Top-Down Attention-based Multiple Instance Learning for Whole Slide Image Analysis

Daniel Reisenbüchler[1,@], Ruining Deng[2], Christian Matek[3], Friedrich Feuerhake[4], and Dorit Merhof[1,5]

[1] Institute of Image Analysis and Computer Vision, University of Regensburg, Regensburg, Germany
[2] Weill Cornell Medicine, New York, USA
[3] Institute of Pathology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[4] Institute of Pathology, Hannover Medical School, Hannover, Germany
[5] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
[@]Correspondence: daniel.reisenbuechler@informatik.uni-regensburg.de

Abstract. Multiple instance learning (MIL) has become the de facto standard approach for whole-slide image analysis in computational pathology (CPath). While instance-wise attention tends to miss correlations between instances, self-attention can capture these interactions, but remains agnostic to the particular task. To address this issue, we introduce **T**op-**D**own **A**ttention-based **M**ultiple **I**nstance **L**earning (TDA-MIL), an architecture that first learns a general representation from the data via self-attention in an initial inference step, then identifies task-relevant instances through a feature selection module, and finally refines these representations by injecting the selected instances back into the attention mechanism for a second inference step. By focusing on task-specific signals, TDA-MIL effectively discerns subtle, yet significant, regions within each slide, leading to more precise classification. Extensive experiments on detecting lymph node metastasis in breast cancer, biomarker screening for microsatellite instability in different organs, and challenging molecular status prediction for HER2 in breast cancer show that TDA-MIL consistently surpasses other MIL baselines, underscoring the effectiveness of our proposed task-relevant refocusing and its broad applicability across CPath tasks. Our implementation is released at https://github.com/agentdr1/TDA_MIL.

Keywords: Computational Pathology · Multiple Instance Learning · Metastasis Detection· Molecular Status Prediction · Biomarker Screening

1 Introduction

Deep learning has led to significant advances in CPath, fuelled by the increasing use of whole-slide images (WSIs). However, WSIs can exceed $100{,}000 \times 100{,}000$ pixels at $20\times$ magnification ($1.5\,\mu m$/pixel), posing significant computational challenges. A common strategy is to partition WSIs into a sequence of manageable patches (e.g., 512×512 pixels), which are then processed into feature

representations in an offline fashion by a vision encoding model. Increasingly, self-supervised pretraining has shown promise in training domain-specific foundation models (FMs) [3], outperforming generic counterparts (e.g., ImageNet pretrained) [8]. However, transforming patch-level features into WSI-level predictions relies on MIL to identify and weight the most salient patches. Early MIL methods employ instance-wise attention [7] to highlight influential patches, but overlook contextual relationships. Recent efforts thus incorporate local neighborhood attention for stronger inductive bias [12], or apply global self-attention mechanisms to model interactions between patches [13,12,19]. Notably, a large-scale study involving a population of over 10,000 patients demonstrated that self-attention not only offers better performance but also faster convergence compared to instance-wise schemes, highlighting the importance of context in WSI analysis [19]. However, although self-attention excels at general context modeling, vanilla formulations in vision transformers tend to focus on broadly discriminative features rather than task-specific cues [14,15]. Moreover, pathologists in clinical practice first capture global information from WSIs and then focus on specific regions of interest, e.g., morphological features or objects, depending on the task. Motivated by the above technical and clinical observations, we propose *TDA-MIL*, a two-step framework that first contextualizes all patches via self-attention, then refines their representations through a dedicated feature selection module. The selected task-relevant tokens are reintroduced into the attention mechanism, further guiding the model to focus on task significant regions. This top-down infusion of task-specific context enables the network to effectively filter and contextualize up to thousands of instances, boosting the accuracy of the final WSI-level predictions. Our main contributions are summarized as follows:

(1) Novel two-step MIL architecture. We propose TDA-MIL, a coarse-to-fine MIL framework that (I) contextualizes all patches with self-attention and weights them through a feature-selection module, and (II) applies top-down re-attention to focus efficiently on task-specific entities.
(2) Performance & Interpretability. TDA-MIL consistently surpasses MIL baselines on diverse CPath benchmarks. Heat-maps show that the feature-selection module pinpoints biomarker-specific ducts overlooked by vanilla self-attention, providing clearer diagnostically relevant insights.

2 Method

We provide an overview of our TDA-MIL pipeline in Fig. 1. The framework consists of an offline *feature compression stage* and an online *aggregation stage* that leverages task-specific top-down attention strategy and the feature selection module. Below, we detail the core components of our algorithm.

(A) Feature Compression Stage. Following standard preprocessing, the WSI is segmented to remove background regions using Otsu's thresholding and tessellated at $20\times$ magnification into n smaller patches, where each $p_i \in \mathbb{R}^{512\times512\times3}$.

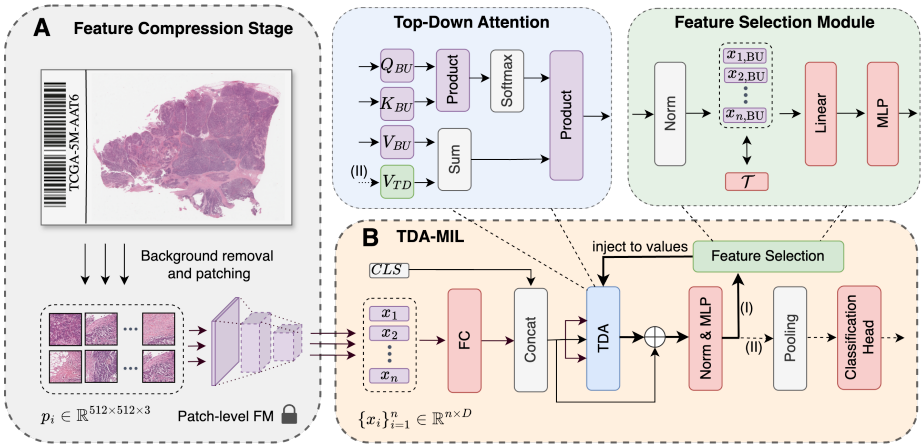

Fig. 1. Overview of the TDA-MIL pipeline. (A) The *offline* feature compression stage removes background areas, tessellates the WSI, and extracts patch-level features via a pathology foundation model. **(B)** The *online* aggregation stage then proceeds in two steps: (I) dimensionality reduction, self-attention-based contextualization, and feature selection, followed by (II) injecting the selected task-relevant features back into the self-attention mechanism for the final WSI-level prediction.

Each patch is then passed through a vision encoding FM, extracting feature embeddings in an offline fashion. Overall, this step reduces the high-dimensional WSIs to a tractable set of patch-level feature vectors $\{x_i\}_{i=1}^n \in \mathbb{R}^{n\times D}$, where D is the resulting latent dimension depending on the used FM.

(B) TDA-MIL. In the online stage, TDA-MIL processes patch features via two sequential inference steps. Given the sequence of n features $\{x_i\}_{i=1}^n$, we first project each feature from dimension D to a lower dimension d using a fully connected (FC) layer. Subsequently, a classification token CLS $\in \mathbb{R}^{1\times D}$ is concatenated to the sequence. In the following, the CLS is treated as other tokens and we continue to denote the sequence as length n for simplicity. We denote the resulting sequence as bottom-up sequence $\{x_{i,\mathrm{BU}}\}_{i=1}^n$.

Inference Step I:
The sequence $\{x_i\}_{i=1}^n$ is next feed through l blocks of self-attention (SA):

$$\mathrm{SA}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

with queries $Q \in \mathbb{R}^{n\times d_k}$, keys $K \in \mathbb{R}^{n\times d_k}$, and values $V \in \mathbb{R}^{n\times d_v}$. These are computed from x by

$$Q = W_Q \cdot x, \quad K = W_K \cdot x, \quad V = W_V \cdot x, \tag{2}$$

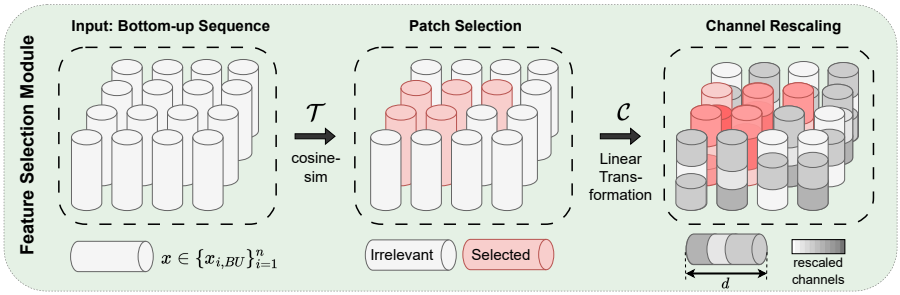

Fig. 2. Illustration of the internal behavior of the feature selection module. Each element (patch of WSI) of the input sequence is compared with a learnable task relevance token $\mathcal{T}$, and the linear transformation $\mathcal{C}$ performs tile-wise channel rescaling.

where $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$ are learnable parameters. Multi-head self-attention (MSA) applies self-attention in h parallel heads, then concatenates and linearly projects them:

$$\begin{aligned} \text{MSA} &= \text{concat}(\text{head}_1, ..., \text{head}_h) \cdot W_O, \\ \text{where head}_i &= \text{SA}(Q^{(j)}, K^{(j)}, V^{(j)}) \text{ for } j \in \{1, ..., h\} \end{aligned}$$

and $W_O \in \mathbb{R}^{hd_v \times d}$ is learnable. A self-attention layer begins with layer normalization, followed by the attention mechanism, and subsequentially with a multi-layer perceptron (MLP). Next, the sequence enters the feature selection module, which refines the output sequence $\{x_i\}_{i=1}^{n}$ by identifying task-relevant patches (dimension n) and channel rescaling (dimension d); see Fig. 2. The tile selection process is defined as

$$\hat{x}_{i,\text{BU}} = \text{clamp}\left(\text{sim}\left(x_{i,\text{BU}}, \mathcal{T}\right)\right), \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, $\text{clamp}\,(\cdot)$ restricts values to $[0, 1]$ and $\mathcal{T} \in \mathbb{R}^d$ is a learnable parameter token encoding task relevance. The top-down sequence is then computed as

$$x_{i,\text{TD}} = \mathcal{C} \cdot \hat{x}_{i,\text{BU}} \cdot x_{i,\text{BU}} \text{ for } i \in \{1, ..., n\} \tag{4}$$

where $\mathcal{C} \in \mathbb{R}^{d \times d}$ is learnable. As outlined in Fig. 2, the parameter $\mathcal{T}$ acts as a task embedding to filter irrelevant tiles in a weighted fashion, while $\mathcal{C}$ performs tile-wise channel rescaling. Finally, another MLP decodes the resulting feature sequence $\{x_{i,\text{TD}}\}_{i=1}^{n}$ before proceeding to the second inference step.

Inference Step II: The selected tiles $x_{i,\text{TD}}$ re-enter the self-attention (Equation 1) from *Inference Step I* for a second pass. To this end, the self-attention module receives the top-down input by adding them to the values, i.e. the values in Equation (2) are infused with information from the feature selection module

(4). Specifically, we add $x_{i,\mathrm{TD}}$ to the values V while leaving the queries Q and keys K unchanged,

$$V = W_V \cdot \big(x_{\mathrm{BU}} + x_{\mathrm{TD}}\big),$$

where x_{BU} is the bottom-up sequence as in the beginning of *Inference Step I*. After self-attention, the classification token CLS is fed to a final fully connected layer, transforming $\mathbb{R}^d$ to $\mathbb{R}^c$ for the prediction, where c is the number of classes.

3 Experiments

We evaluate the effectiveness of TDA-MIL across several key tasks CPath. Below, we introduce the datasets, baselines, evaluation and implementation details.

3.1 Tasks, Datasets and Annotations

Overall, our study comprises 3,719 publicly available WSIs stained from HE stained tissue of the organs breast, colon, rectum, stomach and uterine.

Metastases Detection. We use the CAMELYON17 dataset [1] for lymph node metastasis detection in breast cancer. Following [18,6], we binarize the classes to focus on metastasis presence versus absence (N=500; 182 positive, 318 negative). The challenge dataset is available at camelyon17.grand-challenge.org/Data/.
MSI Screening. We use TCGA-CRC [2] (N=457; 65 positive, 392 negative) and CPTAC-COAD [4] (N=105; 24 positive, 81 negative) to predict microsatellite instability status (MSI) in colorectal cancer. TCGA-STAD involves stomach cancer (N=361; 60 positive, 301 negative) and TCGA-UCEC uterine endometrial carcinoma (N=545; 117 positive, 428 negative). All TCGA datasets are publicly available at portal.gdc.cancer.gov/ and CPTAC at wiki.cancerimagingarchive.net.
Molecular Status Prediction HER2. We employ TCGA-BRCA (N=693; 158 positive, 535 negative) for molecular status prediction of human epidermal growth factor receptor 2 (HER2) in breast cancer. We further add the Breast Cancer Needle Biopsy dataset BCNB [20] (N=1,058; 277 positive, 781 negative) cohort, which is publicly available: paperswithcode.com/dataset/bcdalnmp.
Annotations. CAMELYON17 and BCNB provide their respective ground-truth labels. For TCGA and CPTAC, we matched clinical labels from cbioportal.org.

3.2 Comparable Methods

We benchmark TDA-MIL against comparable MIL methods, including AB-MIL [7] as a classic instance-wise attention baseline, CLAM [10] with clustering-constrained attention, DSMIL [9] employing a dual-stream attention strategy, LA-MIL [12] leveraging local attention, TransMIL [13] relying on transformer based aggregation, GPT [21] using graph transformers, RRT-MIL [17] focusing on feature re-embedding, MHIM [16] applying masked hard instance mining and S4MIL [5] adopting a structured state-space approach. Baselines are retrieved from their official repository and configured according to the published default settings. We used the same feature compression stage (Fig. 1A) for all methods.

3.3 Evaluation

We perform patient-stratified 5-fold cross-validations (CVs) for each task and report results using the area under the receiver operating curve (AUROC) and balanced accuracy (Bal. Acc), which accommodate class imbalances.

3.4 Implementation

Patch extraction is carried out with the CLAM library [10]. We employed the UNI FM [3] for feature extraction, which was not pre-trained on any dataset involved in this study and thus avoid any kind of potential data contamination from pretraining. UNI is based on a ViT-L architecture with output dimension 1024. We used a class-weighted rescaled cross-entropy loss to remedy uneven class distributions during the training process. Training is conducted with a batch size of 1 using the ADAM optimizer at a learning rate of 10^{-5} and weight decay of 10^{-2}, for a maximum of 100 epochs. The learning rate is reduced by a factor of 10 when performance plateaus for 5 subsequent epochs, and early stopping terminates training if performance fails to improve for 10 consecutive epochs. All experiments were performed on a single NVIDIA H100 GPU card.

4 Results

Table 1. Performance across different CPath Tasks and Datasets. Evaluation using AUROC and balanced accuracy (Bal. Acc) metrics. Mean and standard deviation are reported over patient-stratified CV runs. Best in **bold**, second best is underlined.

Task	**Metastases detection**		**HER2 Status**		**MSI Screening**	
Datasets	CAMELYON17		TCGA-BRCA BCNB		TCGA-CRC CPTAC-COAD	
Model/Metric	AUROC	Bal. Acc	AUROC	Bal. Acc	AUROC	Bal. Acc
AB-MIL [7]	94.68 ± 3.9	$\underline{93.35} \pm 4.8$	73.30 ± 1.3	64.82 ± 2.8	$\underline{91.21} \pm 1.5$	78.54 ± 5.7
CLAM [10]	$\underline{95.79} \pm 2.8$	93.10 ± 4.8	$\underline{73.29} \pm 1.2$	66.29 ± 2.4	90.87 ± 2.3	77.87 ± 4.8
DSMIL [9]	82.14 ± 13	78.78 ± 12	60.75 ± 6.7	59.71 ± 5.7	68.37 ± 18	65.97 ± 13
GPT [21]	92.60 ± 4.1	89.96 ± 5.7	67.21 ± 2.4	63.29 ± 1.9	82.02 ± 5.6	74.94 ± 7.3
LA-MIL [12]	85.81 ± 4.1	79.30 ± 5.9	72.51 ± 2.7	$\underline{67.27} \pm 2.7$	90.02 ± 3.0	$\underline{83.29} \pm 3.5$
MHIM [16]	95.69 ± 2.5	93.12 ± 3.1	72.36 ± 1.3	66.14 ± 1.7	89.61 ± 2.1	78.71 ± 5.2
RRT-MIL [17]	92.86 ± 2.9	88.75 ± 4.0	69.91 ± 1.4	63.37 ± 2.0	89.98 ± 1.8	80.56 ± 5.1
S4MIL [5]	90.04 ± 1.0	86.43 ± 1.1	71.45 ± 2.1	64.88 ± 2.0	89.86 ± 1.6	81.73 ± 5.6
TransMIL [13]	92.11 ± 1.4	87.08 ± 1.5	71.46 ± 1.8	62.95 ± 1.8	89.61 ± 1.6	77.24 ± 2.1
TDA-MIL	$\mathbf{97.20} \pm 1.2$	$\mathbf{93.38} \pm 1.8$	$\mathbf{73.66} \pm 1.1$	$\mathbf{68.62} \pm 2.9$	$\mathbf{91.78} \pm 2.3$	$\mathbf{86.45} \pm 1.7$

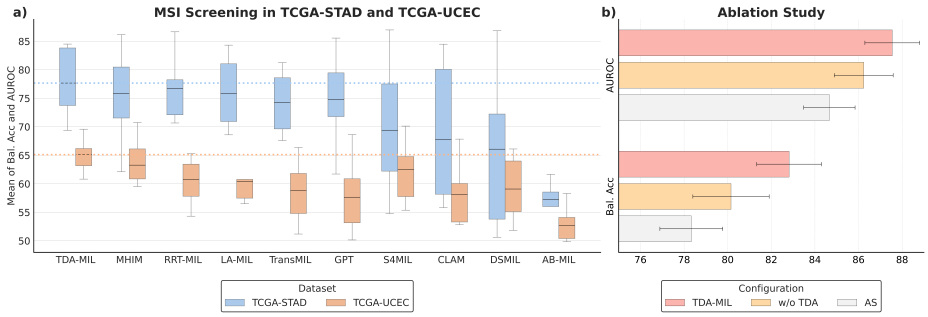

Fig. 3. Evaluation of MSI Screening on TCGA-STAD and TCGA-UCEC, and an Ablation Study. (a) MSI screening results are reported as mean of AUROC and balanced accuracy for each method; bars show patient-stratified cross-validation outcomes and dotted horizontal lines denote the average performance of TDA-MIL. **(b)** Ablation study: replacing the feature selection module with instance-wise attention (AS) and using pure self-attention, i.e., removing TDA (w/o TDA). Results are computed as the mean of the three tasks as in Table 1.

4.1 Performance Analysis

Table 1 reports classification performance across CPath tasks against baselines for CAMELYON17, Molecular status prediction of HER2 in TCGA-BRCA and BCNB, MSI Screening in TCGA-CRC and CPTAC-COAD. Further MSI Screening results in TCGA-STAD and TCGA-UCEC are visualized in Fig. 3a. In metastases detection, TDA-MIL achieves with 97.20 the best AUROC with +1.41 compared to the second best method CLAM. In Bal. Acc, TDA-MIL surpasses AB-MIL with a minor difference. For molecular status prediction HER2, CLAM and LA-MIL provide the second-best results in terms of AUROC and Bal. Acc, respectively; and TDA-MIL performs best overall. In MSI screening in colorectal cancer, TDA-MIL achieves a Bal. Acc of 86.45 with an improvement of 3.16 over the second best LA-MIL model and outperforms AB-MIL marginally in AUROC. However, for MSI prediction in stomach and uterine datasets RRT-MIL and MHIM are the second-best performing methods respectively, which are both outperformed by TDA-MIL. Overall, there is no single second-best baseline across all tasks, while TDA-MIL consistently surpasses the task-winning second-best approaches across various CPath tasks.

4.2 Ablation study

We measured the influence of feature selection and top-down attention in two folds: exchange feature selection with instance-wise attention (AS) as in AB-MIL and ablate top-down and feature selection module (w/o TDA). Fig. 3b) shows the results for exchanging the components or ablating components. Exchanging feature selection with instance-wise attention yields the worst performance across tasks. Ablating the TDA and feature selection part, i.e. using pure self-attention, decrease performance.

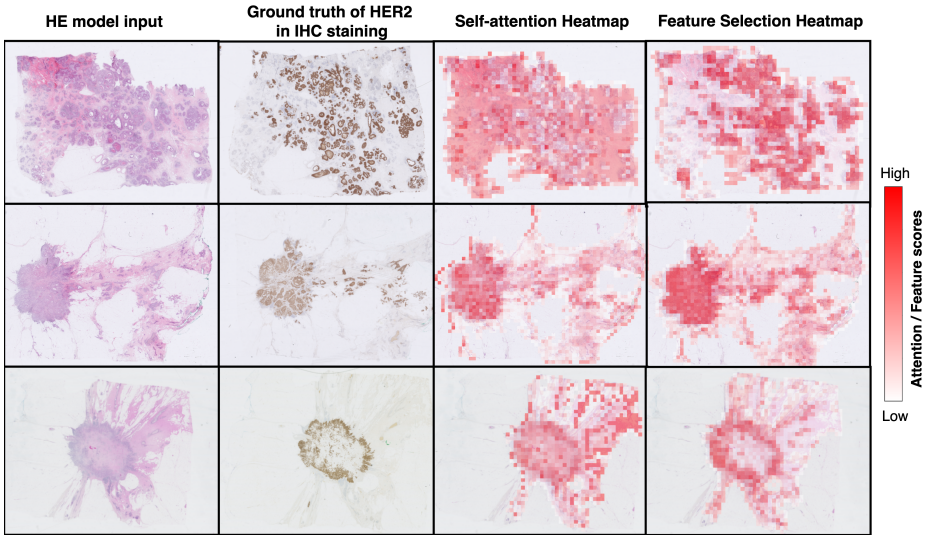

Fig. 4. Attention and feature selection heatmaps for molecular status HER2. Left to right: HE input WSI for TDA-MIL, ground truth IHC staining, heatmap of self-attention values without injection of feature selection values, heatmap of selected tiles from feature selection.

4.3 Qualitative Evaluation and Interpretability.

Fig. 3 shows an HE input for TDA-MIL, an IHC stained image of the same tissue for HER2, self-attention scores and task relevant tiles computed from the feature selection module. Samples in multi-stain are selected for clear IHC marker visibility from external ACROBAT dataset [11]. For the feature selection module, we visualized the normalized scores $\hat{x}_{i,\mathrm{BU}}$ as defined by Equation 3. It can be observed that more task-relevant areas are highlighted and less artifacts are visible in the heatmaps for selected features compared to self-attention values.

5 Conclusion

In this work, we introduced TDA-MIL, a top-down attention-based approach that addresses a limitation of self-attention when used in MIL strategies for whole-slide image analysis. By integrating a feature selection module with self-attention in a two-step inference procedure, TDA-MIL learns a robust general representation and then refocuses the model on task-relevant patches, thereby offering enhanced context modeling leading to refined discriminative power. Extensive evaluations in CPath tasks demonstrate its broad applicability, highlighting the value of task-specific refinement and underscore the potential of TDA-MIL as a generalizable framework for challenging gigapixel image analyses.

Acknowledgements. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaf, DFG) under project number 445703531. The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

Disclosure of Interests. The authors declare that they have no conflicts of interest related to this work.

References

1. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Cetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G.: From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. IEEE Trans. Med. Imaging **38**(2), 550–560 (Feb 2019)
2. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature **487**(7407), 330–337 (Jul 2012)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (Mar 2024). https://doi.org/10.1038/s41591-024-02857-3, http://dx.doi.org/10.1038/s41591-024-02857-3
4. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The CPTAC data portal: A resource for cancer proteomics research. J. Proteome Res. **14**(6), 2707–2713 (Jun 2015)
5. Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.H., Christodoulidis, S.: Structured state space models for multiple instance learning in digital pathology. In: Lecture Notes in Computer Science, pp. 594–604. Lecture notes in computer science, Springer Nature Switzerland, Cham (2023)
6. Fourkioti, O., De Vries, M., Bakal, C.: CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=rzBskAEmoc
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/ilse18a.html
8. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3344–3354 (June 2023)

9. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
10. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
11. Rantalainen, M., Hartman, J.: Acrobat - a multi-stain breast cancer histological whole-slide-image data set from routine diagnostics for computational pathology (2023). https://doi.org/10.48723/W728-P041, https://snd.se/catalogue/dataset/2022-190-1/1
12. Reisenbüchler, D., Wagner, S.J., Boxberg, M., Peng, T.: Local attention graph-based transformer for multi-target genetic alteration prediction. In: Lecture Notes in Computer Science, pp. 377–386. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-16434-7_37, https://doi.org/10.1007/978-3-031-16434-7_37
13. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
14. Shi, B., Darrell, T., Wang, X.: Top-down visual attention from analysis by synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2102–2112 (2023)
15. Shi, B., Gai, S., Darrell, T., Wang, X.: Toast: Transfer learning via attention steering. arXiv preprint arXiv:2305.15542 (2023)
16. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4078–4087 (October 2023)
17. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11343–11352 (June 2024)
18. Tourniaire, P., Ilie, M., Hofman, P., Ayache, N., Delingette, H.: Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images. Medical Image Analysis **85**, 102763 (Apr 2023). https://doi.org/10.1016/j.media.2023.102763, http://dx.doi.org/10.1016/j.media.2023.102763
19. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., Hutchins, G.G., Richman, S.D., Yuan, T., Langer, R., Jenniskens, J.C., Offermans, K., Mueller, W., Gray, R., Gruber, S.B., Greenson, J.K., Rennert, G., Bonner, J.D., Schmolze, D., Jonnagaddala, J., Hawkins, N.J., Ward, R.L., Morton, D., Seymour, M., Magill, L., Nowak, M., Hay, J., Koelzer, V.H., Church, D.N., Matek, C., Geppert, C., Peng, C., Zhi, C., Ouyang, X., James, J.A., Loughrey, M.B., Salto-Tellez, M., Brenner, H., Hoffmeister, M., Truhn, D., Schnabel, J.A., Boxberg, M., Peng, T., Kather, J.N., Church, D., Domingo, E., Edwards, J., Glimelius, B., Gogenur, I., Harkin, A., Hay, J., Iveson, T., Jaeger, E., Kelly, C., Kerr, R., Maka, N., Morgan, H., Oien, K., Orange, C., Palles, C., Roxburgh, C., Sansom, O., Saunders, M., Tomlinson, I.: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell **41**(9), 1650–1661.e4

(Sep 2023). https://doi.org/10.1016/j.ccell.2023.08.002, http://dx.doi.org/10.1016/j.ccell.2023.08.002
20. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in Oncology p. 4133 (2021)
21. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging **41**(11), 3003–3015 (Nov 2022)