GECKO: Gigapixel Vision-Concept Contrastive Pretraining in Histopathology

Saarthak Kapse[1*], Pushpak Pati[3*], Srikar Yellapragada[1†], Srijan Das[2†], Rajarsi R. Gupta[1]
Joel Saltz[1], Dimitris Samaras[1], Prateek Prasanna[1]

[1]Stony Brook University, USA [2]UNC Charlotte, USA [3]Independent Researcher

Abstract

Pretraining a Multiple Instance Learning (MIL) aggregator enables the derivation of Whole Slide Image (WSI)-level embeddings from patch-level representations without supervision. While recent multimodal MIL pretraining approaches leveraging auxiliary modalities have demonstrated performance gains over unimodal WSI pretraining, the acquisition of these additional modalities necessitates extensive clinical profiling. This requirement increases costs and limits scalability in existing WSI datasets lacking such paired modalities. To address this, we propose Gigapixel Vision-Concept Knowledge Contrastive pretraining (GECKO), which aligns WSIs with a Concept Prior derived from the available WSIs. First, we derive an inherently interpretable concept prior by computing the similarity between each WSI patch and textual descriptions of predefined pathology concepts. GECKO then employs a dual-branch MIL network: one branch aggregates patch embeddings into a WSI-level deep embedding, while the other aggregates the concept prior into a corresponding WSI-level concept embedding. Both aggregated embeddings are aligned using a contrastive objective, thereby pretraining the entire dual-branch MIL model. Moreover, when auxiliary modalities such as transcriptomics data are available, GECKO seamlessly integrates them. Across five diverse tasks, GECKO consistently outperforms prior unimodal and multimodal pretraining approaches while also delivering clinically meaningful interpretability that bridges the gap between computational models and pathology expertise. Code is made available at github.com/bmi-imaginelab/GECKO

1. Introduction

There has been a surge in foundational models (FMs) in the histopathology domain [4, 35, 38], scaling both model sizes and the number of whole-slide images (WSIs), which has led to significant improvements in downstream task performance. However, since WSIs are gigapixel in nature, most of these models rely on decomposing a WSI into small patches and encoding them individually into an embedding space. Consequently, unlike in natural imaging, to use existing patch-level FMs for WSI-level analysis, supervisory signals are needed to aggregate patch embeddings into slide-level embeddings; a process that is prevalently achieved by supervised training of a Multiple Instance Learning (MIL)-based aggregator [12, 22].

To overcome this requirement of supervisory signals from human annotations, recent works [20, 38] have explored using a contrastive objective to pretrain the MIL aggregator on patch embeddings extracted from WSIs. This approach enables the derivation of slide-level embeddings in an unsupervised manner. However, this pretraining paradigm faces two distinct challenges that are

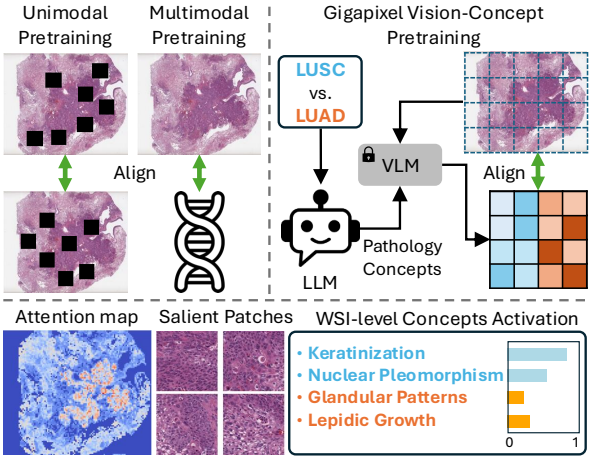

Figure 1. Unlike conventional unimodal or multimodal pretraining of WSI-level MIL aggregators, GECKO aligns a WSI with an interpretable Concept Prior derived from the WSI and task-relevant pathology concepts. Alongside downstream unsupervised and supervised performance benefits, GECKO provides WSI-level pathologist-friendly interpretable descriptors.

*Co-first authors, †Co-second authors

particularly critical to address in real-world applications. (1) the need for additional modalities to facilitate effective MIL pretraining at the WSI-level, and (2) the lack of pathologist-friendly interpretability in the pretrained model's aggregated WSI-level embeddings. Below, we describe these challenges in further detail.

Challenge 1: Need for Additional Modality. Recent studies [14, 15] have raised concerns about unimodal pretraining of MIL solely on WSI data, noting a tendency to overfit on staining artifacts instead of capturing biologically relevant features. This issue arises because the intra-slide invariance objective limits the diversity of the training signal, thereby amplifying the impact of staining variations between training and test distributions. In natural image analysis, integrating an auxiliary modality, such as textual information, into image pretraining serves as an effective strategy to enhance model robustness by mitigating overfitting to spurious visual correlations [26]. Along these lines, recent computational pathology studies have explored multimodal contrastive pretraining, wherein the H&E WSI modality (using MIL aggregator) is aligned with paired transcriptomics [14] or with WSI stained using IHC markers [15], demonstrating superior performance over unimodal pretraining. Nonetheless, their dependency on extra modalities for effective WSI-level pretraining is *limiting*, due to the relatively smaller size of paired modality datasets, the high cost of acquiring them, and the challenges associated with data harmonization and standardization across multiple sources.

Challenge 2: Limited Interpretability. Though both unimodal and multimodal pretraining methods for MIL aggregators bypass the need for WSI-level labels, during testing they yield only an aggregated WSI-level embedding and corresponding patch attention scores (as MIL is often attention-based [12]) for a WSI. Since these deep embeddings are inherently non-interpretable [17, 27], the interpretability of such models is *limited* to highlighting salient regions without revealing the key pathology concepts and insights that drive predictions. In contrast, when diagnosing a cancer patient's H&E WSI, a pathologist not only annotates the salient regions but also explains the diagnosis by identifying key visually discriminative pathology concepts, such as glandular or lepidic growth patterns in lung adenocarcinoma, or keratinization patterns in lung squamous cell carcinoma.

To this end, we ask: *Can we pretrain an effective WSI-level MIL aggregator without an auxiliary data modality, one that also provides WSI-level embeddings interpretable by pathologists?*

The answer is yes. In this paper, we propose the *first effective WSI-level pretraining solution*, that (1) does not require additional clinical profiling (e.g., RNA sequencing) for collecting paired modalities, and (2) provides expert-interpretable WSI-level embeddings (see Figure 1). To this end, we computationally derive a WSI-level Concept Prior and then employ a dual-branch MIL model to align the WSI with this Concept Prior.

First, we propose to derive a new pathology concept-driven prior from H&E WSI that is capable of providing task-specific discriminative signal for pretraining, while being inherently interpretable. This Concept Prior is a cosine similarity matrix computed in the vision-language embedding space. While the vision embeddings are obtained from WSI patches, the language embeddings are extracted from textual descriptions of predefined *visually discriminative pathology concepts* pertinent to each class. Each value in the concept prior matrix encodes the activation of a concept in the corresponding patch, making it inherently interpretable.

Second, we propose **G**igapix**E**l Vision-**C**oncept **K**nowledge C**O**ntrastive Pretraining (GECKO), a pretraining method which employs a dual-branch MIL network comprising a WSI-level deep encoding branch and a concept encoding branch. On one hand, the deep encoding branch aggregates the deep features (patch features) into a WSI-level deep embedding. On the other hand, the concept encoding branch aggregates the concept prior into a WSI-level concept embedding through a linear mapping to preserve its interpretability [17]. Finally, a contrastive objective [26] aligns these modalities, thereby pretraining both MIL branches. While GECKO is robust and does not require additional modalities, it allows for seamless integration of additional modalities (e.g., transcriptomics data), if available, during pretraining. Such multimodal pretraining can always benefit from our concept prior. *Importantly, GECKO's unsupervised pretraining only relies on task-specific class names to build the concept bank, which makes it completely slide-label free.*

The dual-branch MIL model pretrained with GECKO is evaluated in both unsupervised and supervised setup. We propose a pathologist-driven heuristic that leverages interpretable WSI-level concept embeddings to classify a slide by identifying the class with the most dominantly activated concepts; this demonstrates dramatic performance improvement over previous unsupervised baselines [23]. In the supervised setting, we further enhance performance by combining WSI-level deep embeddings with concept embeddings. Notably, GECKO surpasses existing unimodal pretraining methods when only the WSI modality is available. Moreover, when additional modalities like gene data are included, GECKO integrated with gene data exceeds the performance of existing multimodal pretraining approaches [5, 14, 30]. In summary, our main contributions are:

- We derive a task-specific, interpretable *concept prior* from H&E WSI, which enables effective slide-level

pretraining without requiring additional clinical profiling.

- Our dual-branch MIL pretrained with GECKO offers WSI-level deep embedding and interpretable concept embedding, with the latter enabling unsupervised WSI prediction via a pathologist-driven heuristic.
- GECKO allows seamless integration of auxiliary modalities, and consistently delivers state-of-the-art performance on five slide-level tasks while offering pathologist-friendly interpretability for its predictions.

2. Related Work

WSI-level pretraining: Recent advances in pretraining methods [3, 5, 14, 15, 20, 38] for MIL models in histopathology have greatly improved downstream tasks, especially in few-label settings. For example, Giga-SSL [20] uses contrastive learning on gigapixel slides with sparse convolutional layers, removing the need for manual annotations. TANGLE [14] incorporates gene expression profiles to enhance slide representation learning, achieving better few-shot performance than unimodal pretraining on WSI data alone. TANGLE also introduced an unimodal pretraining approach (Intra) that relies solely on WSI data; which is limited by an intra-slide invariance objective that may overfit to staining artifacts [14, 15], resulting in lower performance than multimodal pretraining. MEDELEINE [15] employs different slide stainings of the same tissue with a global-local cross-stain objective to learn comprehensive morphological features, showing the superiority of multimodal pretraining. TITAN [5] employs large patch-level pretraining (8k $\times$ 8k) using DINO [2, 25], followed by contrastive pretraining with CoCa [41] on a large corpus of paired pathology reports and synthetic captions. PANTHER [30] uses Gaussian mixture models to summarize patches into morphological prototypes. Though it is not strictly a pretraining approach, it facilitates unsupervised patch-to-slide level aggregation. Our method, pretrained with only WSI data, significantly outperforms Intra and matches TANGLE. Moreover, when gene data is integrated, our method significantly surpasses previous multimodal approaches.

Interpretability methods: Previously, SI-MIL [17] introduced the first self-interpretable [1, 6, 18, 39] method in histopathology that highlights salient regions in a WSI and also quantifies each handcrafted feature's contribution to predictions. Recently, Concept MIL [31] addressed the limitations of handcrafted features by replacing them with vision-language based concept activation scores, thereby offering more pathologist-friendly interpretability. However, these self-interpretable methods in this domain are limited to supervised setting. In our MIL model, which we aim to pretrain, we employ a linear mapping aggregator motivated by SI-MIL to encode the derived concept prior, thereby preserving interpretability, and then align this linearly aggregated concept prior with the gigapixel WSI. While other works [13, 16, 36, 43] have incorporated concepts to enhance pretraining in histopathology, they focus on patch-level pretraining. Thus, our method marks the first approach to leverage concepts for scaling to WSI-level pretraining.

3. Proposed Method

This section introduces GECKO, our gigapixel image-concept knowledge pre-training strategy, illustrated in Fig. 2. GECKO pretrains a dual-branch MIL consisting of WSI-level deep-encoding and concept-encoding branches, by contrastively aligning them. First, we detail the extraction of the concept prior from a WSI in Sec 3.1. Then, we elaborate on the deep-encoding and concept-encoding branches in Sections Sec. 3.2 and Sec. 3.3, respectively. Finally, we outline the GECKO pre-training and inference strategy in Sec. 3.4.

3.1. Concept Prior Extraction

The concept prior for a WSI is extracted by using well-established task-specific pathology knowledge, ensuring that the resulting derived prior is both pathologist interpretable and provides appropriate discriminative signals for the task at hand. Note that, we focus on task-specific concepts instead of WSI-specific labels.

First, we define the task-specific lexicon by leveraging a LLM, *e.g.*, GPT-4. We define dedicated prompts to query the LLM, generating textual descriptions of *visually discriminative pathology concepts* pertinent to each class $l \in L$ in the downstream task [42]. These concept descriptions across L classes are encoded into embeddings using the text encoder of a pre-trained VLM [9, 11, 24, 29, 37], denoted as $T \in \mathbb{R}^{C \times D}$, where C is the total number of concepts and D is the embedding dimension. Next, each WSI is divided into N patches and each patch is encoded into a feature vector $f_i \in \mathbb{R}^D$ using the vision encoder from the same pre-trained VLM, resulting in $F \in \mathbb{R}^{N \times D}$. Given that both the text and image embeddings reside in a shared embedding space, we compute the pairwise cosine similarity between patch features F and concept description embeddings T, yielding a concept matrix $M \in \mathbb{R}^{N \times C}$. M serves as the concept prior for GECKO. Each element in M quantifies the activation level of a specific concept within a patch, providing an interpretable mapping in the language of pathology. Furthermore, since we select a set of distinct concepts for each class, the concept prior matrix M inherently encodes a task-specific discriminative signal. In particular, the salient regions of a WSI belonging to a given class exhibit higher activation levels for the concepts associated with that class. Note

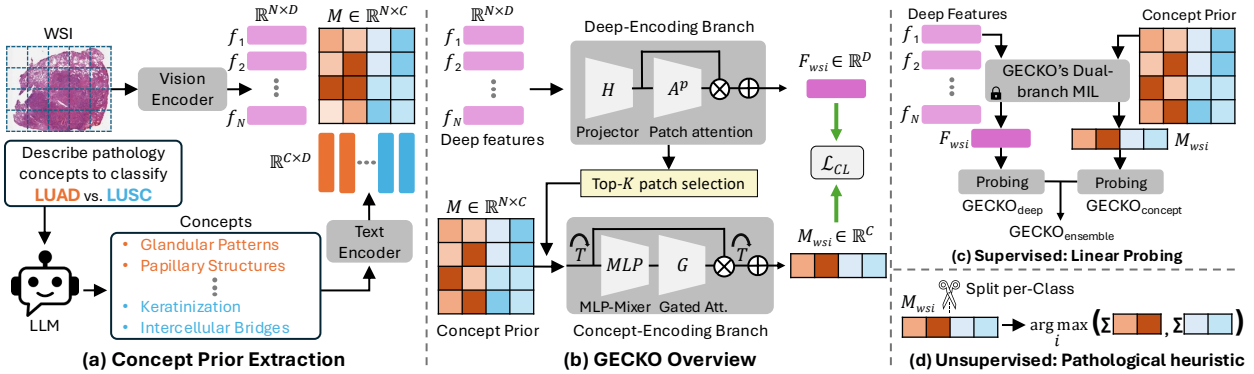

Figure 2. Overview of GECKO: (a) We start by extracting an interpretable, task-relevant Concept Prior from a WSI using a Large Language Model (LLM) and a Vision Language Model (VLM). (b) Next, we pretrain a dual-branch MIL by contrastively aligning WSI-level deep and concept embeddings. (c) These embeddings can be used for supervised learning via linear probing. (d) Additionally, the concept embedding can be directly used for unsupervised learning using a pathologist-driven heuristic.

that, although downstream classes may share some concepts, we only select those that are most visually unique to each class to provide a clear discriminative signal for pretraining.

To highlight, this concept prior extraction can be fully automated using LLMs, unlike other clinical modalities such as gene sequencing, pathology reports, and staining markers which require significant effort.

3.2. WSI-level Deep-Encoding Branch

The WSI-level deep-embedding is learned via a MIL model. The patch-level deep features $f_i \in F$ from a WSI, in Sec. 3.1, are considered as a set of instances and are aggregated via the MIL [12]. As shown in Fig. 2, the MIL first projects the set of f_i on to a feature space using a projector $H(\cdot)$, then employs a patch attention module $A^p(\cdot)$ to compute attention α for each patch, given as,

$$\tilde{f}_i = H(f_i); \quad \alpha_i = A^p(\tilde{f}_i); \quad i \in \{1, 2, \ldots, N\} \quad (1)$$

Here, $A^p(\cdot)$ is a parameterized module with a softmax activation. Next, $\{\tilde{f}_i\}$ is attention-pooled using $\{\alpha_i\}$ to yield the WSI-level deep-embedding $F_{wsi} \in \mathbb{R}^D$ as,

$$F_{wsi} = \sum_{i=1}^{N} \alpha_i \cdot \tilde{f}_i \quad (2)$$

3.3. WSI-level Concept-Encoding Branch

The concept embedding is derived by aggregating concept prior from the most salient patches. It involves identifying key patches, highlighting their salient concepts, and aggregating across the patches to compute the final concept embedding.

The top K salient patches are identified using patch attention scores α_i (eq. 1), employing the differentiable **perturbed Top-K** operator as in [17, 32] for optimal selection. The concept prior M is truncated to these K patches, rendering $\tilde{M} \in \mathbb{R}^{K \times C}$ for further processing.

We highlight the salient concepts by computing concept attention values from $\tilde{M}$. The transposed matrix $\tilde{M}^T$ is processed by an feature attention module, which uses **MLP-Mixer** [33] layers to contextualize spatial information across the K patches and concept activation across the C concepts. A gated attention network $G(\cdot)$ with sigmoid activation is then applied to each row of $\tilde{M}^T$, learning an attention score β_j for each concept activation C_j, as in Eqn. 3. β_j are used to linearly transform $\tilde{M}$ into $\hat{M}$, as in Eqn. 4, emphasizing salient concepts in a data-driven manner. To note, despite the non-linear operations to compute β_j, $\hat{M}$ is only linearly scaled, ensuring the inherent interpretability of $\hat{M}$. Finally, the WSI-level concept embedding $M_{wsi} \in \mathbb{R}^C$ is obtained by average pooling over $\hat{M}$, as in Eqn. 5.

$$\beta_j = G\Big(\text{MLP-mixer}\big(\tilde{M}^T\big)\Big); \quad j \in \{1, ..., C\} \quad (3)$$

$$\hat{M}_{ij} = \beta_j \times \tilde{M}_{ij}; \quad i \in \{1, ..., K\}, \; j \in \{1, ..., C\} \quad (4)$$

$$M_{wsi} = \frac{1}{K} \sum_{i=1}^{K} \hat{M}_i \quad (5)$$

M_{wsi} captures the activation of each concept at the WSI-level, preserving its pathologist-friendly interpretability.

3.4. GECKO: Pre-training and Inference

Pre-training: GECKO follows Vision Concept Model (VCM) to align the WSI-level embeddings $F_{wsi} \in \mathbb{R}^D$ and $M_{wsi} \in \mathbb{R}^C$ from the deep- and concept-encoding branches, respectively, in a shared latent space. To this end, we optimize the symmetric cross-modal CLIP [26] contrastive learning loss $\mathcal{L}_{CL}$:

$$\mathcal{L}_{CL}(a, b) = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(a^i, b^i)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(a^i, b^j)/\tau\right)} \quad (6)$$

Here, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between embeddings, B is batch size, and τ is a temperature hyper-

parameter. Note, a linear layer projects F_{wsi} to match M_{wsi} dimensions. The overall loss is computed as,

$$\mathcal{L} = \frac{1}{2}\Big(\mathcal{L}_{CL}(F_{wsi}, M_{wsi}) + \mathcal{L}_{CL}(M_{wsi}, F_{wsi})\Big) \quad (7)$$

Through this pretraining, the GECKO-pretrained dual-branch MIL is optimized to accurately identify discriminative *patches* and *concepts* within a WSI, effectively contrasting them with other WSIs in the batch. To enhance contrastive pretraining, we use false negative elimination [10] with a keep ratio of r_{keep}. This process excludes a fraction (1 - r_{keep}) of highly similar WSI-embeddings from the contrastive loss, preventing the comparison of similar WSIs.

Inference: After pretraining, the dual-branch MIL provides two embeddings: $F_{wsi} \in \mathbb{R}^D$ from the Deep-Encoding branch, and $M_{wsi} \in \mathbb{R}^C$ from the Concept-Encoding branch. M_{wsi}, being inherently interpretable, can be used for unsupervised prediction; and both F_{wsi} and M_{wsi} can be used for supervised-prediction.

Unsupervised prediction: Let $I_l \subset I$ represent the indices of distinct concepts in C corresponding to class $l \in L$ from the downstream task, where $I = 0, 1, ..., C$. The probability of a WSI belonging to class l is,

$$P(l) = \frac{\sum_{j \in I_l} M_{wsi,j}}{\sum_{k \in I} M_{wsi,k}} \quad (8)$$

This pathologist-driven heuristic ensures that if the aggregated WSI-level concepts for a class l are predominantly activated, the probability $P(l)$ increases, similar to a pathological diagnosis. We refer to our model GECKO-Zero in this setting, as it requires no WSI-level labels during training.

Supervised prediction: Both F_{wsi} and M_{wsi} can be used for labeled classification via linear probing in both few-label and full-supervision setups. In our experiments, we evaluate them in both setups and also report results for an "ensemble" setup, where we average the predicted probabilities from using both embeddings.

4. Experiments and Results

In this section, we provide the evaluation datasets and implementation details of the proposed GECKO-pretrained dual-branch MIL model across multiple WSI classification tasks. We evaluate GECKO under unsupervised settings (with zero labels) as well as under supervised settings when few or all labels are utilized. The supervised setup also includes adapting GECKO to multi-modal scenarios (incorporating gene modality).

4.1. Datasets and Implementation Details

Datasets: We evaluate our framework on three public datasets: TCGA-Lung, TCGA-STAD, and TCGA-BRCA, using the data splits from ConcepPath [42]. Class details and WSI distributions are provided in Supplementary Table 7. In TCGA-STAD, we define three binary classification tasks: EBV vs. Others, MSI vs. Others, and EBV + MSI vs. Others. TCGA-Lung and TCGA-BRCA are used for binary and ternary classifications, respectively. To evaluate generalization on out-of-domain data, we use TCGA-Lung pretrained models for downstream analysis on the CPTAC-Lung dataset. All results are reported for 5-fold cross (mean, standard deviation) on TCGA datasets, with same splits provided by ConcepPath [42]. Note that, for pretraining GECKO and baseline methods Intra and TANGLE, we only use training split for pretraining and conduct this pretraining separately for each fold to avoid any data leakage between train and test split.

Defining Concepts: For each task, we use a LLM with the concept generation technique from ConcepPath [42] to produce concepts per class. In addition, we prompt the LLM to identify the 10 most visually distinct concepts per class from the generated concepts, producing $C = 20$ for subtyping in TCGA-Lung and the three tasks in TCGA-STAD, and $C = 30$ for the three-class TCGA-BRCA. A few example concepts in Lung cancer are provided in Fig. 2, and the full list of concepts for all datasets is available in Supplementary Tables 8- 12.

Patch and Feature Extraction: We use the vision and text encoders from the CONCH [24] model to extract deep features F and the concept prior M. Patches of size 448×448 pixels are extracted at $20\times$ magnification (0.5μm/px). Note that, unless specified otherwise, the CONCH model is used for deep feature extraction in our framework as well as in the baselines. We also perform ablation studies with different patch feature extractors for the deep encoding branch.

MIL Setting: For the deep-encoding branch, we use ABMIL [12] with the architecture from TANGLE [14], featuring a 2-layer MLP projector $H(\cdot)$ with 512 hidden units and a gated-attention network $A^p(\cdot)$, also a 2-layer MLP with 512 hidden units and Sigmoid and Tanh activations. The concept-encoding branch employs 4 layers of MLP-Mixer and a gated-attention network $A^f(\cdot)$ with the same configuration, inspired by SI-MIL [17]. We set $K = 10$ by default. Since the concept-encoding branch input is $\tilde{M} \in \mathbb{R}^{K \times C}$, with $K = 10$ and $C \in \{20, 30\}$, it is much more lightweight than the deep-encoding branch, which uses a 512-dimensional input from the CONCH-extracted features. Further implementation details can be found in Supplementary 9.

4.2. Unsupervised Evaluation Setting

Baselines: We benchmark our unsupervised predictive performance, *i.e.* with **zero** WSI-level annotations, against MI-zero [23] and ConcepPath-Zero, an adapted version of ConcepPath [42]. To enable unsupervised

Methods	Interpretable (patch level-feature level)	LUAD vs. LUSC (530 vs. 512)	EBV+MSI vs. Others (70 vs. 199)	MSI vs. Others (44 vs. 225)	EBV vs. Others (26 vs. 243)	HER2 pos vs. neg vs. equi (164 vs. 583 vs. 186)
MI-Zero [23]	✓- ✗	**96.6** ± 1.4	61.9 ± 6.1	42.3 ± 5.3	74.3 ± 11.9	32.2 ± 1.3
ConcepPath-Zero [42]	✗- ✗	91.0 ± 3.3	74.2 ± 7.2	73.4 ± 7.9	68.3 ± 11.2	37.5 ± 1.1
GECKO-Zero	✓- ✓	95.0 ± 1.7	**83.4** ± 4.9	**77.1** ± 11.4	**82.5** ± 6.3	**60.6** ± 2.4
ConcepPath [42] (supervised)	✗- ✗	98.0 ± 0.7	84.0 ± 5.5	85.0 ± 5.0	90.1 ± 4.0	78.4 ± 1.2

Table 1. Unsupervised classification analysis on TCGA datasets. Table shows AUC from methods using no labels for supervision across tasks. Last row provides upper bound using ConcepPath [42] under full supervision setting.

prediction, we remove the learnable data-driven prompts and the trainable slide adapter from ConcepPath

Results: As observed in Table 1, GECKO-Zero outperforms MI-Zero by a significant 10–30% margin across different tasks, except for the lung subtyping task where performance is comparable. Unlike MI-Zero, which directly uses slide-level classification prompts for cosine-similarity with patch-level features followed by pooling aggregation, ConcepPath-Zero decomposes slide-level prompts into concept prompts, often visible at patch-level, and then aggregates based on the activated concepts. This strategy yields considerable improvements for ConcepPath-Zero over MI-Zero, where relying on activation of slide-level prompts in each patch can introduce noisy aggregation. Notably, GECKO-Zero outperforms ConcepPath-Zero across all tasks for its ability of identifying salient regions and concepts in a WSI via contrastive pretraining and aggregating concepts across the salient patches. To highlight, as our predictions are from WSI-level concept embeddings, the predictions can be directly interpreted by pathologists, and if required, can be corrected via test-time interventions [19].

4.3. Supervised Evaluation Setting

Baselines: Here we benchmark our method against two types of WSI-level pretraining methods: (1) methods using only WSIs, and (2) multimodal pretraining methods. Specifically, our baselines include unimodal pretraining method– Intra established by [14, 15], and TANGLE [14] a multimodal pretraining method incorporating gene data. We extract WSI-level embeddings from the pretrained MIL aggregators and train linear classifiers using labels from the training dataset. We evaluate two scenarios: (1) few-k labels per class, and (2) all training labels. The baseline methods are compared against ours (linear classifiers trained on WSI-level deep embeddings (referred to as $\text{GECKO}_{\text{deep}}$) and interpretable concept embeddings (referred to as $\text{GECKO}_{\text{concept}}$) provided by GECKO-pretrained dual-branch MIL model). We also average the predicted probabilities from $\text{GECKO}_{\text{deep}}$ and $\text{GECKO}_{\text{concept}}$ to define $\text{GECKO}_{\text{ensemble}}$ predictions.

To address variability in the few-label setting, we perform 10 repetitions with shuffled samples from the training set for each k. For in-domain tasks, which involve pretraining on the train split, linear probing on few-k samples per class from the training split, and testing on the test split across all TCGA datasets, this process is executed across all 5-folds (thus resulting in training 50 linear probing classifiers per dataset for each value of k for our and baseline pretraining methods). We then report the average performance on the corresponding test sets over all repetitions and folds. For out-of-domain generalization, we pretrain on the entire TCGA-Lung dataset, conduct linear probing on few-k samples per class from the CPTAC-Lung dataset, and test on all remaining samples in CPTAC-Lung. We perform 10 repetitions by sampling different WSIs for each value of k and testing on all other WSIs in the dataset. The mean and standard deviation of the results are reported in the Table 5 (in Supplementary). We observe that when only WSI modality is available, GECKO significantly outperforms the baseline method Intra, and when gene modality is included, GECKO performs on par with TANGLE.

Few-Labels Setting: As shown in Figure 3, in the unimodal setting with only WSI data (indicated by dashed lines), linear probing on our interpretable WSI-level concept embeddings (dim C) surpasses linear probing on Intra pretrained aggregator embeddings (dim $D >> C$) in several tasks, while also offering interpretable predictions. Moreover, $\text{GECKO}_{\text{ensemble}}$ consistently outperforms the Intra pretraining across all tasks.

In the multimodal setting (indicated by solid lines), $\text{GECKO}_{\text{ensemble}}$ uses TANGLE's gene encoding branch to pretrain with the gene modality alongside WSIs and concept prior. $\text{GECKO}_{\text{ensemble}}$ consistently outperforms the state-of-the-art TANGLE across all tasks. Interestingly, when GECKO is pretrained with the gene modality, the performance of linear probing with the interpretable concept embedding M_{wsi} also improves significantly compared to pretraining with only WSIs. Results for the EBV vs. Others and BRCA datasets in the few-label setting are provided in Supplementary Figure 5.

Full-Supervision Setting: Table 2 presents the results of using all training labels to supervise a linear classifier in a 5-fold cross-validation setting, comparing embeddings from our GECKO-pretrained model against the Intra and TANGLE baseline methods. Even with full supervision, our embeddings outperform Intra in the WSI-only scenario and TANGLE when gene data is available. Remarkably, our dual-branch MIL aggregator, pretrained with GECKO without the additional gene modality, matches or exceeds TANGLE's performance, which relies on additional gene modality data. When the gene

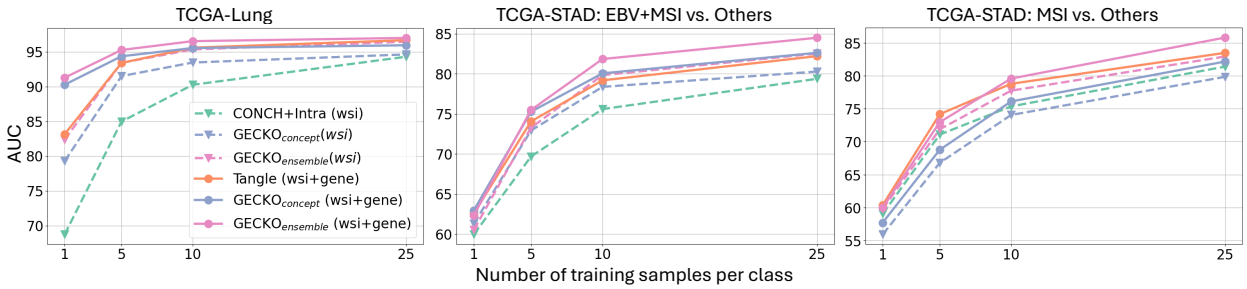

Figure 3. Few-labels (in-domain) classification analysis. The AUC results are obtained through linear probing. Dashed lines indicate pretraining using only WSI data, while solid lines represent multimodal pretraining with additional transcriptomics data. CONCH is utilized for extracting deep features for image patches.

	Methods	Embedding	Interpretable (patch level-feature level)	LUAD vs. LUSC (530 vs. 512)	EBV+MSI vs. Others (70 vs. 199)	MSI vs. Others (44 vs. 225)	EBV vs. Others (26 vs. 243)	HER2 pos vs. neg vs. equi (164 vs. 583 vs. 186)
WSI only	*ConcepPath [42]	concept	✗-✗	**98.0 ± 0.7**	84.0 ± 5.5	85.0 ± 5.0	**90.1 ± 4.0**	**78.4 ± 1.2**
	Intra [14]	deep	✓-✗	97.5 ± 0.6	83.5 ± 8.3	83.9 ± 7.0	85.9 ± 5.2	74.2 ± 1.9
	GECKO	deep	✓-✗	97.5 ± 0.3	85.3 ± 8.8	83.8 ± 4.8	84.3 ± 5.4	75.6 ± 1.5
		concept	✓-✓	96.3 ± 1.2	82.8 ± 6.1	84.4 ± 10.9	83.8 ± 6.1	75.9 ± 1.6
		ensemble	✓-✗	97.6 ± 0.6	**86.4 ± 8.4**	**86.5 ± 7.8**	86.6 ± 6.0	77.2 ± 1.6
WSI + Gene	TANGLE [14]	deep	✓-✗	**97.9 ± 0.5**	85.4 ± 8.0	86.6 ± 5.0	86.8 ± 2.7	75.3 ± 1.3
	GECKO	deep	✓-✗	97.8 ± 0.4	86.1 ± 7.2	88.1 ± 4.3	86.6 ± 3.9	76.4 ± 2.0
		concept	✓-✓	97.3 ± 0.8	85.6 ± 6.3	86.6 ± 6.5	82.7 ± 7.5	76.3 ± 2.3
		ensemble	✓-✗	**97.9 ± 0.5**	**87.1 ± 7.0**	**89.4 ± 5.3**	**87.4 ± 3.7**	**78.4 ± 1.8**

Table 2. Full supervision (in-domain) classification analysis. Table compares AUC from various methods across multiple classification tasks. All results are with linear probing, except for *ConcepPath; In *ConcepPath, all parameters (>100K) are optimized with full supervision compared to just (∼1K) parameters in linear probing. CONCH is utilized for extracting deep features.

modality is included in GECKO, we observe further significant improvements across all tasks, except for the Lung dataset, where performance remains comparable.

Moreover, a linear classifier trained on our WSI-level embeddings using only the WSI modality outperforms ConcepPath in 2 out of 5 tasks, despite its lightweight nature. To note, all parameters of ConcepPath are fully optimized with label supervision, unlike ours, where we only train the linear classifier on slide-embeddings. When pretrained with the gene modality, our model significantly outperforms ConcepPath in 2 tasks and performs on par in 2 others. Although further fine-tuning with label supervision could enhance both our and baseline pretrained MIL backbones [15, 34], our primary goal is to establish a robust WSI pretraining method that consistently outperforms both unimodal and multimodal pretraining approaches under few-labels and even full supervision conditions.

4.4. Interpretability analysis

In Figure 1, we present the patch attention heatmap and the top four attended patches for a query WSI using our GECKO-pretrained dual-branch MIL. Utilizing the interpretable WSI-level concept embedding in an unsupervised setting, we highlight the two concepts with the highest and lowest aggregated activations and their activation strengths. Unlike previous MIL pretraining methods [14, 15, 20] that could only provide patch attention maps, our approach uniquely identifies the pathology concepts driving predictions. In a user study, a pathologist reviewed WSIs with saliency maps and ranked concepts, confirming that the most activated concepts matched the WSIs' predictions as LUSC/LUAD, consistent with the pathological prior. Additional examples are in Supplementary Sec. 11. Table 3 presents a quantitative analysis of the accuracy of the salient concepts identified by our pretrained model in both unsupervised and supervised setting.

Unsupervised setting: Using M_{wsi}, we extract the top-j most activated concepts from a WSI-level concept embedding and evaluate their accuracy based on overlap with the respective ground truth class concepts. The assessment is performed across 5-folds on the test split, reporting mean and standard deviation. Notably, the top-1 concept shows 81.4% accuracy in Lung cancer and 54.0% in the challenging MSI vs. Others task. Analysis across various j reveals task complexity patterns: Lung cancer subtyping is relatively easy, making top-1 concept identification easy, but increasing j mixes concepts across classes, reducing accuracy. Conversely, for MSI vs. Others, identifying the top-1 concept is challenging, but increasing j improves accuracy by capturing correct class concepts. Importantly, concept selection excludes activation strengths, which explains why MSI vs. Others shows lower concept selection accuracy but higher prediction AUC in Table 1. This capability sup-

		Unsupervised	Fully-Supervised
LUAD vs. LUSC	$j=1$	81.4 ± 2.5	99.9 ± 0.2
	$j=3$	75.8 ± 2.3	98.1 ± 1.4
	$j=5$	71.6 ± 1.2	95.1 ± 3.5
MSI vs. Others	$j=1$	54.0 ± 1.2	83.3 ± 16.9
	$j=3$	61.1 ± 3.3	76.1 ± 7.0
	$j=5$	61.0 ± 3.9	80.5 ± 3.3

Table 3. Accuracy of salient concepts identified by our pretrained model. Here, j denotes the correctness of the top-j concepts, determined by their alignment with the concepts associated with the ground-truth class.

ports biomarker discovery in clinical settings with unlabeled WSI data, enabling hypothesis testing by defining relevant concepts, pretraining a dual-branch MIL with GECKO, and analyzing salient concepts at the slide or class level using the interpretable M_{wsi}. GECKO's interpretable-by-design architecture fundamentally sets it apart from existing interpretability methods.

Supervised setting: We multiply the interpretable concept embedding at the WSI-level M_{wsi}, with the weights w derived from the fully supervised trained linear classifier. When evaluating a test WSI, the classifier's prediction determines the selection of concepts. If the prediction is class 0, we choose the concepts with the lowest values in $w \times M_{wsi}$. If the prediction is class 1, we choose the concepts with the highest values in $w \times M_{wsi}$. The reasoning for this is that a binary logistic regression classifier learns weights that push features related to class 0 towards lower values and features related to class 1 towards higher values, optimizing the sigmoid-based classification. Notably, we observe that in supervised setting, our method can identify top-1 concept with 99.9% accuracy in Lung cancer and 83.3% for the challenging MSI vs. Others task. The improved concept selection, compared to unsupervised setting, across different j can be attributed to the strong label supervision.

4.5. Ablations

Comparison with additional WSI-level encoding methods: Thus far in this study, we have extensively compared GECKO against Intra in unimodal setting and TANGLE when gene data is available, consistently using the CONCH feature extractor. Table 4 further compares GECKO with PANTHER [30] and TITAN [5]. Since TITAN utilizes the CONCH v1.5 encoder, we also base PANTHER on CONCH v1.5 embeddings and use CONCH v1.5 to extract deep features for GECKO. In addition, we evaluate WSI-level embeddings from the unimodal Giga-SSL [20, 21], training a linear classifier. Giga-SSL provides WSI-level embeddings based on pretraining on three different patch-level foundation models, Phikon [7], Gigapath [38], and H-Optimus-0 [28]. Our method significantly outperforms PANTHER and

	Methods	Embedding	LUAD vs. LUSC $k=10$	LUAD vs. LUSC $k=25$	EBV+MSI vs. Others $k=10$	EBV+MSI vs. Others $k=25$
WSI only	PANTHER	deep	91.2 ± 0.7	95.2 ± 0.8	78.5 ± 2.8	84.0 ± 5.3
	Giga-SSL	Phikon	84.7 ± 1.7	89.5 ± 1.7	74.4 ± 5.6	79.4 ± 6.9
		Gigapath	92.4 ± 1.0	94.7 ± 1.0	78.0 ± 5.1	82.7 ± 6.2
		H-Optimus	92.8 ± 1.1	94.8 ± 1.0	77.5 ± 4.0	82.0 ± 5.4
	GECKO	concept	95.4 ± 0.7	95.6 ± 0.7	75.9 ± 1.9	79.8 ± 3.8
		ensemble	$\mathbf{96.4 \pm 0.6}$	$\mathbf{96.7 \pm 0.6}$	$\mathbf{82.1 \pm 4.7}$	$\mathbf{84.6 \pm 5.5}$
Multi-modal	TITAN	deep	$\mathbf{97.5 \pm 0.9}$	$\mathbf{97.7 \pm 0.8}$	78.7 ± 3.6	84.7 ± 4.6
	GECKO	concept	96.2 ± 1.3	96.5 ± 1.0	79.3 ± 5.1	81.7 ± 5.7
		ensemble	97.0 ± 1.1	97.2 ± 0.8	$\mathbf{84.4 \pm 5.4}$	$\mathbf{86.0 \pm 5.8}$

Table 4. Comparison with additional WSI-level encoding methods. All AUCs reported are with linear probing. CONCH v1.5 is used for extracting deep features.

Giga-SSL in WSI-only settings across both datasets, with a notably larger performance gap at lower k values.

In the multimodal setting, TITAN is pretrained with closed-source pathology reports and synthetic captions. For a fair comparison, we evaluate GECKO using gene data and derived concept priors. While TITAN slightly surpasses our method by 0.5% on the Lung dataset for $k = 10, 25$, our approach significantly excels on the challenging EBV+MSI vs. Others dataset, with improvements of 6.7% at $k = 10$ and 1.3% at $k = 25$. Notably, TITAN is pretrained on over 100K paired multimodal samples, whereas GECKO uses only about 800 WSIs for the Lung dataset and around 200 WSIs for the EBV+MSI vs. Others STAD dataset. Furthermore, GECKO can incorporate the pathology reports used in TITAN's pretraining as an extra modality, potentially boosting performance further.

Additional ablations. In Supp. sec 8, we provide additional experiments to study (1) choice of WSI Concept-Encoding branch architecture, and (2) the effect of the keep ratio (r_{keep}) in false negative elimination.

5. Conclusion

We introduce GECKO, a novel pretraining framework that learns robust WSI-level representations without the need for auxiliary data modalities, while also delivering inherently interpretable WSI-level concept embeddings. Additionally, GECKO effectively incorporates additional modalities when available, outperforming existing unimodal and multimodal methods and achieving consistent performance enhancements. In future research, we aim to explore pan-cancer pretraining with an expanded concept bank that encompasses a wide range of cancer sites and subtypes, setting the stage for a foundational slide-level GECKO capable of interpretable pan-cancer zero-shot prediction. Furthermore, our WSI-level concept embedding could be utilized in generative models [8, 40], offering greater control in generating gigapixel images.

6. Acknowledgment

This research was partially supported by NIH and NCI grants 1R21CA258493-01A1, 1R01CA297843-01, 3R21CA258493-02S1, NSF grants 2442053, 2123920, 2212046, as well as Stony Brook Profund 2022 seed funding and private support from Bob Beals and Betsy Barton. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lió, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, pages 1801–1825. PMLR, 2023. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[3] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022. 3

[4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 1

[5] 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. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*, 2024. 2, 3, 8

[6] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022. 3

[7] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023. 8

[8] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 8

[9] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 3

[10] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2785–2795, 2022. 5, 12

[11] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36: 37995–38017, 2023. 3

[12] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 4, 5, 12

[13] Guillaume Jaume, Pushpak Pati, Behzad Bozorgtabar, Antonio Foncubierta, Anna Maria Anniciello, Florinda Feroce, Tilman Rau, Jean-Philippe Thiran, Maria Gabrani, and Orcun Goksel. Quantifying explainers of graph neural networks in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8106–8116, 2021. 3

[14] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Thomas Peeters, Andrew H Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9632–9644, 2024. 2, 3, 5, 6, 7, 12, 13

[15] Guillaume Jaume, Anurag Vaidya, Andrew Zhang, Andrew H. Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Faisal Mahmood. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024. 2, 3, 6, 7

[16] Sajid Javed, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, and Mohammed Bennamoun. Cplip: zero-shot learning for histopathology with comprehensive vision-language alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11450–11459, 2024. 3

[17] Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R Gupta, and Prateek Prasanna. Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 11226–11237, 2024. 2, 3, 4, 5, 12

[18] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 3

[19] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 6

[20] Tristan Lazard, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. Giga-ssl: Self-supervised learning for gigapixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023. 1, 3, 7, 8

[21] Tristan Lazard, Marvin Lerousseau, Sophie Gardrat, Anne Vincent-Salomon, Marc-Henri Stern, Manuel Rodrigues, Etienne Decencière, and Thomas Walter. Democratizing computational pathology: optimized whole slide image representations for the cancer genome atlas. *bioRxiv*, pages 2023–12, 2023. 8

[22] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1

[23] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023. 2, 5, 6

[24] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 3, 5

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 4

[27] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1 (5):206–215, 2019. 2

[28] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. 8

[29] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024. 3

[30] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11566–11578, 2024. 2, 3, 8

[31] Susu Sun, Leslie Tessier, Frédérique Meeuwsen, Clément Grisi, Dominique van Midden, Geert Litjens, and Christian F Baumgartner. Label-free concept based multiple instance learning for gigapixel histopathology. *arXiv preprint arXiv:2501.02922*, 2025. 3

[32] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022. 4

[33] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34: 24261–24272, 2021. 4

[34] Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025. 7, 13

[35] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023. 1

[36] Hasindri Watawana, Kanchana Ranasinghe, Tariq Mahmood, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Hierarchical text-to-vision self supervised alignment for improved histopathology representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 167–177. Springer, 2024. 3

[37] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, pages 1–10, 2025. 3

[38] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. 1, 3, 8

[39] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept

bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 3

[40] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5182–5191, 2024. 8

[41] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[42] Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus CF Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. *npj Digital Medicine*, 7 (1):383, 2024. 3, 5, 6, 7, 13

[43] Xiao Zhou, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pretraining for computational pathology. In *European Conference on Computer Vision*, pages 345–362. Springer, 2024. 3