieve higher retrieval quality than in the WSI-level setting, as expected for direct local patch comparison. Foundation model descriptors still give clear improvements. SISH+CONCH increases MAP@5 from 87.8% to 91.7%, and SISH+Virchow further increases it to 95.9%. The class-wise MAP@5 analysis shows consistent improvements across all six anatomical labels. The original SISH descriptor performs weakest on stomach, bronchus/lung, and colon patches, whereas Virchow substantially reduces these class-wise gaps.

Kather100k is close to saturation for all methods because many tissue categories are visually distinctive at $224 \times 224$ resolution. Even in this high-performance regime, SISH+CONCH improves MAP@5 from 95.5% to 98.0%, and SISH+Virchow further improves it to 99.4%. The class-wise MAP@5 results show that Virchow reaches near-perfect retrieval quality across all nine tissue categories. The largest remaining differences between methods are observed for morphologically heterogeneous categories such as stroma, smooth muscle, and normal mucosa.

WSSS4LUAD is smaller and contains variable-size lung cancer patches, but the same trend is observed. SISH+CONCH improves macro MAP@5 from 90.5% to 93.6%, while SISH+Virchow gives the best result with 95.4%. The class-wise MAP@5 values indicate that one tissue category remains more challenging than the others for all methods, whereas the remaining classes reach higher retrieval quality. Overall, the WSSS4LUAD results confirm that the benefit of foundation model descriptors is preserved even on a smaller dataset with variable patch sizes.

Across all datasets, the largest relative improvement appears in the TCGA-300-WSI setting, where retrieval requires aggregation of evidence from multiple heterogeneous regions. In contrast, Kather100k is already close to saturation, so the absolute gains are smaller but still consistent. These results indicate that pathology-specific representations improve both slide-level and patch-level retrieval while preserving the scalable retrieval framework.

### 5.2 Runtime Analysis

Table 2 reports the average CPU-only runtime per patch for database construction and retrieval.

**Table 2.** Average CPU-only runtime per patch for database construction and retrieval, reported as mean ± standard deviation.

| Setting | Method | Construction time, s | Retrieval time, s |
|---|---|---|---|
| 1024 × 1024 | SISH | $0.761 \pm 0.029$ | $0.188 \pm 0.078$ |
| | SISH+CONCH | $1.074 \pm 0.109$ | $0.270 \pm 0.001$ |
| | SISH+Virchow | $4.827 \pm 0.012$ | $0.365 \pm 0.014$ |
| 224 × 224 | SISH | $0.824 \pm 0.038$ | $0.098 \pm 0.168$ |
| | SISH+CONCH | $0.401 \pm 0.039$ | $0.085 \pm 0.026$ |
| | SISH+Virchow | $0.540 \pm 0.003$ | $0.170 \pm 0.017$ |

For 1024 × 1024 patches, SISH+CONCH increases construction time moderately compared with the original SISH descriptor, while SISH+Virchow is substantially more computationally expensive. This reflects the larger number of forward passes required by the Virchow preprocessing strategy for large patches. Nevertheless, this additional cost mainly affects the database construction stage, which can be performed offline and does not directly determine the latency of interactive retrieval. Retrieval remains below one second per patch for all methods, indicating that all evaluated variants remain practical for search-time use.

For the 224 × 224 setting, the foundation-model variants require less construction time than the original SISH descriptor, while retrieval time remains comparable across methods. SISH+CONCH is the fastest method in this setting. SISH+Virchow remains slower but still practical. Since all timing measurements were obtained on CPU, the reported values represent a conservative estimate of computational cost. Retrieval-time latency remains below one second per patch for all evaluated methods, suggesting that search itself can be performed efficiently even without dedicated GPU infrastructure. However, database construction is more computationally demanding, especially for the Virchow-based variant on 1024 × 1024 patches. Therefore, for large-scale archives, feature extraction and database construction are preferably performed on GPU.

### 5.3 Qualitative Retrieval Examples

Figures 3 and 4 show representative retrieval examples at the patch and whole-slide levels. In the patch-level example, the returned patches are visually similar to the query and are ranked only by increasing normalized descriptor distance. In contrast, the whole-slide example illustrates slide-level retrieval, where the final ranking is obtained after aggregating evidence from several query mosaics. For the retrieved whole-slide cases, both the retrieval bag entropy and the normalized descriptor distance are reported, because entropy is used only at the slide level to characterize the uncertainty of the aggregated retrieval evidence. These qualitative examples support the quantitative results by showing that SISH+Virchow retrieves visually coherent tissue patterns in both local patch retrieval and aggregated whole-slide retrieval scenarios.

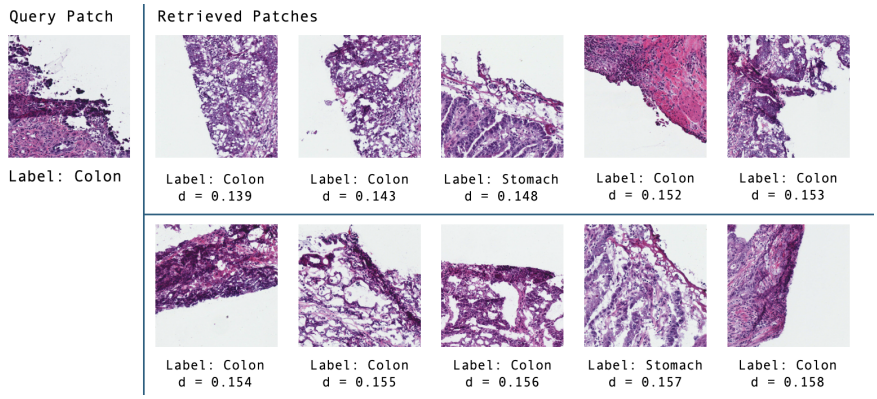


**Fig. 3.** Example SISH+Virchow retrieval query in the patch-to-patch mode for the TCGA-300-Patch dataset. The query is shown on the left, and the five retrieved similar patches are shown on the right, ranked by increasing Hamming distance $d$, normalized by the descriptor length.

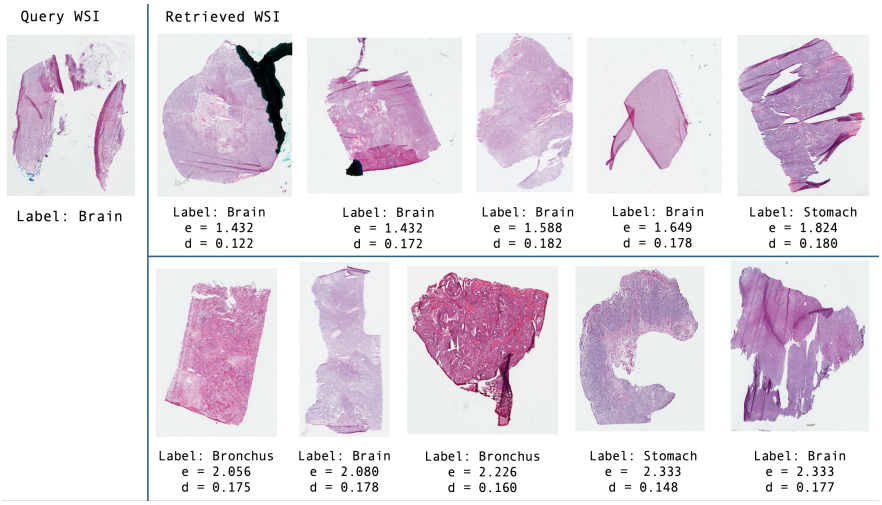


**Fig. 4.** Example SISH+Virchow retrieval query in the whole-slide mode for the TCGA-300-WSI dataset. The query slide is shown on the left, and the retrieved whole-slide cases are shown on the right. Each retrieved case is annotated with the retrieval bag entropy (e) and normalized Hamming distance (d). Results are ranked first by entropy and then by distance.

## 6 Discussion

The experiments lead to several observations. SISH-like variants improve scalable histology retrieval over the original DenseNet121-based SISH baseline while preserving the core candidate-search mechanism: VQ-VAE-based discrete indexing and van Emde Boas tree retrieval. The changes are limited to descriptor extraction, model-specific preprocessing, and descriptor-specific Hamming-distance thresholds. This suggests that most of the gain comes from stronger visual representations rather than from redesigning the indexing strategy.

Virchow achieves the best performance in most settings. Its advantage is likely related to large-scale pathology-specific pretraining, which helps the model encode tissue morphology more relevant to histological similarity than the ImageNet-based DenseNet121 descriptor. This effect is most visible in whole-slide retrieval, where the final result depends on matches aggregated across heterogeneous tissue regions.

The gain comes at a computational cost. SISH+Virchow is the most accurate variant, but also the most expensive, especially for $1024 \times 1024$ patches, where each patch is represented by sixteen $224 \times 224$ crops. CONCH gives a smaller but more efficient improvement, making it a practical choice when database construction time or computational resources are limited. These results also show that foundation-model descriptors require input adaptation rather than direct insertion into an existing retrieval pipeline.

The framework has several limitations. TCGA-300-WSI is small and balanced, which makes it useful for controlled comparison but insufficient for claims about clinical deployment readiness. The evaluation uses anatomical or tissue-type labels rather than expert similarity judgments, so it measures class-consistent retrieval rather than all diagnostically meaningful forms of similarity. In the patch-level evaluation, the query patch is excluded from the search results, but other patches from the same parent slide remain in the database. These results should therefore be interpreted as class-consistent local retrieval rather than fully slide-independent or patient-independent retrieval. For $1024 \times 1024$ patches, the sparse crop grids used by the CONCH- and Virchow-based variants preserve the native input resolution of each foundation model and reduce feature extraction cost, but they do not densely cover the full patch area. Denser or overlapping crop sampling may further improve descriptor quality at the cost of additional computation. Finally, all foundation models are used in a zero-shot setting. Task-specific calibration, learned crop aggregation, and learned binary hashing may further improve retrieval quality.

## 7 Conclusion

This study evaluates whether SISH-like retrieval variants with pathology foundation model descriptors can improve retrieval quality. The original SISH baseline is defined as the DenseNet121-based configuration with VQ-VAE discrete indexing, van Emde Boas tree candidate search, Hamming-distance filtering, and

uncertainty-based WSI ranking. In the proposed SISH-like variants, these indexing and ranking components are retained, while the visual descriptor is replaced with CONCH or Virchow and the preprocessing and distance thresholds are adapted to the selected encoder.

The updated preprocessing uses four $448 \times 448$ crops for CONCH and sixteen stride-based $224 \times 224$ crops for Virchow when processing $1024 \times 1024$ images. For datasets with $224 \times 224$ or otherwise small images, the foundation-model variants avoid the original SISH upsampling to $1024 \times 1024$ and encode images directly. This keeps the networks close to their native input scale. Experiments on TCGA-300-WSI, TCGA-300-Patch, Kather100k, and WSSS4LUAD demonstrate consistent improvements of the SISH-like foundation-model variants over the original DenseNet121-based SISH baseline across several top-$K$ retrieval settings, with Virchow achieving the strongest overall performance. These results indicate that foundation model descriptors can improve visual search in digital pathology while preserving the scalable retrieval mechanisms inherited from SISH.

**Acknowledgements** This work was supported by the Russian Science Foundation, project no. 26-11-00118.

## References

1. Alfasly, S., Alabtah, G., Hemati, S., Kalari, K.R., Garcia, J.J., Tizhoosh, H.: Validation of histopathology foundation models through whole slide image retrieval. Scientific reports **15**(1), 3990 (2025)
2. Campanella, G., Chen, S., Singh, M., Verma, R., Muehlstedt, S., Zeng, J., Stock, A., Croken, M., Veremis, B., Elmas, A., et al.: A clinical benchmark of public self-supervised pathology foundation models. Nature Communications **16**(1), 3640 (2025)
3. Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F.: Fast and scalable search of whole-slide images via self-supervised deep learning. Nature biomedical engineering **6**(12), 1420–1434 (2022)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
5. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: A multimodal whole-slide foundation model for pathology. Nature medicine pp. 1–13 (2025)
6. Han, C., Lin, J., Mai, J., Wang, Y., Zhang, Q., Zhao, B., Chen, X., Pan, X., Shi, Z., Xu, Z., et al.: Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Medical Image Analysis **80**, 102487 (2022)
7. Hegde, N., Hipp, J.D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C.J., Amin, M.B., Mermel, C.H., et al.: Similar image search for histopathology: Smily. NPJ digital medicine **2**(1), 56 (2019)
8. Kalra, S., Tizhoosh, H.R., Choi, C., Shah, S., Diamandis, P., Campbell, C.J., Pantanowitz, L.: Yottixel–an image search engine for large archives of histopathology whole slide images. Medical Image Analysis **65**, 101757 (2020)

9. Kalra, S., Tizhoosh, H.R., Shah, S., Choi, C., Damaskinos, S., Safarpoor, A., Shafiei, S., Babaie, M., Diamandis, P., Campbell, C.J., et al.: Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. NPJ digital medicine **3**(1), 31 (2020)
10. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019)
11. Lahr, I., Alfasly, S., Nejat, P., Khan, J., Kottom, L., Kumbhar, V., Alsaafin, A., Shafique, A., Hemati, S., Alabtah, G., et al.: Analysis and validation of image search engines in histopathology. IEEE Reviews in Biomedical Engineering **18**, 350–367 (2024)
12. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
13. Neidlinger, P., El Nahhas, O.S., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., et al.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature biomedical engineering pp. 1–11 (2025)
14. Tizhoosh, H.R., Pantanowitz, L.: On image search in histopathology. Journal of Pathology Informatics **15**, 100375 (2024)
15. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine **30**(10), 2924–2935 (2024)
16. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)