Transcriptomics-guided Slide Representation Learning in Computational Pathology

Supplementary Material

1. Model & training

iBOT-Tox pre-training: iBOT-Tox is the first vision encoder for toxicologic pathology targeting non-human samples. It uses a Vision Transformer Base (ViT-B) [3] as backbone to learn 768-dimensional embeddings from 224×224 pixels image patches. ViTs are based on the self-attention paradigm to encode spatial interactions among small regions (called tokens) of the input image. iBOT-Tox is trained using the iBOT recipe [9], a state-of-the-art training strategy based on student-teacher knowledge distillation [1]. iBOT combines contrastive and reconstruction objectives: (1) a self-distillation objective to align different views of the input image based on image crop and augmentation. This objective helps to encode contextual and semantic information from the image, allowing for the creation of representations that are invariant to staining or rotation; and (2) a masked image modeling objective that aims to reconstruct image tokens from the other tokens. This objective helps to encode the image structure and is analogous to masked language modeling in Large Language Model training, such as BERT[2]. To train the network, we relied on the public implementation of iBOT*. iBOT-Tox was trained on 15 million patches extracted from different 47,227 WSIs (liver and kidney slides). We trained the network for 1,176,640 iterations or 80 epochs. The specific hyperparameters used for training are listed in Table 4. Most parameters were adapted from ImageNet-22K pre-training.

ABMIL architecture: TANGLE is using an ABMIL architecture [6, 7], which is composed of three components: a pre-attention MLP, consisting of 2 layers with 768 hidden units, layer normalization, GELU activation, and 0.1 dropout; a gated-attention network, consisting of 2-layer MLP with 512 hidden units, with Sigmoid and Tanh activation respectively and 0.25 dropout; and a post-attention network, consisting a linear layer with 768 units.

TANGLE pre-training: We pre-trained TANGLE with AdamW optimizer and a batch size of 128 for 50 epochs. The learning rate is linearly ramped up during a 5-epoch warmup from 1e-8 to 1e-4. Then, we employed a cosine scheduler to reach the final learning rate of 1e-8 after 50 epochs. To increase training diversity and simplify batch processing, we sample a fixed and random subset of patches per slide. In TG-GATEs, we sample 4,096 patches, and in TCGA-BRCA and TCGA-NSCLC, we sample 2,048 patches per slide. In slides with fewer patches, we perform random over-sampling.

*https://github.com/bytedance/ibot

2. Data

TG-GATEs transcriptomics pre-processing: The raw transcriptomics consists of microarrays (Affymetrix GeneChip) with 31,042 probes. Data were downloaded from the toxigates portal* that aggregates all omics data acquired as part of The Japanese Toxicogenomics Project [5]. All data followed probe-wise normalization using log2 fold change with respect to a control group. Log2 fold change quantifies the proportional difference, on a logarithmic scale, between the expression levels of a particular probe under two conditions: a control group (on average 22 slides per study in TG-GATEs) and a sample group (a defined set of compound, time and sacrifice). Each probe was then mapped to a unique gene identifier using SynGoPortal†, resulting in 13,404 gene expression measurements per sample. Finally, studies from the train set with compounds or chemicals known to induce liver injury were selected (n=74) to extract the 1,000 genes with the largest log2 fold change, used for our analysis. The log2 fold change gene expression values were not further normalized before processing by the deep learning system. In total, we obtained 6,597 transcriptomic samples used for training.

Histology data overview: A summary of the liver data (TG-GATEs), Breast carcinoma (BRCA), and Lung carcinoma (NSCLC) is presented in Table 1, Table 2 and Table 3.

Table 1. **TG-GATEs data split overview.** Normal refers to benign WSIs without lesions. Positive refers to WSIs with lesions as reported by toxicologic pathologists.

	Samples	Normal	Positive
iBOT-Tox pre-training	47,227	–	–
TANGLE pre-training	6,597	5,204	1,393
Few-shot train	2,783	2,322	461
Independent test	4,584	3,858	726

3. Results

Lesion-wise TG-GATEs few-shot performance: To better understand TANGLE few-shot performance on TG-GATEs for lesion classification in rat liver, we provide per-lesion classification performance, namely, on cellular infiltration,

*https://toxygates.nibiohn.go.jp/toxygates/

†https://www.syngoportal.org/convert

Table 2. **BRCA data split overview.** All (S+E) pre-training slides were included for few-shot training.

	Samples	IDC	ILC
TANGLE pre-training	1,041	831	210
Few-shot train	1,041	831	210
Independent test	1,265	982	283

Table 3. **NSLCL data split overview.** All (S+E) pre-training slides were included for few-shot training.

	Samples	LUAD	LUSC
TANGLE pre-training	1,033	528	505
Few-shot train	1,033	528	505
Independent test	1,946	1,621	325

fatty change, (hepatocellular) hypertrophy, increased mitosis, (hepatocellular) necrosis, and proliferation of bile duct and oval cells. These lesions can take various sizes, *e.g.,* necrosis can be focal (located in a small region) or diffuse (scattered all over the tissue). Lesions can also have different morphologies, such as hepatocellular hypertrophy that can be accompanied by eosinophilic or basophilic degeneration. As presented in Table 5, large lesions such as fatty change and hypertrophy are easier to detect than smaller ones like cellular infiltration and necrosis. This may be due to the expression profiles not expressing focal lesions, for instance, because the amount of tissue that includes the lesion of interest is too small.

Loss ablation: We conduct three types of ablations on TG-GATEs: (1) ablation of the TANGLE loss, (2) ablation of the INTRA loss, and (3) experiments where we combine TANGLE and INTRA (see Figure 1).

First, we compare the symmetric contrastive objective of TANGLE with a one-sided objective (image $\rightarrow$ expression). Adding a symmetric loss leads to a consistent performance boost. We also tested with a mean-squared error (L2) and an L1 objective, both leading to a performance drop of 7.0% and 6.7% AUC, respectively. In addition, we compare the gain of using both a local-global and local-local contrastive alignment in INTRA. Both objectives bring complementary information and lead to a performance loss when only one is employed. Finally, we combine TANGLE objective with an INTRA objective based on contrasting the average token (Contrast w. Avg.) and a random view (Contrast w. Random View). Combining both leads to a performance drop of -2.0% AUC.

Model ablation: TANGLE uses an attention-based MIL (ABMIL) as backbone. We compare the performance of TANGLE when replacing it with TransMIL [8] (see Figure 2). This modification leads to a performance drop of 3.92% AUC. We hypothesize that (1) the tasks we focus on (TG-GATEs lesion classification and TCGA lung and breast subtyping) are predominantly morphological, thereby reducing the utility of context modeling, (2) ABMIL training can use larger batch sizes due to reduced memory requirements; and (3) our ABMIL implementation uses "modern tricks" such as a deeper pre-attention network and LayerNorm (see implementation).

Hyper-parameter search: Figure 3 presents a series of experiments with different hyper-parameters known to influence contrastive pre-training, namely, the batch size, the Softmax temperature, and the number of patches sampled per slide. Batches larger than 64 seem to perform equally well. Softmax temperatures that are too high lead to a severe performance drop. Finally, the number of tokens (or patches) sampled per slide has relatively little influence on the downstream few-shot performance.

4. Interpretability

Rank analysis: Following [4], we use the rank as a fast and cheap measure of the quality of the underlying latent space learned during SSL pre-training. Here, we compute the rank as the entropy of the d (assuming $d < n$) L1-normalized singular values of the slide embedding matrix $H \in \mathbb{R}^{n \times d}$. Specifically, we have:

$$\text{RankMe}(H) = \exp(-\sum_{k=1}^{d} p_k \log(p_k)) \ , \tag{1}$$

$$p_k = \frac{\sigma_k(H)}{|\sigma(H)|_1} + \epsilon \tag{2}$$

where σ_k denotes the $k-$th singular of H (sorted from large to low), and ϵ is small constant set to $1e-7$ for numerical stability. Figure 4 presents the smooth rank of the slide embeddings obtained with different methods on the three independent test cohorts.

Attention heatmaps: We also present attention heatmaps of TANGLE when pre-trained on breast (Figure 5, top) and lung (Figure 5, bottom). Interestingly, the attention is assigned to regions that overlap with tumor, a property that naturally emerges from multimodal pre-training without explicit training.

[†]50 or maximal available labeled samples per class

Table 4. **iBOT-Tox pre-training hyperparameters.** 8 × 80GB NVIDIA A100 GPUs were used for training. Batch size refers to the total batch size across GPUs.

Hyperparameter	Value
Layers	12
Heads	12
Patch size	16
Head activation	GELU
Embedding dimension	768
Drop path rate	0.1
Global crop scale	0.32, 1.0
Global crop number	2
Local crop scale	0.05, 0.32
Local crop number	10
Partial prediction shape	Block
Partial prediction ratio	0.0, 0.3
Partial prediction variance	0, 0.2
Gradient clipping	0.3
Normalize last layer	✓
Shared head	✓
AdamW β	(0.9, 0.999)
Batch size	1024
Freeze last layer epochs	3
Warmup epochs	5
Warmup teacher temperature epochs	30
Max epochs	80
Learning rate schedule	Cosine
Learning rate (start)	0
Learning rate (post warmup)	5e-4
Learning rate (final)	2e-6
Teacher temperature (start)	0.04
Teacher temperature (final)	0.07
Teacher momentum (start)	0.996
Teacher momentum (final)	1.000
Weight decay (start)	0.04
Weight decay (end)	0.4
Automatic mixed precision	fp16

Table 5. **Lesion-wise few-shot linear probing performance of TANGLE in rat liver.** TANGLE is tested on an independent test cohort comprising 4,584 slides, without any data leakage (slide- or study-level) from unimodal and multimodal pre-training. Average AUC and standard deviation are reported over five runs.

Lesion	k=1(↑)	k=5(↑)	k=10(↑)	k=25(↑)	k=50[†](↑)
Cellular infiltration	56.9 ± 14.5	60.3 ± 14.1	69.8 ± 2.3	71.5 ± 3.2	74.9 ± 3.7
Fatty change	74.6 ± 23.3	74.3 ± 21.5	89.8 ± 2.6	92.7 ± 1.8	94.6 ± 0.5
Hypertrophy	84.6 ± 7.7	86.3 ± 10.4	90.0 ± 2.5	92.1 ± 1.5	91.3 ± 1.8
Increased mitosis	75.5 ± 7.2	89.9 ± 2.9	89.7 ± 1.5	89.7 ± 1.1	89.7 ± 0.4
Necrosis	56.4 ± 15.8	75.6 ± 5.9	74.9 ± 6.3	79.8 ± 2.0	78.1 ± 2.8
Proliferation	84.4 ± 5.0	93.9 ± 2.6	94.0 ± 1.3	91.8 ± 2.3	92.8 ± 2.7
Mean	72.1 ± 11.6	80.1 ± 11.3	84.7 ± 9.0	86.3 ± 7.9	86.9 ± 7.6

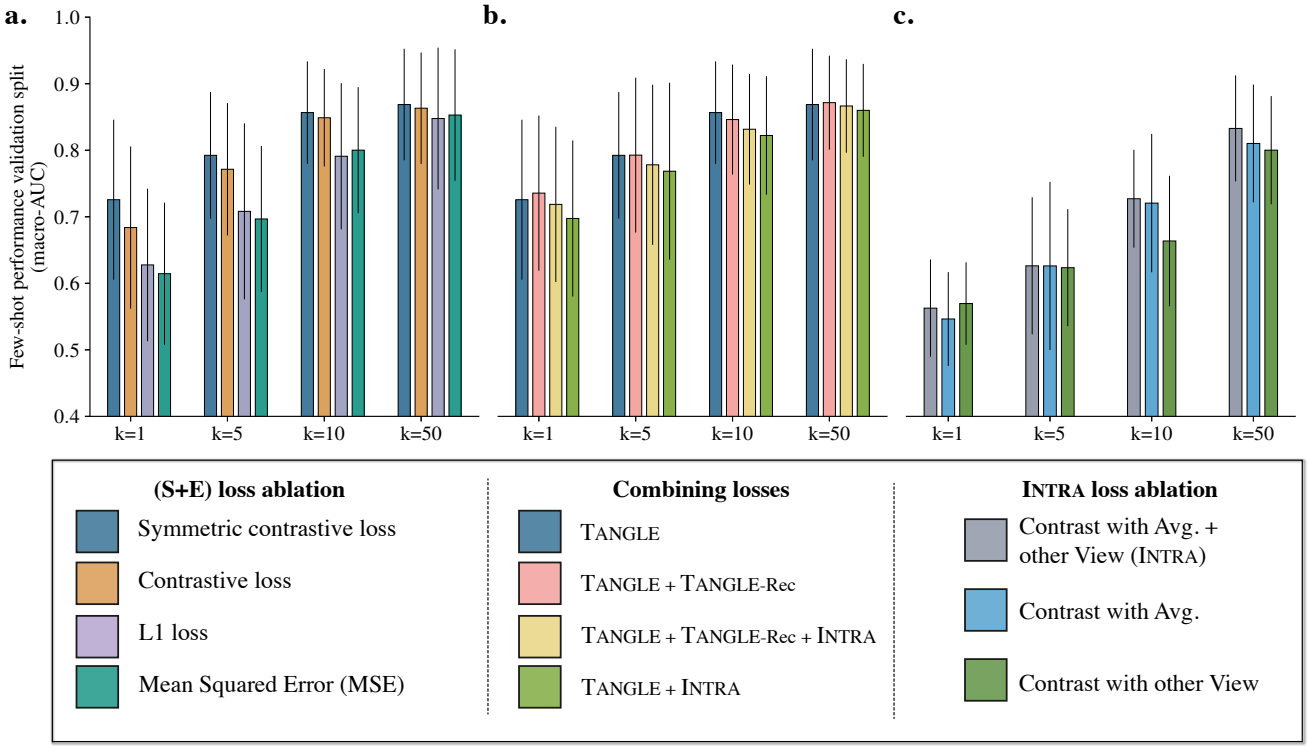

Figure 1. **Ablation study on TG-GATES. a.** Ablation of the (S+E) loss of TANGLE. We compare a symmetric contrastive loss with its non-symmetric counterpart, an L1 loss, and a Mean Squared Error loss. **b.** Combining TANGLE loss with TANGLE-Rec and INTRA. **c.** INTRA loss ablation using the average patch embedding, a random other view based on a different patch set, or a combination of both.

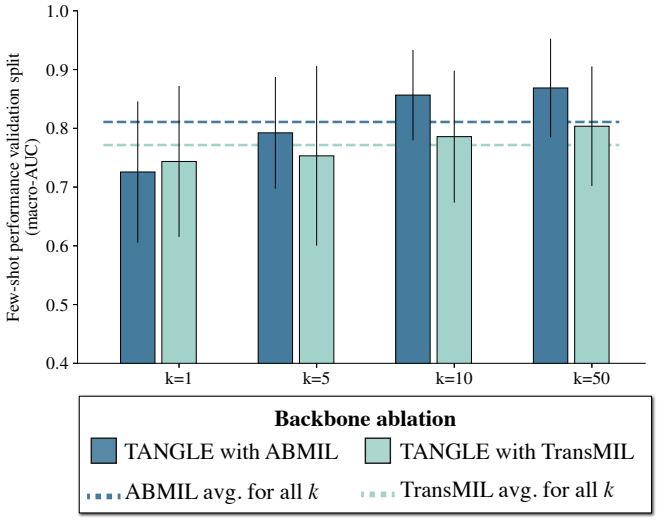

Figure 2. **Model ablation on TG-GATEs.** TANGLE training when replacing the ABMIL backbone by TransMIL.

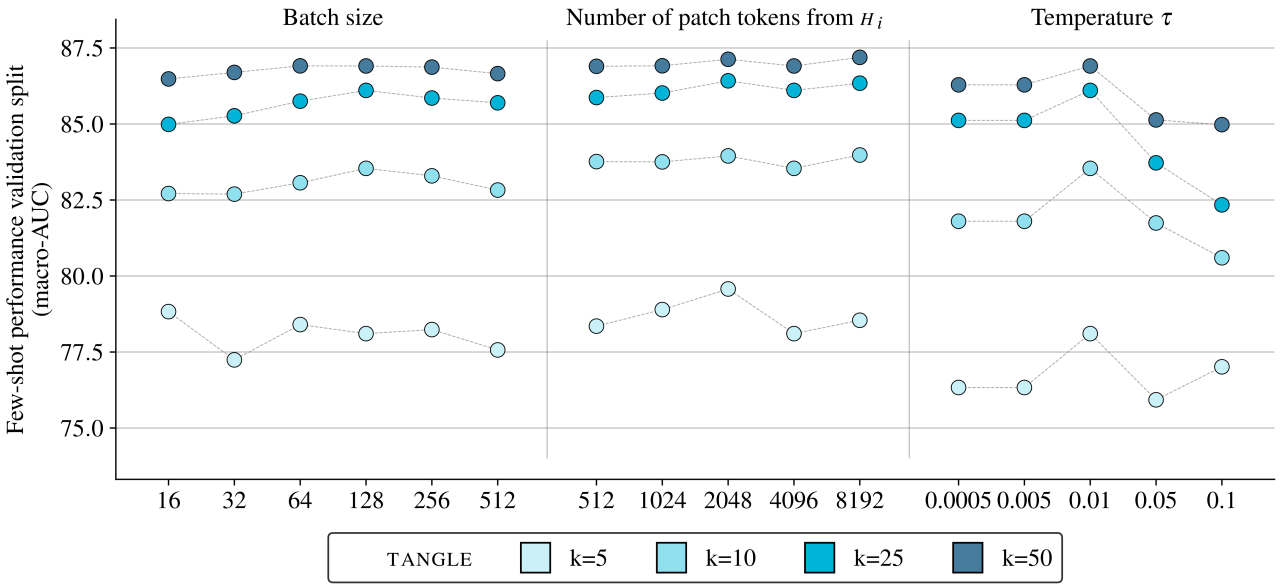

Figure 3. **Hyper-parameter search on TG-GATES.** We assess the influence of the batch size, number of patches sampled per slide, and the Softmax temperature.

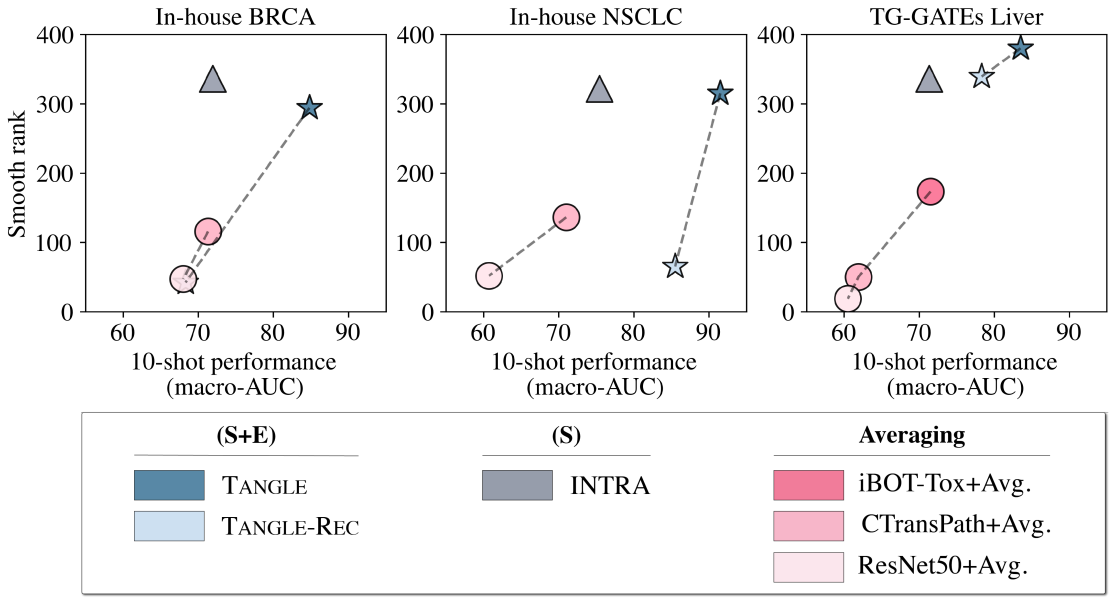

Figure 4. **Few shot performance *vs.* smooth rank.** TANGLE linear probing performance (k=10) and baselines, plotted against the smooth rank of the slide embedding matrix of the independent test cohorts. Test cohorts tested on BRCA subtyping (human breast, n=1,265 WSIs), NSCLC subtyping (human lung, n=1,946 WSIs), and TG-GATEs lesion classification (rat liver, n=4,584 WSIs). For each family of methods, we observe a strong positive correlation between performance and rank.

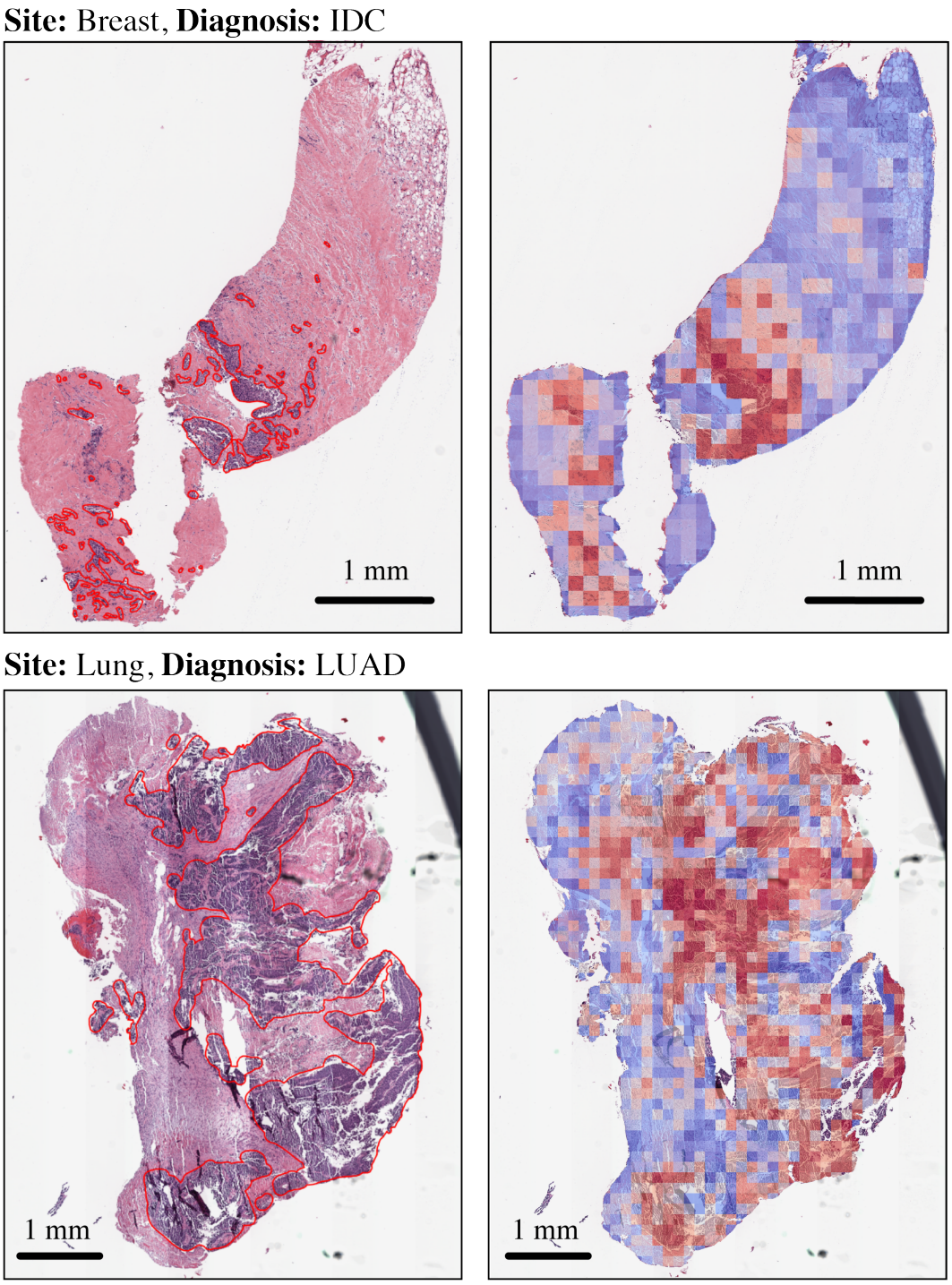

Figure 5. **TANGLE attention heatmaps of a lung and breast slide.** Attention weights of the (frozen) ABMIL slide encoder pre-trained with TANGLE overlaid on randomly chosen samples for our in-house cohorts. The network focuses mostly on tumor regions (marked in red) in both the breast and lung samples. This is a remarkable property of (S+E) pre-training as the network was not explicitly trained for tumor-related tasks, such as subtyping or grading.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2018. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[4] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, 2022. 2

[5] Yoshinobu Igarashi, Noriyuki Nakatsu, Tomoya Yamashita, Atsushi Ono, Yasuo Ohno, Tetsuro Urushidani, and Hiroshi Yamada. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research*, 43(D1):D921–D927, 2014. 1

[6] 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

[7] 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

[8] 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. 2

[9] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 1