# Unsupervised Foundation Model-Agnostic Slide-Level Representation Learning

Tim Lenz[1 *], Peter Neidlinger[1 *], Marta Ligero[1], Georg Wölflein[1,2], Marko van Treeck[1], Jakob N. Kather[1,3,4]
[1]EKFZ for Digital Health TU Dresden, [2]University of St Andrews,
[3]Heidelberg University Hospital, [4]University Hospital Dresden
{tim.lenz,peter.neidlinger,jakob_nikolas.kather}@tu-dresden.de

## Abstract

*Representation learning of pathology whole-slide images (WSIs) has primarily relied on weak supervision with Multiple Instance Learning (MIL). This approach leads to slide representations highly tailored to a specific clinical task. Self-supervised learning (SSL) has been successfully applied to train histopathology foundation models (FMs) for patch embedding generation. However, generating patient or slide level embeddings remains challenging. Existing approaches for slide representation learning extend the principles of SSL from patch level learning to entire slides by aligning different augmentations of the slide or by utilizing multimodal data. By integrating tile embeddings from multiple FMs, we propose a new single modality SSL method in feature space that generates useful slide representations. Our contrastive pretraining strategy, called* COBRA*, employs multiple FMs and an architecture based on Mamba-2.* COBRA *exceeds performance of state-of-the-art slide encoders on four different public Clinical Protemic Tumor Analysis Consortium (CPTAC) cohorts on average by at least* $+4.4\%$ *AUC, despite only being pretrained on 3048 WSIs from The Cancer Genome Atlas (TCGA). Additionally,* COBRA *is readily compatible at inference time with previously unseen feature extractors. Code available at https://github.com/KatherLab/COBRA.*

## 1. Introduction

In recent years, self-supervised learning (SSL) has emerged as a foundational approach in Computational Pathology (CPath), providing the basis for weakly supervised models to achieve remarkable results in diagnostic, prognostic, and treatment response prediction tasks [3, 4, 10, 11, 19, 24, 26–28, 32, 37, 39, 41, 44, 47]. By capturing informative, low-dimensional representations from unannotated whole-slide images (WSIs), SSL has enabled weakly supervised models to use these features for downstream tasks, effectively bridging the gap between high-resolution data and the limited availability of fully annotated datasets. SSL excels in generating low-dimensional feature representations for gigapixel WSIs, which can reach dimensions of 150,000 $\times$ 150,000 pixels (px), making them challenging to process with Vision Transformers (ViTs) [8] due to memory constraints. Consequently, most CPath approaches tessellate WSIs into smaller patches and extract low-dimensional embeddings for these patches using pretrained histopathology foundation models (FMs) [23]. Typically, these patch embeddings are used in weakly-supervised models for downstream classification tasks via multiple-instance learning (MIL) [7, 16, 35].

In addition to patch-based representations, SSL can also generate slide-level embeddings without any human annotations [20, 21, 46]. Pretrained SSL models can be leveraged to achieve impressive results on downstream tasks with minimal labeled data for task-specific fine-tuning, offering practical advantages like reduced labeling costs, elimination of noisy labels inherent to inter-observer variability, and improved generalizability through label-free representations. Central to SSL is the alignment of multiple representations of WSIs or related modalities (e.g., morphological text descriptions) into a shared latent space using contrastive learning or other similarity-based pretraining methods. However, generating effective augmentations to create these representations remains challenging. While image-level augmentations have been widely explored for patch-based learning, they may fail to produce diverse feature augmentations, as many modern FMs are designed to be invariant to these transformations [29, 43]. Other approaches, such as using different stainings (e.g., hematoxylin and eosin (H&E) combined with immunohistochemistry (IHC)), have shown potential but are limited by the availability of multi-stained tissue samples [18]. Similarly, aligning multiple modalities, such as text or gene expression data, has produced promising results but is constrained by the limited availability of such datasets and requires additional compute to process the

*Equal contribution

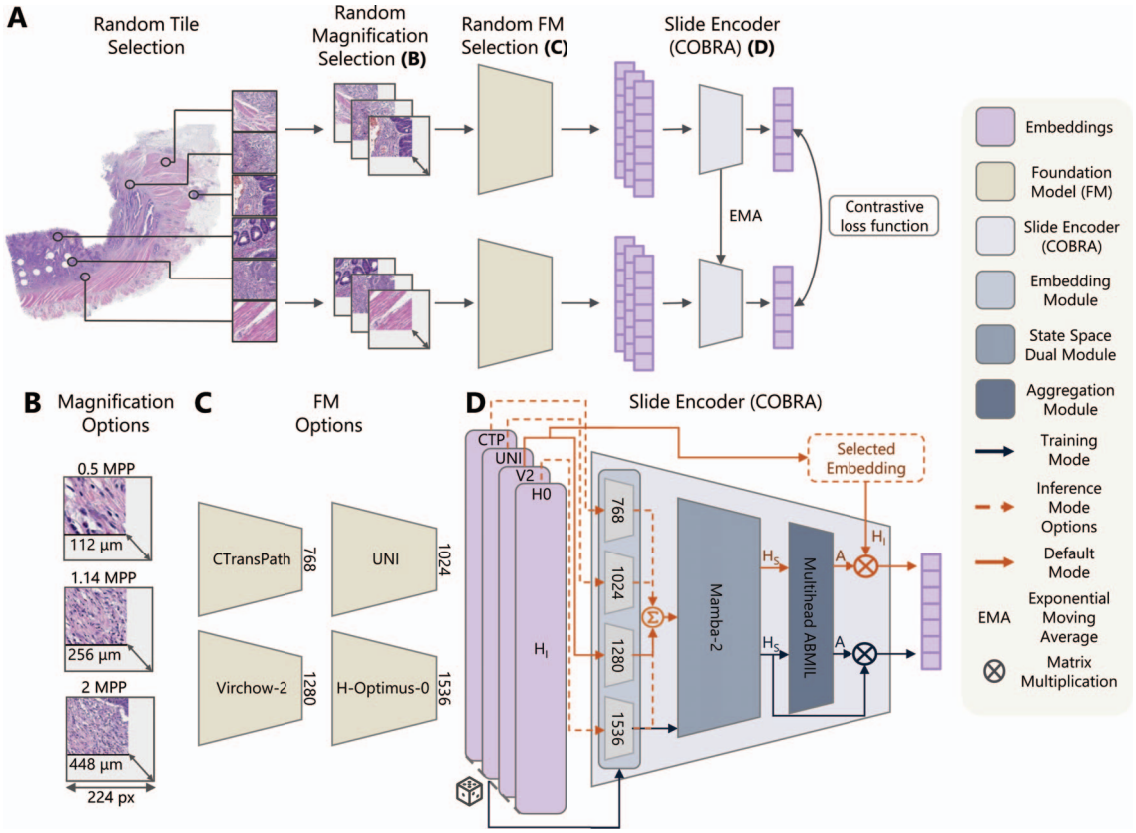


Figure 1. **COBRA overview** for self-supervised slide representation learning (**A**). A WSI is tessellated into patches at different magnifications (**B**) and encoded using different foundation models (FMs) (**C**) to produce tile embeddings. The magnifications (**B**) and foundation models (FMs) (**C**) serve as feature space augmentations to pretrain the COBRA slide encoder (**D**) using contrastive self-supervised learning.

different modalities [17, 34, 42].

To address these challenges, we propose a novel SSL method for image-only slide representation learning called COntrastive Biomarker Representation Alignment (COBRA). COBRA integrates tile embeddings from multiple FMs to generate augmentations directly in feature space, which can then be used to train a slide- or patient-level encoder. By employing Mamba-2 [6] followed by multi-head gated attention [18] and a contrastive loss objective, COBRA produces robust slide-level embeddings. Our contributions are summarized as follows:

- We propose an unsupervised single-modality contrastive slide encoder framework (COBRA) that avoids the need for stochastic image augmentations as it is trained and deployed on frozen patch embeddings. Extensive evaluations across 15 downstream classification tasks on three tissue types with external validation demonstrate COBRA's superiority over existing slide encoders.
- Our patient level encoder produces state-of-the-art (SOTA) unsupervised slide representations with unprecedented data efficiency, outperforming existing approaches with only a fraction of the pretaining data (3048 WSIs across four tissue types).
- We show that COBRA can turn patch level FMs, including ones not encountered during training, into better slide level feature extractors *without any additional finetuning*, making it particularly valuable as new FMs emerge.
- COBRA can be deployed across different WSI magnifications, where lower magnifications yield significant gains in computational efficiency with minimal sacrifice of downstream classification performance.

## 2. Related work

**Patch representation learning** Most works applying SSL focus on creating embeddings from image patches. Training a ViT with an SSL method like Dino-v2 [30] is now the preferred approach for learning task-agnostic image representations in CPath. SOTA FMs usually combine alignment- and reconstruction-based objectives trained with a student-teacher learning paradigm. These FMs are trained on increasingly large datasets and architectures (e.g. ViT-Giant [44] or trained on up to 3M WSIs [47]).

Table 1. **Slide encoder overview**. Abbreviations are as follows: # Ps refers to the number of parameter and # WSI refers to the number of WSIs the slide encoder was pretrained on.

| Model | # Ps[M] | # WSI[K] | Patch FM |
|---|---|---|---|
| Gigapath-SE [44] | 86 | 171 | Gigapath [44] |
| CHIEF [42] | 1 | 60 | CTransPath [41] |
| PRISM [34] | 513 | 587 | Virchow [39] |
| MADELEINE [18] | 5 | 21 | CONCH [24] |
| COBRA | 15 | 3 | CTransPath [41], UNI [4], Virchow2 [47], H-Optimus-0 [32] |

Besides image-only FMs, vision-language pathology FMs have recently emerged which rely on large-scale paired data [15, 24].

**Multiple instance learning** The SOTA approach for WSI classification is generating tile embeddings using FMs and then using these embeddings in a MIL approach to train an aggregator model for a specific downstream task. In particular, Attention-based MIL (ABMIL) [16] and many extensions thereof have been proposed [10, 23, 35, 40, 45]. While MIL approaches are prevalent for WSI classification, they are typically supervised and tailored to specific tasks.

**Slide representation learning** In contrast to MIL, slide representation learning constructs embeddings in an unsupervised manner and is task-agnostic. This next frontier in representation learning of histology images has been proposed in several works. In early work, Chen et al. proposed a hierarchical self-distillation approach for learning unsupervised WSI-level representations [3]. Lazard et al. used augmented patches to create many embeddings of the same input image to enable contrastive learning with slide embeddings [21]. In GigaPath, Xu et al. trained a masked autoencoder on the embeddings of their patch encoder to obtain slide representations [44]. More recent work applied vast amounts of multimodal data to pretrain aggregation models [18, 34, 42]. Differing from previous methodologies, we achieve state-of-the-art WSI-patient-level encoding by performing self-supervised contrastive learning on frozen vision features with a fraction of the data volume. None of the mentioned studies used less than 10K WSIs for WSI-level encoder pretraining [3, 18, 21, 42], while PRISM [34] and Gigapath [44] were trained on over 100K WSIs. COBRA surpasses the performance of earlier work, even though it is trained on only 3K publicly available WSIs (see Table 1).

# 3. Method

COBRA is an unsupervised slide representation learning framework. Given a set of WSIs $\{\mathbf{X}_i | \mathbf{X}_i \in \mathbb{R}^{d_x \times d_y \times 3}\}$ belonging to a single patient, it produces a single $d$-dimensional feature vector $\boldsymbol{z} \in \mathbb{R}^d$ representing that patient. We provide a brief overview of COBRA below and in Fig. 1, before going into detail in the following subsections.

COBRA operates on preprocessed patch embeddings (Sec. 3.1) from a set of CPath FMs. Its architecture consists of a Mamba-2 [6] encoding module, a multi-head attention-based pooling module for learning a patient-level slide embedding (Sec. 3.2) and an embedding module that learns to align multiple FMs into the same embedding space. COBRA can be deployed in various different modes, which makes it very flexible to adapt to different FMs (see Sec. 3.3). We train COBRA using a contrastive loss [38] (Sec. 3.4) and evaluate it on a variety of external validation tasks (Sec. 4).

## 3.1. Preprocessing

Given a histology slide ($\mathbf{X}_i \in \mathbb{R}^{d_x \times d_y \times 3}$), we tessellate the slide into ($224 \times 224$) px patches and remove background tiles by employing Canny background detection [31]. Next, we extract patch embeddings with pretrained FMs and pool the resulting feature vectors into a slide embedding. We use $fe_n$ to refer to the $n^{\text{th}}$ FM, $fe_n \in \{\text{CTP}, \text{UNI}, \text{V2}, \text{H0}\}$ denoting CTransPath [41], UNI [4], Virchow2 [47], and H-optimus-0 [32], respectively. By integrating FMs of different sizes and with different strengths, we aim to capture a diverse set of morphological features and ensure that our slide representations are robust and that COBRA is adaptable to other FMs. We obtain the patch embeddings $\boldsymbol{H}^{fe_n} \in \mathbb{R}^{N_t \times d_n}$ with $N_t$ and $d_n$ denoting the number of tiles and the embedding dimension $d_n \in ds = \{768, 1024, 1280, 1536\}$. We extract patch embeddings at 0.5, 1.14 and 2 microns per pixel (MPP) using 3048 WSIs from 2848 patients in TCGA BRCA, CRC, LUAD, LUSC and STAD. The use of multiple magnifications acts as a form of data augmentation in feature space, enriching the model's learning by providing multiscale contextual information. This approach enhances the model's ability to learn scale-invariant representations and improves its generalization across different tasks.

## 3.2. Architecture

The slide encoder consists of individual embedding MLPs for the different FMs and two Mamba-2 layers [6] followed by multihead gated attention [16, 18]. The embedding module is a layer norm [1] followed by an MLP with one hidden layer and SiLU activation [14]. It projects the different embedding dimensions of the FMs to the shared embedding space of the slide encoder. Inspired by MambaMIL [45], we use two Mamba [12] layers to efficiently encode the feature embeddings. We opt for the Mamba-2 state space dual (SSD) modules as they scale substantially better for higher state-space dimensions compared to original Mamba modules [6]. Additional information on the hyperparame-

ters used can be found in Appendix A.

Formally, the architecture may be described as follows: Let $f_{SE} : \mathbb{R}^{N_t \times ds} \rightarrow \mathbb{R}^d$ denote the slide encoder consisting of three submodules $f_E : \mathbb{R}^{N_t \times ds} \rightarrow \mathbb{R}^{N_t \times d}$, $f_S : \mathbb{R}^{N_t \times d} \rightarrow \mathbb{R}^{N_t \times d}$ and $f_A : \mathbb{R}^{N_t \times d} \rightarrow \mathbb{R}^d$, given by

$$\boldsymbol{z} = f_{SE}(\boldsymbol{H}^{fe_n}) = f_A(f_S(f_E(\boldsymbol{H}^{fe_n}))),\ \boldsymbol{H}^{fe_n} \in \mathbb{R}^{N_t \times d_n}, \tag{1}$$

where $f_E$,$f_S$,$f_A$ denote the *embedding module*, the *state-space dual module* and the *aggregation module*, respectively, and $d_n \in ds = \{768, 1024, 1280, 1536\}$ and $\boldsymbol{H}^{fe_n}$ refers to the patch embedding of the $n^{\text{th}}$ FM. The *embedding module* $f_E$ is defined as follows:

$$\boldsymbol{H}_E = f_E(\boldsymbol{H}^{fe_n}) = \text{Lin}(\text{SiLU}(\text{Lin}(\text{LN}(\boldsymbol{H}^{fe_n})))), \tag{2}$$

where Lin denotes a linear layer and LN denotes layer norm. The *state-space dual module* $f_S$ is specified as:

$$\boldsymbol{H}_S = f_S(\boldsymbol{H}_E) = \text{Lin}(\text{SSD}(\text{SSD}(\boldsymbol{H}_E) + \boldsymbol{H}_E) + \boldsymbol{H}_E). \tag{3}$$

The *aggregation module* $f_A$ consists of multi-head gated attention [16, 18] to aggregate the input embeddings into a single feature vector via a weighted average. For multi-head gated attention, the encoded embeddings are split into $M$ parts for the $M$ heads: $\boldsymbol{H}_S = \{\boldsymbol{H}_S^m\}_{m \in \{1,\dots,M\}}$ with $\boldsymbol{H}_S^m \in \mathbb{R}^{N_t \times \frac{d}{M}}$. The *aggregation module* $f_A$ is given by

$$\begin{aligned} \boldsymbol{z} = f_A(\boldsymbol{H}_S) &= \sum_{k=1}^{N_t} a_k(\boldsymbol{H}_{S,k}) \cdot \boldsymbol{H}_{S,k}; \\ a_k(\boldsymbol{H}_{S,k}) &= \frac{1}{M} \sum_{m=1}^{M} a_k^m(\boldsymbol{H}_{S,k}^m), \end{aligned} \tag{4}$$

with $\boldsymbol{H}_{S,k} \in \mathbb{R}^d$ and $a_k^m \in \mathbb{R}$ is defined as:

$$a_k^m(\boldsymbol{H}_{S,k}^m) = \frac{\exp\left(\boldsymbol{w}_m^\top\left(\tanh\left(\boldsymbol{V}_m(\boldsymbol{H}_{S,k}^{m\top})\right) \odot \sigma\left(\boldsymbol{U}_m \boldsymbol{H}_{S,k}^{m\top}\right)\right)\right)}{\sum_i^{N_t} \exp\left(\boldsymbol{w}_m^\top\left(\tanh\left(\boldsymbol{V}_m \boldsymbol{H}_{S,i}^{m\top}\right) \odot \sigma\left(\boldsymbol{U}_m \boldsymbol{H}_{S,i}^{m\top}\right)\right)\right)}, \tag{5}$$

with $\sigma$ denoting the sigmoid function and $\boldsymbol{w} \in \mathbb{R}^{p \times 1}, \boldsymbol{U} \in \mathbb{R}^{p \times d}, \boldsymbol{V} \in \mathbb{R}^{p \times d}$ as learnable parameters and $p$ being the attention dimension.

### 3.3. Inference modes

During self-supervised pretraining, the slide encoder learns to map the patch embeddings ($\boldsymbol{H}^{fe_n}$) of different slides, patches, FMs and magnifications from the same patient to be close in slide embedding space ($\boldsymbol{z}$). For this purpose, encoded embeddings are aggregated to a single feature vector.

**Single-FM inference mode** In line with Wang et al. [42], we found it beneficial at inference time to compute the weighted average in Eq. (4) using the original patch embeddings ($\boldsymbol{H}^{fe_n}$) instead of the encoded embeddings ($\boldsymbol{H}_S$) to obtain the slide-level representation (see Appendix D.1). Importantly, we still use the encoded embeddings to compute the weighting $a_k(\boldsymbol{H}_{S,k})$ of that average. Specifically, at inference time, Eq. (4) becomes

$$\boldsymbol{z} = f_{A_{\text{inf}}}(\boldsymbol{H}_S, \boldsymbol{H}^{fe_n}) = \sum_k^{N_t} a_k(\boldsymbol{H}_{S,k}) \cdot \boldsymbol{H}_k^{fe_n}. \tag{6}$$

We refer to this as the *single-FM inference mode* of COBRA and provide an ablation for the choice of Eq. (4) *vs*. Eq. (6) in Appendix D.1. Unless stated otherwise, we will denote as COBRA the *single-FM inference mode* version using Virchow2 patch embeddings as input, which is given by

$$\boldsymbol{z} = f_{SE_{\text{inf}}}(\boldsymbol{H}^{V2}, \boldsymbol{H}^{V2}). \tag{7}$$

**Multi-FM inference mode** Additionally, one can use feature vectors from multiple different FMs and average the embeddings after the embedding module to extract patient-level features which incorporate the knowledge of the different FMs simultaneously with $f_{SE_{\text{inf}}}^\dagger : \mathbb{R}^{N_t \times ds} \times \mathbb{R}^{N_t \times d_k} \rightarrow \mathbb{R}^d$ ($d_k \in ds$):

$$\begin{aligned} \boldsymbol{z}^\dagger &= f_{SE_{\text{inf}}}^\dagger(\{\boldsymbol{H}^{fe_n}\}_{n \in \{1,\dots,N_{FM}\}}, \boldsymbol{H}^{fe_l}) \\ &= f_{A_{\text{inf}}}\left(f_S\left(\frac{\sum_n^{N_{FM}} f_E^\dagger(\boldsymbol{H}^{fe_n})}{N_{FM}}\right), \boldsymbol{H}^{fe_l}\right). \end{aligned} \tag{8}$$

Here, $N_{FM}$ denotes the number of FMs used for pretraining and $\boldsymbol{H}^{fe_l}$ refers to the patch embeddings that are aggregated during inference. Additional information on the inference modes can be found in Appendix B.

### 3.4. Contrastive loss function

Following He et al. [13], we interpret contrastive learning as training an encoder for a *dictionary look-up task*:

Consider a set of encoded samples, denoted as $K = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \dots, \boldsymbol{k}_N\}$, which represent the keys of a dictionary. For a given query $\boldsymbol{q}$, there exists exactly one matching key $\boldsymbol{k}^+ \in K$. The contrastive loss is minimized when $\boldsymbol{q}$ closely matches $\boldsymbol{k}^+$ and diverges from all other keys. The InfoNCE [38] loss function is defined as

$$\mathcal{L}_{\mathbf{q}} = -\log \frac{\psi(\mathbf{q}, \mathbf{k}^+)}{\sum_{i=1}^{N} \psi(\mathbf{q}, \mathbf{k_i})}, \tag{9}$$

where $\boldsymbol{q}$ and the corresponding $\boldsymbol{k}^+$ represent feature vectors produced by a randomly selected pretrained encoder, sampling patches from WSIs of the same patient and $N$ is the

batch size or the length of the memory queue. The function $\psi$ is defined as follows:

$$\psi(\mathbf{x_1}, \mathbf{x_2}) = \exp(\text{sim}(\mathbf{x_1}, \mathbf{x_2})/\tau), \tag{10}$$

where $\tau$ denotes the temperature parameter and the cosine similarity function is depicted as $\text{sim}(\cdot)$. To avoid feature collapse, the keys and queries should be generated by distinct encoders. Let $\theta_q$ denote the parameters of the query encoder with the dense projection head, then the parameters of the key encoder $\theta_k$ are updated as follows:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{11}$$

where $m \in [0, 1)$ is the momentum coefficient. With the key encoder as the exponential average of the query encoder, the key representations stay more consistent, which enables a more stabilized training process. We adapted the public MoCo-v3 [5] repository for our experiments to align the embedding space of the slide embeddings generated with tile embeddings from different FMs.

# 4. Experiments & results

## 4.1. Dataset

**TCGA** We collected 3048 WSIs from 2848 patients using the cohorts TCGA [36] Breast Invasive Carcinoma (TCGA-BRCA, 1112 WSIs), TCGA Colorectal Carcinoma (TCGA-CRC, 566 WSIs), TCGA Lung Adenocarcinoma (TCGA-LUAD, 524 WSIs), TCGA Lung Squamous Cell Carcinoma (TCGA-LUSC, 496 WSIs), and TCGA Stomach Adenocarcinoma (TCGA-STAD, 350 WSIs). See Appendix C for detailed information. These cohorts were used for pretraining COBRA and for training the downstream classifiers and linear regression models. We emphasize that neither COBRA nor any FMs used in this study were pretrained on datasets included in the evaluation of the downstream tasks, precluding any data leakage.

**CPTAC** We collected 1604 WSIs from 444 patients using the cohorts CPTAC [9] Breast Invasive Carcinoma (CPTAC-BRCA, 395 WSIs), CPTAC Colon Adenocarcinoma (CPTAC-COAD, 233 WSIs), CPTAC Lung Adenocarcinoma (CPTAC-LUAD, 498 WSIs), and CPTAC Lung Squamous Cell Carcinoma (CPTAC-LUSC, 478 WSIs). These cohorts were exclusively used for external validation.

## 4.2. Pretraining setup

We trained COBRA on patch embeddings derived from slides of 2848 patients, using a batch size of 1024 across four NVIDIA A100 GPUs for 2000 epochs, which took approximately 40 hours. In total, we used 36576 extracted feature embeddings consisting of 3048 WSIs for each of the four FMs models and each of the three magnifications included into the pretraining. Additional information about the hyperparameters used for the training of COBRA can be found in the Appendix Tab. 5.

## 4.3. Tasks

CPath is used for different task categories. One important such category is biomarker prediction. Here, we focused on *STK11*, *EGFR*, *KRAS* and *TP53* mutation prediction in LUAD, *ESR1*, *PGR* and *ERBB2* expression, and *PIK3CA* mutation prediction in BRCA, and MSI status, *BRAF*, *KRAS*, *PIK3CA* mutation prediction in COAD. We also included classification of phenotypic subtypes, Non-Small Cell Lung Cancer (NSCLC) Subtyping and Sidedness prediction of COAD. Finally, we added N-Status prediction in COAD, a task that goes beyond the tissue itself and tries to classify whether the tumor has infiltrated lymph nodes, thereby influencing prognostication. We report area under the receiver operating characteristic (AUC) results in the main text, additional metrics such as F1 score, area under the precision recall characteristic (AUPRC) and the balanced accuracy of all experiments can be found in Appendix D. Unless indicated otherwise, all results are reported for 0.5 MPP (20× WSI magnification). In general, we conducted our evaluation experiments for three different WSI magnifications: 0.5 MPP (20×), 1.14 MPP (9×) and 2 MPP (5×). Additional information about the downstream experiments can be found in Appendix A.1.

## 4.4. Evaluation of patient embeddings

**MLP downstream classification** We evaluated COBRA's patient-level slide embeddings following standard practice in CPath using 5-fold cross-validation on the TCGA training cohort followed by deploying all five classifiers on the full external validation set CPTAC. The classifier is a simple MLP. Generating a slide embedding and then training a small MLP is much more efficient than current MIL approaches using tile embeddings. We compare COBRA to all mean patch embeddings of FMs used in this study and to the slide encoders MADELEINE [18], PRISM [34], GigaPath [44] and CHIEF [42] (see Tab. 2). All slide encoders except GigaPath and MADELEINE manage to outperform the mean patch embeddings of the FM they are based upon. However, COBRA is the only model that manages to reach a higher macro-AUC than Virchow2 mean patch embeddings. Nevertheless, it should be noted that MADELEINE was trained only on BRCA slides. Still, COBRA also substantially outperforms MADELEINE on most BRCA tasks (*ESR1*: +9.5%, *PGR* +5.5%, *ERBB2* +4.9%, *PIK3CA* -1.3% AUC). Overall, COBRA improves over PRISM by +4.4% average AUC and over the mean of the patch embeddings of Virchow2 by +1.5%. Especially on the COAD downstream tasks, MSI and *BRAF*, COBRA achieves substantial performance increases over the other slide encoders of at

Table 2. **Comparison of different slide encoders and mean baselines.** AUC performance of downstream tasks trained on TCGA and deployed on CPTAC. ST denotes Subtyping, SE denotes Slide Encoder. Overline indicates mean over patch embeddings, Concatenated refers to concatenated mean embeddings of all FMs involved in COBRA's pretraining, and Ensemble Prediction refers to the average of predictions from the mean patch embeddings of the training FMs. **Bold** indicates the best performance, and underline indicates the second-best performance.

| AUROC[%] | NSCLC | LUAD | | | | BRCA | | | | COAD | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | ST | STK11 | EGFR | TP53 | KRAS | ESR1 | PGR | ERBB2 | PIK3CA | MSI | BRAF | LN | KRAS | Side | PIK3CA | |
| CTransPath [41] | $87.2_{1.5}$ | $62.8_{2.5}$ | $59.3_{7.4}$ | $70.1_{2.3}$ | $52.4_{5.9}$ | $68.1_{2.5}$ | $66.5_{2.1}$ | $48.6_{1.5}$ | $56.3_{3.0}$ | $76.1_{4.6}$ | $59.8_{2.3}$ | $59.8_{1.0}$ | $55.9_{7.7}$ | $52.5_{2.6}$ | $56.3_{6.3}$ | $62.1_{4.2}$ |
| Virchow [39] | $89.4_{0.6}$ | $76.5_{6.8}$ | $60.3_{2.0}$ | $70.7_{1.7}$ | $54.3_{6.9}$ | $66.9_{4.8}$ | $60.5_{3.9}$ | $51.3_{5.7}$ | $63.5_{3.7}$ | $62.1_{6.7}$ | $65.0_{2.6}$ | $58.2_{5.1}$ | $53.9_{7.1}$ | $52.3_{3.1}$ | $52.4_{5.7}$ | $62.5_{4.9}$ |
| CONCH [24] | $96.5_{0.3}$ | $66.0_{10.3}$ | $62.0_{7.6}$ | $74.6_{1.6}$ | $59.0_{7.4}$ | $85.3_{1.5}$ | $80.3_{2.0}$ | $58.8_{11.0}$ | $63.2_{3.1}$ | $79.2_{0.5}$ | $57.5_{3.4}$ | $\mathbf{67.3}_{2.0}$ | $55.7_{8.6}$ | $53.4_{2.4}$ | $63.2_{5.3}$ | $68.1_{5.7}$ |
| UNI [4] | $95.8_{1.1}$ | $69.4_{2.4}$ | $70.1_{12.1}$ | $73.9_{0.8}$ | $50.7_{4.7}$ | $87.4_{3.1}$ | $74.9_{2.3}$ | $64.0_{3.4}$ | $62.3_{5.0}$ | $89.0_{1.6}$ | $73.0_{3.4}$ | $62.0_{7.8}$ | $63.5_{2.8}$ | $59.4_{3.7}$ | $63.5_{6.2}$ | $70.6_{4.9}$ |
| H-Optimus [32] | $97.2_{0.4}$ | $78.5_{2.7}$ | $78.2_{3.5}$ | $71.3_{1.1}$ | $58.1_{3.7}$ | $85.2_{2.7}$ | $74.9_{3.3}$ | $51.4_{4.5}$ | $59.5_{4.9}$ | $94.7_{0.7}$ | $77.1_{7.1}$ | $55.5_{3.6}$ | $59.2_{4.2}$ | $\mathbf{62.2}_{3.4}$ | $62.4_{9.0}$ | $71.0_{4.3}$ |
| GigaPath [44] | $96.6_{0.7}$ | $71.3_{1.9}$ | $75.7_{6.8}$ | $75.4_{1.3}$ | $56.9_{6.2}$ | $85.7_{1.1}$ | $75.9_{1.8}$ | $64.5_{2.7}$ | $62.2_{5.9}$ | $93.3_{1.6}$ | $77.5_{2.6}$ | $61.7_{2.9}$ | $56.1_{5.0}$ | $60.0_{1.5}$ | $59.4_{7.9}$ | $71.5_{4.0}$ |
| Virchow2 [47] | $95.8_{0.7}$ | $79.6_{5.5}$ | $78.3_{4.6}$ | $72.1_{0.7}$ | $\mathbf{60.9}_{5.6}$ | $89.2_{2.8}$ | $79.3_{2.7}$ | $71.3_{1.8}$ | $63.2_{4.8}$ | $94.9_{1.2}$ | $81.6_{4.5}$ | $63.0_{1.9}$ | $59.3_{6.2}$ | $56.3_{3.8}$ | $62.7_{11.6}$ | $73.8_{4.7}$ |
| Ensemble Prediction | $97.2_{0.3}$ | $77.2_{3.7}$ | $78.5_{4.1}$ | $73.3_{0.6}$ | $59.5_{5.1}$ | $87.6_{2.8}$ | $77.2_{2.7}$ | $65.6_{2.0}$ | $63.3_{4.1}$ | $94.7_{1.0}$ | $78.9_{5.4}$ | $62.5_{3.9}$ | $\mathbf{64.1}_{3.0}$ | $60.5_{2.1}$ | $64.5_{9.1}$ | $73.6_{4.0}$ |
| Concatenated | $97.4_{0.4}$ | $75.7_{3.0}$ | $\mathbf{80.2}_{2.2}$ | $72.5_{0.8}$ | $57.6_{4.8}$ | $89.6_{1.4}$ | $79.1_{3.4}$ | $67.5_{3.9}$ | $61.8_{4.0}$ | $\mathbf{95.0}_{1.1}$ | $82.2_{4.3}$ | $61.6_{2.5}$ | $59.7_{5.8}$ | $62.0_{2.4}$ | $\mathbf{70.2}_{4.1}$ | $74.1_{3.3}$ |
| GigaPath-SE [44] | $90.9_{1.3}$ | $67.0_{4.4}$ | $65.4_{4.4}$ | $73.7_{1.4}$ | $57.1_{5.2}$ | $72.9_{0.9}$ | $71.9_{3.3}$ | $55.4_{4.7}$ | $60.5_{4.6}$ | $66.2_{2.1}$ | $56.7_{4.5}$ | $54.6_{5.2}$ | $51.3_{2.9}$ | $45.8_{3.1}$ | $53.2_{5.7}$ | $62.8_{3.9}$ |
| MADELEINE [18] | $94.0_{0.6}$ | $72.2_{8.7}$ | $64.0_{6.7}$ | $72.0_{2.8}$ | $51.9_{3.9}$ | $80.1_{1.7}$ | $73.7_{1.3}$ | $66.7_{2.7}$ | $64.9_{1.6}$ | $68.6_{9.1}$ | $54.2_{6.7}$ | $60.3_{7.3}$ | $58.9_{6.6}$ | $50.5_{1.6}$ | $59.5_{8.6}$ | $66.1_{5.5}$ |
| CHIEF [42] | $93.6_{0.8}$ | $64.2_{10.7}$ | $62.8_{10.9}$ | $73.4_{1.5}$ | $50.1_{5.0}$ | $83.0_{0.5}$ | $77.5_{0.3}$ | $63.4_{2.3}$ | $65.4_{1.5}$ | $75.1_{4.8}$ | $63.6_{4.3}$ | $58.0_{1.7}$ | $58.4_{3.8}$ | $48.2_{4.2}$ | $56.6_{3.2}$ | $66.2_{4.9}$ |
| PRISM [34] | $\mathbf{99.2}_{0.1}$ | $\mathbf{87.6}_{1.6}$ | $70.7_{2.4}$ | $78.2_{0.5}$ | $52.9_{8.5}$ | $\mathbf{92.2}_{0.7}$ | $\mathbf{84.2}_{0.5}$ | $64.5_{6.0}$ | $\mathbf{69.4}_{2.1}$ | $79.1_{1.5}$ | $59.9_{1.4}$ | $67.2_{2.4}$ | $54.6_{6.2}$ | $52.2_{1.8}$ | $52.1_{6.8}$ | $70.9_{3.8}$ |
| COBRA | $98.1_{0.2}$ | $84.0_{2.9}$ | $80.0_{2.4}$ | $\mathbf{78.4}_{2.9}$ | $59.2_{6.2}$ | $89.6_{2.0}$ | $79.2_{2.4}$ | $\mathbf{71.6}_{2.2}$ | $63.6_{6.2}$ | $94.1_{0.5}$ | $\mathbf{87.8}_{2.0}$ | $65.7_{2.5}$ | $62.1_{10.4}$ | $58.3_{1.9}$ | $57.6_{7.5}$ | $\mathbf{75.3}_{4.4}$ |

Table 3. **Ablation over different inference modes.** AUC performance of COBRA embeddings compared to mean embeddings of the FMs involved. Overline indicates mean over patch embeddings, † indicates that embeddings of all four training FMs were used to generate the weighting vector (Eq. (8)). For the other COBRA entries, we used the inference mode from (Eq. (6)). **Bold** indicates the best performance, and underline indicates the second-best performance. The abbreviations are as follows: CTP: CTransPath [41], H0: H-Optimus-0 [32], V2: Virchow-2 [47], GP: GigaPath [44]. The different magnifications (5×, 9×, 20×) indicate which magnification of the WSIs was used to extract the embeddings.

| AUC[%] | NSCLC | LUAD | | | | BRCA | | | | COAD | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | ST | STK11 | EGFR | TP53 | KRAS | ESR1 | PGR | ERBB2 | PIK3CA | MSI | BRAF | LN | KRAS | Side | PIK3CA | |
| CTransPath [41] | $87.2_{1.5}$ | $62.8_{2.5}$ | $59.3_{7.4}$ | $70.1_{2.3}$ | $52.4_{5.9}$ | $68.1_{2.5}$ | $66.5_{2.1}$ | $48.6_{1.5}$ | $56.3_{3.0}$ | $76.1_{4.6}$ | $59.8_{2.3}$ | $59.8_{1.0}$ | $55.9_{7.7}$ | $52.5_{2.6}$ | $56.3_{6.3}$ | $62.1_{4.2}$ |
| UNI [4] | $95.8_{1.1}$ | $69.4_{2.4}$ | $70.1_{12.1}$ | $73.9_{0.8}$ | $50.7_{4.7}$ | $87.4_{3.1}$ | $74.9_{2.3}$ | $64.0_{3.4}$ | $62.3_{5.0}$ | $89.0_{1.6}$ | $73.0_{3.4}$ | $62.0_{7.8}$ | $63.5_{2.8}$ | $59.4_{3.7}$ | $63.5_{6.2}$ | $70.6_{4.9}$ |
| H-Optimus [32] | $97.2_{0.4}$ | $78.5_{2.7}$ | $78.2_{3.5}$ | $71.3_{1.1}$ | $58.1_{3.7}$ | $85.2_{2.7}$ | $74.9_{3.3}$ | $51.4_{4.5}$ | $59.5_{4.9}$ | $94.7_{0.7}$ | $77.1_{7.1}$ | $55.5_{3.6}$ | $59.2_{4.2}$ | $\mathbf{62.2}_{3.4}$ | $62.4_{9.0}$ | $71.0_{4.3}$ |
| GigaPath [44] | $96.6_{0.7}$ | $71.3_{1.9}$ | $75.7_{6.8}$ | $75.4_{1.3}$ | $56.9_{6.2}$ | $85.7_{1.1}$ | $75.9_{1.8}$ | $64.5_{2.7}$ | $62.2_{5.9}$ | $93.3_{1.6}$ | $77.5_{2.6}$ | $61.7_{2.9}$ | $56.1_{5.0}$ | $60.0_{1.5}$ | $59.4_{7.9}$ | $71.5_{4.0}$ |
| Virchow2 [47] | $95.8_{0.7}$ | $79.6_{5.5}$ | $78.3_{4.6}$ | $72.1_{0.7}$ | $\mathbf{60.9}_{5.6}$ | $89.2_{2.8}$ | $79.3_{2.7}$ | $71.3_{1.8}$ | $63.2_{4.8}$ | $94.9_{1.2}$ | $81.6_{4.5}$ | $63.0_{1.9}$ | $59.3_{6.2}$ | $56.3_{3.8}$ | $62.7_{11.6}$ | $73.8_{4.7}$ |
| COBRA-CTP | $95.9_{0.6}$ | $65.0_{10.5}$ | $66.0_{5.5}$ | $74.8_{1.8}$ | $49.2_{4.4}$ | $78.6_{0.7}$ | $72.2_{0.6}$ | $62.0_{1.4}$ | $60.9_{3.6}$ | $80.3_{2.2}$ | $73.2_{3.0}$ | $61.3_{1.6}$ | $52.4_{4.2}$ | $48.1_{2.3}$ | $56.3_{4.3}$ | $66.4_{4.0}$ |
| COBRA-UNI | $98.8_{0.3}$ | $79.4_{2.5}$ | $76.5_{5.3}$ | $78.9_{1.2}$ | $52.2_{4.2}$ | $88.1_{1.7}$ | $80.5_{3.0}$ | $65.1_{4.3}$ | $63.9_{5.2}$ | $89.1_{1.1}$ | $82.8_{1.5}$ | $64.6_{2.2}$ | $59.0_{8.4}$ | $57.4_{2.1}$ | $\mathbf{64.3}_{5.8}$ | $73.4_{3.9}$ |
| COBRA-H0 | $\mathbf{99.4}_{0.2}$ | $86.5_{1.8}$ | $79.9_{2.9}$ | $80.1_{2.4}$ | $54.3_{4.7}$ | $87.1_{1.0}$ | $74.0_{4.2}$ | $64.2_{4.9}$ | $55.7_{2.3}$ | $\mathbf{96.0}_{0.6}$ | $86.2_{3.3}$ | $58.2_{2.5}$ | $62.2_{4.4}$ | $57.2_{1.9}$ | $62.9_{4.7}$ | $73.6_{3.2}$ |
| COBRA-V2 | $98.1_{0.2}$ | $84.0_{2.9}$ | $80.0_{2.4}$ | $78.4_{2.9}$ | $59.2_{6.2}$ | $\mathbf{89.6}_{2.0}$ | $79.2_{2.4}$ | $71.6_{2.2}$ | $63.6_{6.2}$ | $94.1_{0.5}$ | $87.8_{2.0}$ | $65.7_{2.5}$ | $62.1_{10.4}$ | $58.3_{1.9}$ | $57.6_{7.5}$ | $75.3_{4.4}$ |
| COBRA†-CTP | $95.9_{0.6}$ | $68.1_{5.1}$ | $69.2_{4.7}$ | $75.1_{1.6}$ | $46.7_{4.1}$ | $77.9_{0.8}$ | $71.3_{1.3}$ | $59.3_{1.5}$ | $59.2_{1.3}$ | $80.3_{1.5}$ | $73.5_{3.4}$ | $60.6_{2.5}$ | $55.2_{4.5}$ | $48.3_{3.4}$ | $54.3_{5.2}$ | $66.3_{3.2}$ |
| COBRA†-UNI | $99.1_{0.2}$ | $79.1_{2.8}$ | $76.2_{4.6}$ | $80.2_{0.7}$ | $55.0_{5.7}$ | $86.0_{1.7}$ | $78.1_{3.2}$ | $60.3_{4.9}$ | $62.3_{3.1}$ | $89.1_{0.7}$ | $83.5_{1.5}$ | $65.7_{2.1}$ | $65.3_{4.2}$ | $57.2_{2.1}$ | $61.9_{2.1}$ | $73.3_{3.1}$ |
| COBRA†-H0 | $\mathbf{99.4}_{0.1}$ | $\mathbf{86.9}_{2.0}$ | $80.9_{3.4}$ | $79.9_{1.8}$ | $56.7_{3.6}$ | $87.8_{1.2}$ | $72.8_{3.4}$ | $59.9_{2.1}$ | $58.0_{0.9}$ | $95.2_{1.1}$ | $84.9_{3.7}$ | $58.1_{2.6}$ | $59.7_{7.2}$ | $58.5_{2.4}$ | $61.2_{3.4}$ | $73.3_{3.1}$ |
| COBRA†-V2-5× | $99.0_{0.2}$ | $79.0_{1.2}$ | $\mathbf{82.9}_{2.7}$ | $79.9_{1.8}$ | $59.4_{3.1}$ | $89.0_{1.3}$ | $81.6_{1.5}$ | $62.1_{4.7}$ | $\mathbf{67.2}_{2.9}$ | $94.1_{0.7}$ | $75.5_{2.7}$ | $\mathbf{68.1}_{8.5}$ | $\mathbf{69.6}_{3.4}$ | $51.4_{7.8}$ | $61.3_{1.5}$ | $74.7_{3.7}$ |
| COBRA†-V2-9× | $98.9_{0.2}$ | $79.5_{1.2}$ | $76.7_{3.9}$ | $80.0_{1.8}$ | $53.0_{4.4}$ | $\mathbf{89.6}_{1.6}$ | $\mathbf{83.6}_{1.7}$ | $70.6_{2.6}$ | $65.8_{4.8}$ | $95.1_{0.9}$ | $82.5_{2.5}$ | $61.7_{0.5}$ | $61.2_{3.2}$ | $61.9_{2.9}$ | $58.4_{12.8}$ | $74.6_{4.2}$ |
| COBRA†-V2-20× | $98.4_{0.2}$ | $84.6_{1.9}$ | $78.9_{3.6}$ | $78.4_{2.6}$ | $55.9_{7.1}$ | $\mathbf{89.6}_{1.7}$ | $80.0_{2.3}$ | $\mathbf{72.2}_{1.7}$ | $65.1_{4.5}$ | $94.2_{0.6}$ | $\mathbf{88.7}_{1.6}$ | $64.8_{2.3}$ | $60.6_{5.7}$ | $58.6_{1.7}$ | $61.5_{4.1}$ | $\mathbf{75.4}_{3.3}$ |
| COBRA†-GP | $98.9_{0.3}$ | $81.5_{2.3}$ | $78.7_{4.2}$ | $\mathbf{80.9}_{1.1}$ | $56.9_{5.3}$ | $87.8_{1.2}$ | $77.5_{1.1}$ | $65.4_{1.3}$ | $64.6_{3.8}$ | $93.5_{1.2}$ | $85.6_{2.0}$ | $64.7_{2.4}$ | $59.2_{6.6}$ | $57.4_{2.1}$ | $56.9_{9.4}$ | $74.0_{3.8}$ |

least +15% and +24.2% average AUC, respectively.

**Linear probing few-shot classification** We also evaluate COBRA in a few-shot setting across 10 runs for high-performance tasks, where the mean patch embeddings of at least one FM scores an average macro-AUC of $> 0.7$ across the five folds of the full classification and where the TCGA cohorts contain at least 50 cases per class. These tasks are NSCLC Subtyping, *STK11*, *EGFR* and *TP53* mutation in LUAD, *ESR1*, *PGR* and *ERBB2* expression in BRCA, and *BRAF* mutation, Sidedness and MSI status prediction in COAD (see Fig. 2). Although COBRA was only trained on very few samples and with only one modality, we observe that it is still robust enough to achieve high few-shot performance compared to the other slide encoders. On the BRCA tasks, it slightly outperforms the competition, while it substantially exceeds the results of the other models on the COAD tasks. We provide further results and information about the few shot experiments in Appendix D.2. Interestingly, slide encoders demonstrate greater robustness with fewer training samples compared to their mean patch embedding baselines, as all slide encoders except GigaPath consistently outperform their corresponding baselines.

## 4.5. Inference ablations

**Foundation models** As COBRA is FM-agnostic, it can be used to improve small, inferior tile level FMs like CTransPath as COBRA-CTP improves over all slide encoders but PRISM (see Tabs. 2 and 3). This substantially improves efficiency, as CTransPath has approximately 30M

Table 4. **Evaluation of the magnification augmentation during pretraining** AUC performance of downstream tasks trained on TCGA and deployed on CPTAC. † indicates that embeddings of all four training FMs were used to generate the weighting vector (Eq. (8)), ‡ indicates that COBRA was only pretrained on 0.5 MPP. ST denotes Subtyping. **Bold** indicates the best performance, and underline indicates the second-best performance.

| | AUC[%] | NSCLC | LUAD | | | | BRCA | | | | COAD | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model | ST | STK11 | EGFR | TP53 | KRAS | ESR1 | PGR | ERBB2 | PIK3CA | MSI | BRAF | LN | KRAS | Side | PIK3CA | |
| 5× | COBRA‡-CTP | $88.7_{1.5}$ | $63.8_{10.5}$ | $66.2_{2.7}$ | $69.9_{2.0}$ | $51.7_{3.9}$ | $82.8_{1.2}$ | $75.0_{7.7}$ | $58.5_{3.1}$ | $61.5_{4.1}$ | $75.4_{1.4}$ | $62.5_{10.3}$ | $63.1_{2.5}$ | $53.8_{8.8}$ | $51.3_{1.4}$ | $51.7_{4.0}$ | $65.1_{5.4}$ |
| | COBRA‡-UNI | $90.3_{0.5}$ | $72.3_{9.9}$ | $68.5_{4.0}$ | $69.2_{2.2}$ | $53.3_{7.5}$ | $81.4_{1.2}$ | $75.9_{1.4}$ | $56.0_{2.9}$ | $64.1_{4.8}$ | $77.8_{1.6}$ | $71.9_{2.6}$ | $60.4_{2.4}$ | $57.4_{8.0}$ | $55.3_{2.3}$ | $51.3_{10.4}$ | $67.0_{5.2}$ |
| | COBRA‡-H0 | $89.5_{1.2}$ | $80.4_{4.2}$ | $69.1_{2.5}$ | $67.9_{1.6}$ | $52.5_{3.2}$ | $77.7_{1.8}$ | $73.6_{2.2}$ | $56.2_{2.8}$ | $62.0_{4.3}$ | $79.8_{2.8}$ | $71.4_{3.4}$ | $62.9_{1.9}$ | $56.3_{3.6}$ | $56.3_{2.1}$ | $58.0_{4.1}$ | $67.6_{2.9}$ |
| | COBRA†-CTP | $96.6_{0.4}$ | $70.5_{3.9}$ | $70.5_{2.1}$ | $74.3_{1.0}$ | $53.2_{2.4}$ | $82.2_{0.9}$ | $77.1_{0.8}$ | $65.7_{2.9}$ | $66.0_{2.6}$ | $79.3_{1.3}$ | $67.9_{2.8}$ | $60.6_{3.2}$ | $53.0_{8.8}$ | $47.2_{2.4}$ | $51.6_{2.3}$ | $67.7_{3.2}$ |
| | COBRA†-H0 | $97.6_{0.4}$ | $78.5_{4.0}$ | $71.3_{4.0}$ | $73.1_{1.3}$ | $51.8_{5.2}$ | $81.8_{1.1}$ | $74.9_{1.8}$ | $56.7_{2.0}$ | $63.1_{4.1}$ | $82.3_{1.6}$ | $71.4_{1.8}$ | $59.6_{3.1}$ | $56.0_{6.6}$ | $50.6_{2.4}$ | $55.7_{5.3}$ | $68.3_{3.5}$ |
| | COBRA†-UNI | $97.1_{0.4}$ | $77.1_{1.6}$ | $71.5_{2.5}$ | $75.2_{1.1}$ | $\underline{57.3}_{2.0}$ | $82.7_{0.7}$ | $74.3_{1.5}$ | $57.1_{3.2}$ | $\mathbf{68.2}_{4.2}$ | $78.8_{1.9}$ | $70.5_{2.2}$ | $55.9_{4.3}$ | $54.6_{7.0}$ | $49.3_{6.8}$ | $58.4_{4.4}$ | $68.5_{3.5}$ |
| | COBRA‡-V2 | $97.3_{0.4}$ | $81.3_{3.2}$ | $72.9_{2.6}$ | $77.0_{1.6}$ | $\mathbf{58.1}_{5.2}$ | $\mathbf{91.2}_{1.2}$ | $\mathbf{81.3}_{1.2}$ | $\underline{71.9}_{1.8}$ | $65.0_{3.1}$ | $87.2_{1.5}$ | $78.0_{1.8}$ | $64.2_{3.6}$ | $57.2_{12.4}$ | $60.3_{1.5}$ | $62.6_{8.8}$ | $73.7_{4.6}$ |
| | COBRA†-V2 | $99.0_{0.2}$ | $81.6_{1.5}$ | $75.5_{2.7}$ | $79.9_{1.8}$ | $51.4_{7.8}$ | $89.0_{1.3}$ | $79.0_{1.2}$ | $67.2_{2.9}$ | $62.1_{4.7}$ | $94.1_{0.7}$ | $82.9_{2.7}$ | $61.3_{1.5}$ | $\mathbf{68.1}_{8.5}$ | $59.4_{3.1}$ | $\mathbf{69.6}_{3.4}$ | $\underline{74.7}_{3.7}$ |
| 9× | COBRA‡-CTP | $93.7_{1.1}$ | $68.7_{17.7}$ | $70.9_{3.8}$ | $72.3_{1.5}$ | $51.0_{6.8}$ | $81.0_{0.7}$ | $74.5_{0.8}$ | $50.2_{3.8}$ | $57.3_{2.2}$ | $72.4_{2.1}$ | $64.9_{1.7}$ | $58.9_{4.8}$ | $56.8_{5.5}$ | $54.2_{1.7}$ | $\underline{63.4}_{2.4}$ | $66.0_{5.6}$ |
| | COBRA†-CTP | $96.4_{0.3}$ | $75.5_{3.5}$ | $69.4_{6.7}$ | $74.4_{1.4}$ | $52.1_{3.8}$ | $81.6_{0.6}$ | $76.2_{0.2}$ | $65.7_{1.5}$ | $62.0_{2.1}$ | $81.3_{1.0}$ | $72.1_{2.5}$ | $59.9_{2.2}$ | $52.6_{8.1}$ | $49.4_{4.4}$ | $55.4_{8.0}$ | $68.3_{4.0}$ |
| | COBRA‡-H0 | $98.0_{0.6}$ | $83.9_{1.7}$ | $74.6_{4.0}$ | $75.9_{1.7}$ | $47.9_{3.1}$ | $83.0_{1.2}$ | $76.8_{1.1}$ | $61.1_{4.7}$ | $60.9_{2.0}$ | $76.5_{2.0}$ | $69.1_{2.3}$ | $\mathbf{66.4}_{2.5}$ | $58.6_{4.6}$ | $57.5_{1.7}$ | $58.4_{7.2}$ | $69.9_{3.2}$ |
| | COBRA‡-UNI | $97.3_{0.5}$ | $80.2_{3.1}$ | $77.6_{4.6}$ | $75.7_{1.2}$ | $53.5_{3.5}$ | $86.3_{1.2}$ | $79.1_{0.8}$ | $60.0_{5.5}$ | $\underline{66.7}_{2.5}$ | $78.3_{2.2}$ | $72.0_{0.9}$ | $60.1_{3.0}$ | $56.9_{11.5}$ | $58.8_{1.9}$ | $59.7_{6.1}$ | $70.8_{4.3}$ |
| | COBRA‡-V2 | $97.5_{0.7}$ | $84.3_{1.5}$ | $77.2_{4.2}$ | $77.7_{2.1}$ | $52.0_{6.8}$ | $88.4_{1.9}$ | $79.4_{1.3}$ | $68.6_{4.4}$ | $62.1_{2.2}$ | $81.3_{1.6}$ | $72.8_{0.4}$ | $62.6_{2.3}$ | $58.5_{9.7}$ | $60.3_{1.1}$ | $59.4_{11.1}$ | $72.1_{4.7}$ |
| | COBRA†-H0 | $\underline{99.3}_{0.2}$ | $83.4_{2.3}$ | $73.6_{3.3}$ | $78.7_{2.1}$ | $52.4_{5.5}$ | $84.1_{2.0}$ | $77.2_{0.9}$ | $66.7_{1.7}$ | $62.3_{3.6}$ | $91.4_{0.7}$ | $82.5_{3.3}$ | $63.8_{2.7}$ | $56.2_{4.5}$ | $57.3_{2.8}$ | $58.3_{2.5}$ | $72.5_{2.9}$ |
| | COBRA†-UNI | $98.9_{0.3}$ | $71.6_{16.2}$ | $74.8_{3.4}$ | $\mathbf{80.5}_{1.7}$ | $56.3_{4.1}$ | $87.2_{0.7}$ | $79.1_{0.9}$ | $65.5_{2.6}$ | $66.0_{3.8}$ | $89.5_{1.4}$ | $\underline{85.2}_{1.9}$ | $59.9_{4.9}$ | $62.0_{8.7}$ | $58.4_{3.9}$ | $56.5_{5.7}$ | $72.8_{5.6}$ |
| | COBRA†-V2 | $98.9_{0.2}$ | $83.6_{1.7}$ | $76.7_{3.9}$ | $80.0_{1.8}$ | $53.0_{4.4}$ | $\underline{89.6}_{1.6}$ | $79.5_{1.2}$ | $70.6_{2.6}$ | $65.8_{4.8}$ | $\underline{95.1}_{0.9}$ | $82.5_{2.5}$ | $61.7_{0.5}$ | $58.4_{12.8}$ | $\underline{61.9}_{2.9}$ | $61.2_{3.2}$ | $74.6_{4.2}$ |
| 20× | COBRA‡-CTP | $94.1_{0.9}$ | $69.0_{4.7}$ | $67.9_{11.3}$ | $77.0_{0.9}$ | $51.3_{7.3}$ | $76.5_{1.7}$ | $70.7_{0.8}$ | $58.0_{1.5}$ | $51.3_{5.7}$ | $81.9_{0.7}$ | $67.1_{2.4}$ | $54.7_{4.6}$ | $54.6_{7.1}$ | $51.2_{1.6}$ | $59.4_{5.4}$ | $65.6_{4.8}$ |
| | COBRA†-CTP | $95.9_{0.6}$ | $68.1_{5.1}$ | $69.2_{4.7}$ | $75.1_{1.6}$ | $46.7_{4.1}$ | $77.9_{0.8}$ | $71.3_{1.3}$ | $59.3_{1.5}$ | $59.2_{1.3}$ | $80.3_{1.5}$ | $73.5_{3.4}$ | $60.6_{2.5}$ | $55.2_{4.5}$ | $48.3_{3.4}$ | $54.3_{5.2}$ | $66.3_{3.2}$ |
| | COBRA‡-UNI | $97.8_{0.6}$ | $75.5_{3.7}$ | $78.5_{4.9}$ | $\underline{80.4}_{1.7}$ | $53.6_{5.7}$ | $83.1_{2.4}$ | $73.6_{2.2}$ | $63.4_{3.7}$ | $59.2_{2.7}$ | $86.4_{1.6}$ | $76.5_{1.0}$ | $62.0_{3.2}$ | $57.5_{6.4}$ | $61.1_{1.4}$ | $60.6_{2.4}$ | $71.3_{3.3}$ |
| | COBRA‡-H0 | $98.8_{0.2}$ | $\underline{85.1}_{2.9}$ | $78.0_{8.4}$ | $79.5_{1.0}$ | $55.8_{3.9}$ | $86.7_{1.2}$ | $75.5_{2.7}$ | $66.3_{4.2}$ | $50.9_{2.7}$ | $91.0_{0.8}$ | $79.0_{2.7}$ | $56.0_{4.1}$ | $56.3_{6.3}$ | $\mathbf{64.3}_{0.9}$ | $62.8_{3.4}$ | $72.4_{3.7}$ |
| | COBRA†-H0 | $\mathbf{99.4}_{0.1}$ | $\mathbf{86.9}_{2.0}$ | $\underline{\mathbf{80.9}}_{3.4}$ | $79.9_{1.8}$ | $56.7_{3.6}$ | $87.8_{1.2}$ | $72.8_{3.4}$ | $59.9_{2.1}$ | $58.0_{0.9}$ | $\mathbf{95.2}_{1.1}$ | $84.9_{3.7}$ | $58.1_{2.6}$ | $59.7_{7.2}$ | $58.5_{2.4}$ | $61.2_{3.4}$ | $73.3_{3.1}$ |
| | COBRA†-UNI | $99.1_{0.2}$ | $79.1_{2.8}$ | $76.2_{4.6}$ | $80.2_{0.7}$ | $55.0_{5.7}$ | $86.0_{1.7}$ | $78.1_{3.2}$ | $60.3_{4.9}$ | $62.3_{3.1}$ | $89.1_{0.7}$ | $83.5_{1.5}$ | $\underline{65.7}_{2.1}$ | $\underline{65.3}_{4.2}$ | $57.2_{2.1}$ | $61.9_{2.1}$ | $73.3_{3.1}$ |
| | COBRA‡-V2 | $96.9_{0.3}$ | $83.4_{4.0}$ | $\underline{\mathbf{80.9}}_{3.4}$ | $78.8_{1.5}$ | $56.7_{4.4}$ | $88.3_{1.7}$ | $77.8_{1.9}$ | $70.7_{4.7}$ | $58.1_{1.3}$ | $91.8_{0.8}$ | $80.6_{2.3}$ | $62.2_{1.8}$ | $54.7_{11.9}$ | $61.3_{1.9}$ | $61.0_{7.2}$ | $73.5_{4.4}$ |
| | COBRA†-V2 | $98.4_{0.2}$ | $84.6_{1.9}$ | $78.9_{3.6}$ | $78.4_{2.6}$ | $55.9_{7.1}$ | $\underline{89.6}_{1.7}$ | $\underline{80.0}_{2.3}$ | $\mathbf{72.2}_{1.7}$ | $65.1_{4.5}$ | $94.2_{0.6}$ | $\mathbf{88.7}_{1.6}$ | $64.8_{2.3}$ | $60.6_{5.7}$ | $58.6_{1.7}$ | $61.5_{4.1}$ | $\mathbf{75.4}_{3.3}$ |

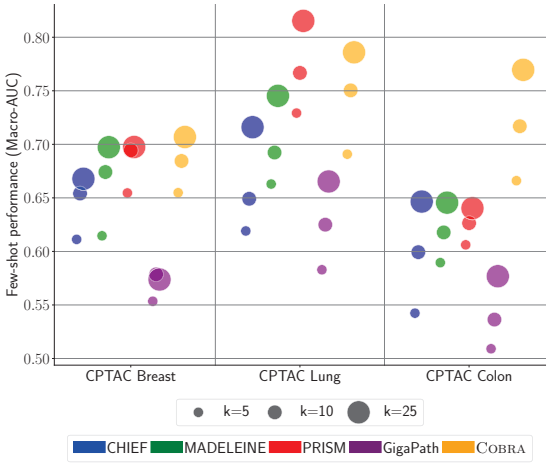


Figure 2. **Few shot linear probing classification**. Linear probing macro-AUC performance comparison for $k$ samples per class.

parameters, compared to more than 600M in Virchow2 and more than 1B in H-optimus-0.

**Magnifications** Another way to achieve efficiency improvements is to reduce the magnification of the WSIs for the patch embeddings, which in turn significantly reduces the number of tiles that need to be extracted and embedded. In particular, this change does not result in a significant performance drop as COBRA†-V2-5× and COBRA†-V2-9× achieve performance gains over the next best slide encoder PRISM at 0.5 MPP of +3.8% and +3.7% average AUC, respectively (see Tabs. 2 and 3), which we attribute to our multiscale alignment during pretraining.

**Combined inference and unseen FMs** In a combined inference mode (indicated by † in Tab. 3), where embeddings from all pretrained FMs are used, performance is slightly better for COBRA-V2, though it does not notably improve the downstream classification performance for 0.5 MPP. In general, performance is comparable to the single-FM mode. However, for smaller magnifications, the benefits of the combined inference mode become more notable (see Appendix Tabs. 10 and 11). Especially on 2 MPP, COBRA exhibits gains over the single-FM mode (on average +0.53% AUC). Furthermore, COBRA remains useful for future FMs as it can aggregate embeddings from unseen FMs and improve their performance over the mean baseline. We show evidence for this by deploying COBRA on GigaPath patch embeddings, which improves over the mean baseline of +2.5% average AUC and +3.1% average AUC over the next best slide encoder PRISM (Tabs. 2 and 3).

### 4.6. Pretraining ablation

**Single-magnification pretraining** We analyze the performance of a single magnification training on only 0.5 MPP embeddings and find that the use of all three magnifications results in an average AUC improvement of +1.73% AUC across all models when comparing the multi-FM inference mode (Eq. (8)). Furthermore, the three-magnification setup yields substantial gains in NSCLC subtyping at 5× magnification, with improvements of +6.8%


Figure 3. **COBRA Unsupervised Heatmap**. Visualization of the weighting scores for the tiles of a WSI generated by COBRA for Patient-ID TCGA-CA-6716 from TCGA-CRC.

AUC for UNI, +7.9% for CTransPath, +8.1% for H-optimus-0 and +1.7% for Virchow2 (Tab. 4). These results indicate that the use of multiple magnifications can enhance performance in certain cases and does not negatively impact model performance.

### 4.7. Interpretability

COBRA enables unsupervised interpretability as it is an aggregation method of patch embeddings that calculates a weighted average by assigning each tile a softmaxed value, which can be interpreted as an attention value. By visualizing these weightings for WSIs, we observe that the model shows high attention values for the tumor regions in the slide (see Fig. 3). It is worth mentioning that for these heatmaps, no GradCam [33] is required, and they are generated only based on patch embeddings, so each tile only receives one value instead of pixel-level attention that can be achieved with other methods. However, this extremely simple approach is sufficient to identify important tumor regions in detail without any supervision such as targeted segmentation training. More examples and detailed explanations can be found in Appendix E.

Furthermore, we visualized COBRA's embedding space using uniform manifold approximation and projection (UMAP) [25] plots of COBRA's patient-level slide embeddings extracted at 0.5 MPP for TCGA and CPTAC (see Fig. 4). We observe a decent separation between the different tissue types involved in this study, indicating that COBRA learned meaningful representations that can distinguish between tissue types without supervision.

## 5. Conclusion

In this paper, we introduced COBRA, a novel FM- and task-agnostic approach for slide representation learning. Trained on only 3048 WSIs from TCGA, COBRA achieves SOTA performance, even surpassing multimodal slide encoders. This is particularly valuable for medical imaging, where acquiring large annotated datasets is challenging due to privacy concerns and annotation costs. While additional data might enhance performance, our results indicate that COBRA is highly effective even in low-data regimes. These results highlight the potential of SSL in leveraging the strengths of histopathology FMs. Future work includes exploring SSL objectives that extend beyond contrastive approaches, as well as incorporating more cancer types, pretraining data and a larger variety of FMs into COBRA.

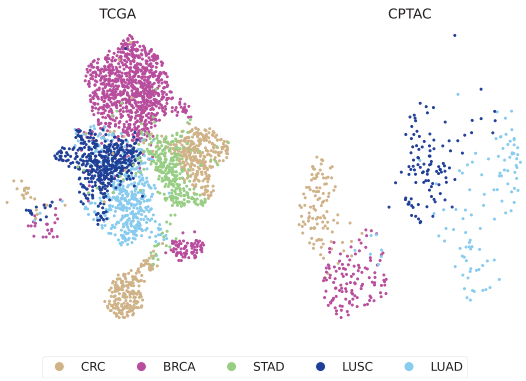


Figure 4. UMAP visualization of COBRA's patient-level slide embeddings for TCGA and CPTAC datasets at 0.5 MPP. Each color represents a different tissue type, with five tissue types in total.

**Acknowledgments** The authors gratefully acknowledge the GWK's support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden. We also acknowledge the TCGA Research Network and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), which generated the data on which the results shown in this study are based. GW is supported by Lothian NHS.

## References


[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[2] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Serdar O Sumer, Bülent A Aksoy, Anders Jacobsen, Christina J Byrne, Michael L Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Allen P Goldberg, Chris Sander, and Nikolaus Schultz. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012. 2

[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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16134, 2022. 1, 3

[4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 1, 3, 6, 4, 5, 7, 8, 9, 10, 11, 12

[5] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 5

[6] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. 2, 3

[7] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 1

[9] NJ Edwards, M Oberti, RR Thangudu, S Cai, PB McGarvey, S Jacob, S Madhavan, and KA Ketchum. The cptac data portal: A resource for cancer proteomics research. *Journal of Proteome Research*, 14(6):2707–2713, 2015. Epub 2015 May 4. 5

[10] Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nature Protocols*, 2024. 1, 3

[11] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. 1

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. 3

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 4

[14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. 3, 1

[15] Z. Huang, F. Bianchi, M. Yuksekgonul, et al. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29:2307–2316, 2023. 3

[16] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018. 1, 3, 4

[17] Guillaume Jaume, Lukas Oldenburg, Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Thomas Peeters, Andrew H. Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology, 2024. 2

[18] Guillaume Jaume, Anurag Jayant Vaidya, Andrew Zhang, Andrew H Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Mahmood Faisal. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*. Springer, 2024. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

[19] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, 2023. 1

[20] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations, 2020. 1

[21] Tristan Lazard, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. Giga-ssl: Self-supervised learning for gigapixel images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4305–4314, 2023. 1, 3

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1

[23] 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, 3

[24] Ming-Yu Lu, Bo Chen, Drew F.K. Williamson, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 1, 3, 6, 4, 5, 7, 8, 9, 10, 11, 12

[25] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. 8

[26] Patience Mukashyaka, Todd B. Sheridan, Ali Foroughi pour, and Jeffrey H. Chuang. Sampler: unsupervised repre-

sentations for rapid analysis of whole slide tissue images. *eBioMedicine*, 99:104908, 2024. 1

[27] O. S. M. El Nahhas, C. M. L. Loeffler, Z. I. Carrero, et al. Regression-based deep-learning predicts molecular biomarkers from pathology slides. *Nature Communications*, 15:1253, 2024.

[28] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology, 2024. 1

[29] Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology, 2024. 1

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2

[31] Weibin Rong, Zhanjing Li, Wei Zhang, and Lining Sun. An improved canny edge detection algorithm. In *2014 IEEE international conference on mechatronics and automation*, pages 577–582. IEEE, 2014. 3

[32] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. 1, 3, 6, 4, 5, 7, 8, 9, 10, 11, 12

[33] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 8, 3

[34] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. Prism: A multi-modal generative foundation model for slide-level histopathology, 2024. 2, 3, 5, 6, 4, 7, 8, 9, 10, 11, 12

[35] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 1, 3

[36] The Cancer Genome Atlas Research Network, J Weinstein, E Collisson, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013. 5

[37] Michaela Unger and Jakob Nikolas Kather. A systematic analysis of deep learning in genomics and histopathology for precision oncology. *BMC Medical Genomics*, 17(1):48, 2024. 1

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 3, 4

[39] E. Vorontsov, A. Bozkurt, A. Casson, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 2024. 1, 3, 6, 4, 5, 7, 8, 9, 10, 11, 12

[40] SJ Wagner, D Reisenbüchler, NP West, JM Niehues, J Zhu, S Foersch, GP Veldhuizen, P Quirke, HI Grabsch, PA van den Brandt, GGA Hutchins, SD Richman, T Yuan, R Langer, JCA Jenniskens, K Offermans, W Mueller, R Gray, SB Gruber, JK Greenson, G Rennert, JD Bonner, D Schmolze, J Jonnagaddala, NJ Hawkins, RL Ward, D Morton, M Seymour, L Magill, M Nowak, J Hay, VH Koelzer, DN Church, TransSCOT consortium, C Matek, C Geppert, C Peng, C Zhi, X Ouyang, JA James, MB Loughrey, M Salto-Tellez, H Brenner, M Hoffmeister, D Truhn, JA Schnabel, M Boxberg, T Peng, and JN Kather. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41(9):1650–1661.e4, 2023. Epub 2023 Aug 30. 3

[41] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022. 1, 3, 6, 4, 5, 7, 8, 9, 10, 11, 12

[42] X. Wang, J. Zhao, E. Marostica, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 2024. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

[43] Georg Wölflein, Dyke Ferber, Asier R. Meneghetti, Omar S. M. El Nahhas, Daniel Truhn, Zunamys I. Carrero, David J. Harrison, Ognjen Arandjelović, and Jakob Nikolas Kather. Benchmarking pathology feature extractors for whole slide image classification, 2024. 1

[44] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024. 1, 2, 3, 5, 6, 4, 7, 8, 9, 10, 11, 12

[45] Shu Yang, Yihui Wang, and Hao Chen. MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer Nature Switzerland, 2024. 3

[46] Zhimiao Yu, Tiancheng Lin, and Yi Xu. Slpd: Slide-level prototypical distillation for wsis, 2023. 1

[47] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Sever-

son. Virchow2: Scaling self-supervised mixed magnification models in pathology, 2024. 1, 2, 3, 6, 4, 5, 7, 8, 9, 10, 11, 12