SUPPLEMENTARY MATERIAL
HistoMILKD: A Multiple Instance Learning based Multi-Teacher Knowledge Distillation Framework for Whole Slide Image Classification

Mayur Mallya
mayur.mallya@ubc.ca

Ali Khajegili Mirabadi
ali.mirabadi@ubc.ca

Hossein Farahani
h.farahani@ubc.ca

Ali Bashashati
ali.bashashati@ubc.ca

University of British Columbia

1. Choice of Foundation Models

While the early histopathology foundation models (FM) in the literature such as CTransPath [11], Lunit-Dino [5], Phikon [3], etc. were trained on highly overlapping public datasets (mostly the TCGA database), the more recent FMs such as UNI, H-Optimus-0, Prov-GigaPath, Virchow2, and Hibou-L (all FMs used in this work) are pretrained on non-overlapping and private institutional cohorts. In Tab. 1, we present the heterogeneity of FMs in terms of the pretraining data cohorts used to train the FMs used in our work.

With such heterogeneous and non-overlapping pretraining data cohorts, each FM learns unique representations that are shaped by its own data distributions. This includes the distinct patient populations, staining protocols, and imaging pipelines. Leveraging multiple such FMs pretrained on heterogeneous datasets together (in the case of ensembles or **HistoMILKD**) can mitigate the risk of data biases which can occur if a FM is pretrained specifically in a demographic population or geographic region. This aggregation of diverse FMs ensures that the resulting model is more robust and generalizable than a single FM. This is further supported by our results in Tab. 1 of the main paper where the inclusion of multiple FMs in the training phase significantly improved the prediction performance.

Furthermore, we encourage the future works to extend this framework by including other diverse FMs in terms of data modality and framework. This includes FMs such as CONCH [7] and PLIP [4] which include textual data in its pretraining. Furthermore, the recent slide-level FMs such as PRISM [9] and TITAN [2] with multimodal pretraining cohorts are also an interesting direction for furthering the **HistoMILKD** framework, as the FMs can produce slide-level representations without needing an MIL adapter.

2. Experimental Details

2.1. Datasets

We evaluate the proposed method on five publicly available WSI classification datasets.

1. **TCGA BRCA** [12]: Breast Invasive Carcinoma subtyping between invasive ductal carcinoma (IDC, 837 slides) and invasive lobular carcinoma (ILC, 211 slides).
2. **TCGA NSCLC**: Non-Small Cell Lung Cancer subtyping between lung adenocarcinoma (LUAD, 530 slides) and lung squamous cell carcinoma (LUSC, 511 slides).
3. **TCGA RCC**: Renal Cell Carcinoma subtyping between clear renal cell (KIRC or CCRCC, 455 slides), papillary renal cell (KIRP or PRCC, 298 slides), and chromophobe renal cell carcinoma (KICH or CHRCC, 121 slides)
4. **OCEAN** [1]: Multi-institutional Ovarian Cancer subtyping between high-grade serous (HGSC, 216 slides), low-grade serous (LGSC, 41 slides), endometrioid (EC, 119), clear cell (CC, 94 slides), and mucinous carcinoma (MC, 41 slides)
5. **TCIA OBR** [10]: Ovarian Bevacizumab Response prediction between effective (157 slides) and invalid (108 slides) treatments.

2.2. Preprocessing

Given the gigapixel resolution of the WSIs, we extract $k = 500$ tissue patches of size 224×224 at 20X magnification from each WSI at random. The patches are stain normalized using the Macenko normalization technique [8] to account for the staining disparities across different WSIs. These patches are then forward-passed through pretrained and frozen FMs to extract WSI features $\mathbf{F}$ that form the inputs to the **HistoMILKD** model, as shown in Fig. 2 of the paper.

Foundation Model	Data size (WSIs)	Data source	Backbone	Method
UNI	100,000	Mass General Brigham	ViT-L	DINOv2
H-Optimus-0	500,000	Bioptimus proprietary dataset	ViT-G	DINOv2
Prov-GigaPath	171,189	Providence Healthcare	ViT-G	DINOv2 + Masked autoencoder
Virchow2	3,134,922	Memorial Sloan Kettering Cancer Center	ViT-H	DINOv2 + Regularization
Hibou-L	1,141,581	HistAI proprietary dataset	ViT-L	DINOv2

Table 1. Diversity of selected foundation models with respect to pretraining data cohort, model backbone, and the training framework. Data size and data source refer to the pretraining data the foundation models are trained on.

2.3. Training Setup

For each dataset, we first set aside 20% of the patients as the test set. With the remaining patients, we make three folds of training (60%) and validation (20%) sets and train three independent models, with all three models finally evaluated on the held-out test set. Additionally, for each fold, we train the models with 5 different random seeds for generalization purposes and report the averaged results across all 15 experiments (3 folds $\times$ 5 seeds).

2.4. Optimization

We use the Adam optimizer [6] with a learning rate and weight decay of 1×10^{-3} each. We use the loss weighting with $\alpha = 1$ and $\beta = 100$, inspired by [13] and further by the hyperparameter optimization as presented in Tab. 2. All models are trained for a total of 25 epochs and the model at the best validation performance epoch is used for testing on the held-out test set. To handle the class imbalance in our datasets, we use the balanced accuracy score as the metric to evaluate our models. All models in our experiments are trained using a combination of NVIDIA GeForce RTX 3090 and A6000 GPUs.

2.5. Choice of loss weights α and β

Inspired by the previous work, CAMKD [13], we use the similar set of hyperparameters in this work. While the previous work achieved best results with $\alpha = 1$ and $\beta = 50$, we observed that the choice of $\alpha = 1$ and $\beta = 100$ led to better results on the TCGA BRCA dataset. Tab. 2 shows the hyperparameter sensitivity of our framework for α and β.

Table 2. Hyperparameter optimization results on the TCGA BRCA dataset for optimal choice of α and β. The values denoted represent the balanced accuracy scores. *Nan* indicates the model failed to converge.

β	α			
	0.5	1	2	5
10	85.66	86.51	86.25	*Nan*
50	84.30	87.22	86.70	75.34
100	88.21	**88.65**	86.67	*Nan*
200	81.53	82.67	80.30	*Nan*

3. HistoMILKD training pseudocode

We provide the pseudocode for the training loop of **HistoMILKD** in Algorithm 1.

Algorithm 1 Simplified **HistoMILKD** training loop.

Require: teachers $T_1..T_K$ (frozen), student S; data batches $(X_{T_1}, .., X_{T_K}, X_S, y)$;

for each training batch **do**
  Teacher forwards:
  for $k = 1..K$: $(\hat{y}_{T_k}, h_{T_k}) \leftarrow T_k(X_{T_k})$
  Student forward: $(\hat{y}_S, h_S) \leftarrow S(X_S)$
  Supervised: $\mathcal{L}_{\text{CE}_S} \leftarrow \text{CE}(y, \hat{y}_S)$
  KD (teacher $\rightarrow$ student):
  $p_S \leftarrow \text{softmax}(\hat{y}_S),\ p_{T_k} \leftarrow \text{softmax}(\hat{y}_{T_k})$
  $\mathcal{L}_{\text{KL}}^{(k)} \leftarrow \text{KL}(p_{T_k} \,\|\, p_S)$
  $w_{T_k} \leftarrow H(p_{T_k})$
  $\mathcal{L}_{\text{KD}_S} \leftarrow \sum_{k=1}^{K} w_{T_k}\, \mathcal{L}_{\text{KL}}^{(k)}$
  Teacher-projection hint:
  for $k = 1..K$: $h'_{T_k} \leftarrow T_k(h_S)$
  $\mathcal{L}_{\text{MSE}_y} \leftarrow \sum_{k=1}^{K} \|h_{T_k} - h'_{T_k}\|_2^2$
  Aux CE via teacher heads:
  for $k = 1..K$: $\hat{y}'_{S,k} \leftarrow \text{Head}_{T_k}(h'_{T_k})$
  $\mathcal{L}_{\text{CE}_{S'}} \leftarrow \sum_{k=1}^{K} \text{CE}(y, \hat{y}'_{S,k})$
  Total: $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}_S} + \alpha\, \mathcal{L}_{\text{KD}_S} + \beta\, \mathcal{L}_{\text{MSE}_y} + \mathcal{L}_{\text{CE}_{S'}}$
  Update: ZEROGRAD(); BACKPROP($\mathcal{L}$); STEP()
end for

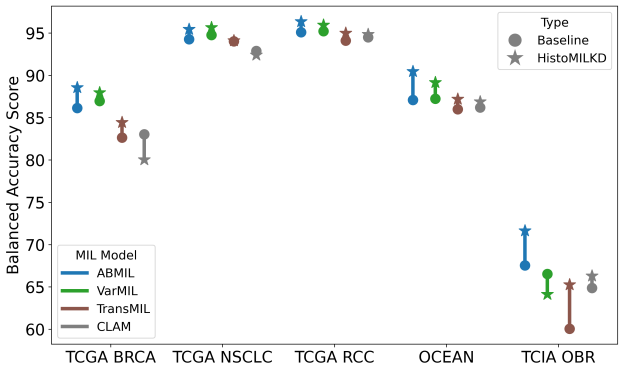

Figure 1. Analysis of MIL model performance in the **HistoMILKD**. In each case, we use UNI as the student and H-Optimus-0, Hibou, Prov-GigaPath, and Virchow2 as the teachers.

MIL	Size (M)	GFLOPs	Time (hrs)
ABMIL	1.97	31.79	0.55
VarMIL	2.95	32.10	0.63
TransMIL	14.66	657.55	2.32
CLAM	5.28	167.97	0.96

Table 3. Computational impact of choice of MIL head in the **HistoMILKD** framework. Size (M) and GFLOPs refer to the number of model parameters (in millions) and floating point operations (in billions) contributed by all the MIL heads in our framework. We also report the GPU training time (in hours) of our framework averaged across all 5 datasets with the respective MIL.

4. Additional analysis

4.1. Impact of MIL Model

In this section, we analyzed the impact of the MIL used in the **HistoMILKD** framework. Specifically, we compare the performance of our primary choice of MIL, ABMIL, against other popular MIL models in computational pathology literature, such as CLAM, VarMIL, and TransMIL. We present the results of this analysis across all five datasets in Fig. 1. In all the experiments under this section, we fix UNI as the student model and use 4 FM teachers (H-Optimus-0, Hibou-L, Prov-GigaPath, and Virchow2) for knowledge distillation. Additionally, we also compare the computational overhead of our framework caused by the choice of MIL model, the results of which are presented in Tab. 3.

We observed that ABMIL, despite being the smallest MIL among others, has the most consistent performance across all datasets in our framework. VarMIL, with a similar attention-based aggregation as ABMIL, has a better baseline performance compared to ABMIL on TCGA BRCA and NSCLC, but the improvements obtained from **HistoMILKD** with VarMIL are relatively lower. TransMIL, on the other hand, leads to a bulky framework and performs sub-optimally in our case. Similarly, CLAM does not suit our framework as it is highly sensitive to the model hyperparameters and often runs into convergence issues.

4.2. Interpretability

In this section, we provide more examples of the attention heatmaps. We also include the failure cases to highlight the limitations of our approach. Fig. 2 and Fig. 3 show the attention heatmaps of the cases where **HistoMILKD** improves the prediction over the student UNI model. On the other hand, Fig. 4 and Fig. 5 present the cases where our method fails to improve the prediction.

Additionally, we also include the UMAP feature space representations to compare the feature spaces of the student UNI model with that of the **HistoMILKD** UNI model. In Fig. 6, we present the UMAP comparison of the two methods across five datasets.

References

[1] Maryam Asadi-Aghbolaghi, Hossein Farahani, Allen Zhang, Ardalan Akbari, Sirim Kim, Ashley Chow, Sohier Dane, OCEAN Challenge Consortium, OTTA Consortium, David G Huntsman, et al. Machine learning-driven histotype diagnosis of ovarian carcinoma: insights from the ocean ai challenge. *medRxiv*, pages 2024–04, 2024. 1

[2] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*, 2024. 1

[3] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *MedRxiv*, pages 2023–07, 2023. 1

[4] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 1

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

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[7] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024. 1

[8] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology

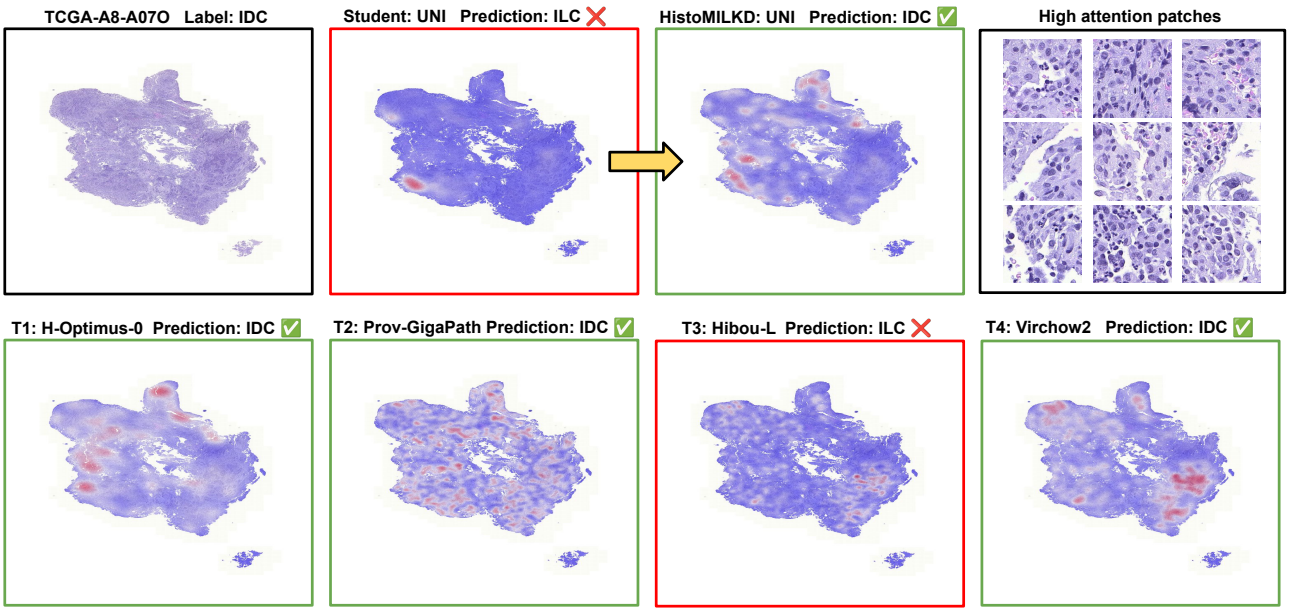

Figure 2. Attention heatmaps for the IDC case TCGA-A8-A07O, where **HistoMILKD** improves prediction over the student UNI. Top row compares the heatmaps of the baseline student with the **HistoMILKD** student, along with the high-attention patches from the latter model. Bottom row presents the heatmaps of the teachers (denoted by T1, T2, etc.). Green and red bounding boxes denote correct and wrong predictions respectively. We observe similarly distributed high-attention regions between the **HistoMILKD** student and the correctly predicting teachers (H-Optimus-0, Prov-GigaPath and Virchow2).

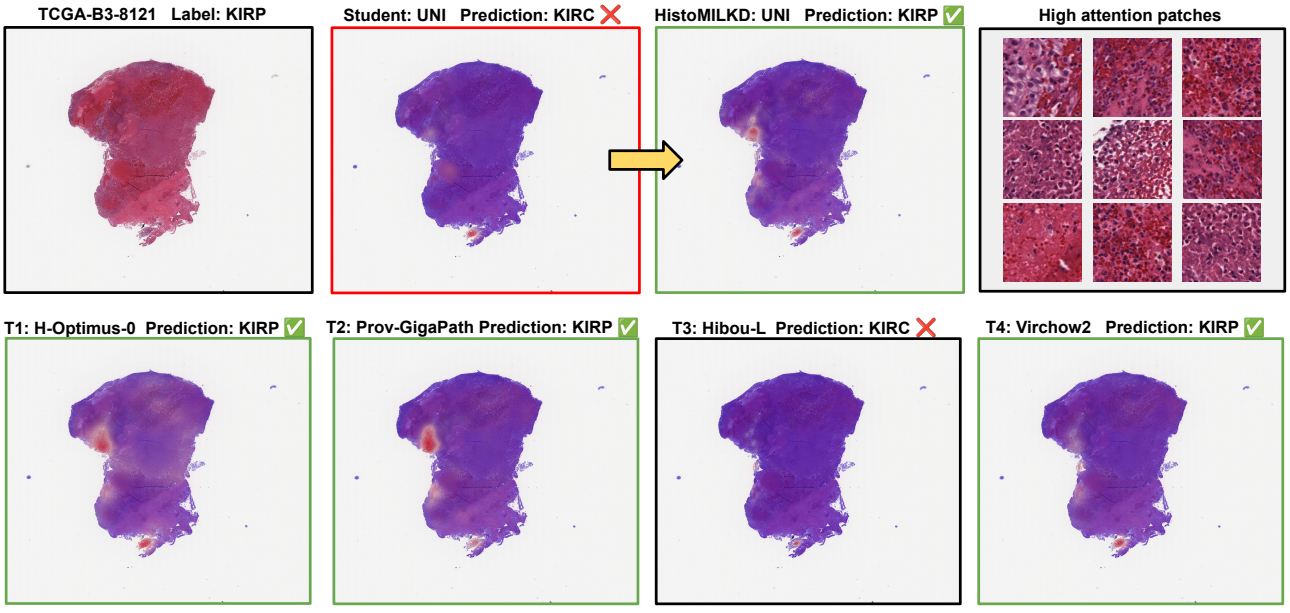

Figure 3. Attention heatmaps for the KIRP case TCGA-B3-8121, where **HistoMILKD** improves prediction over the student UNI. We observe similar high-attention regions between the **HistoMILKD** student and the correctly predicting teachers (H-Optimus-0, Prov-GigaPath and Virchow2). The incorrectly predicted models (UNI and Hibou-L) have highly concentrated attention regions and thereby miss out on the relevant tumor information.

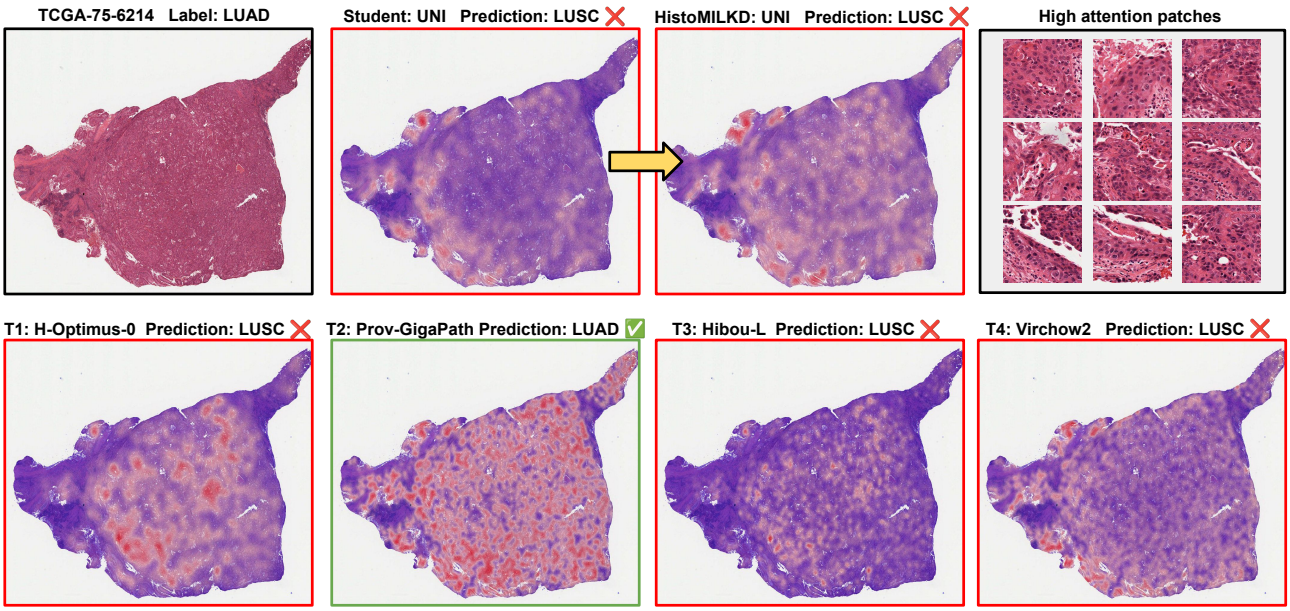

Figure 4. Attention heatmaps for the LUAD case TCGA-75-6214, where **HistoMILKD** fails to improve its prediction over the student UNI. We observe three of the four teachers also incorrectly predict the case to be LUSC and accordingly, the **HistoMILKD** UNI model predicts the case to be LUSC.

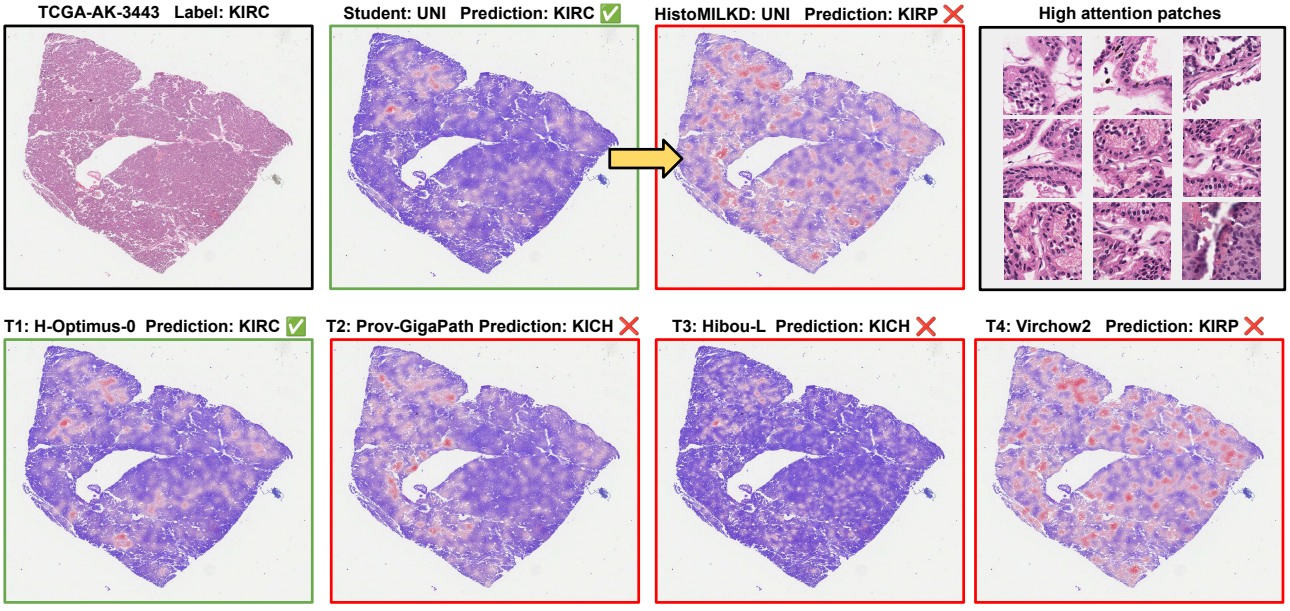

Figure 5. Attention heatmaps for the KIRC case TCGA-AK-3443, where **HistoMILKD** predicts the wrong subtype despite the baseline UNI model predicting the correct subtype. We observe three of the four teachers also incorrectly predict the case to be either KICH or KIRP and as a result of incorrect teachers, the **HistoMILKD** UNI model predicts the case incorrectly.

slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009. 1

[9] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*,

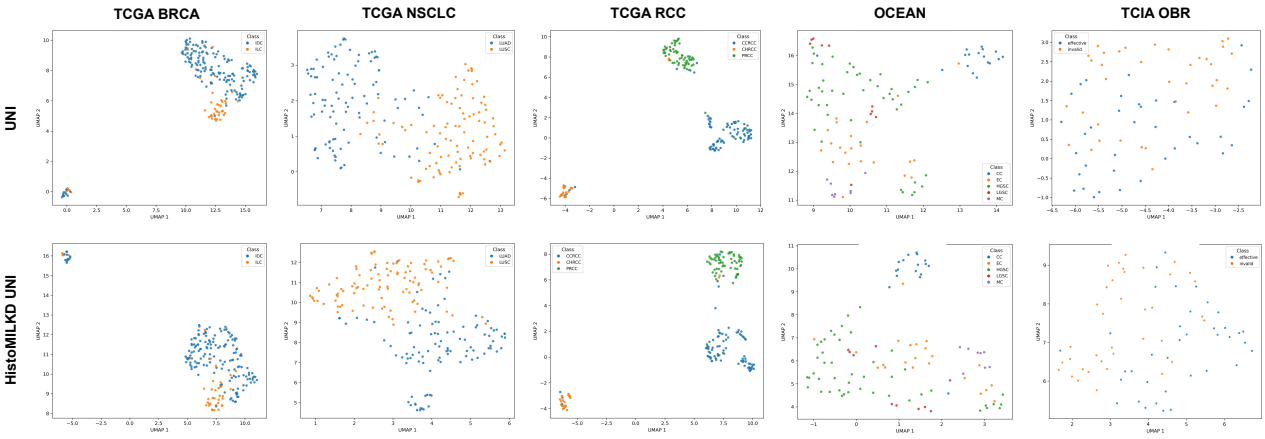

Figure 6. UMAP plots comparing the feature representations of the baseline UNI model (top row) with that of the **HistoMILKD** UNI model (bottom row) across five datasets. The similarity of embedding spaces between the two methods indicate that UMAP may not fully reflect the improvements achieved by **HistoMILKD**.

2024. 1

[10] Ching-Wei Wang, Cheng-Chang Chang, Yu-Ching Lee, Yi-Jia Lin, Shih-Chang Lo, Po-Chao Hsu, Yi-An Liou, Chih-Hung Wang, and Tai-Kuang Chao. Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images. *Computerized Medical Imaging and Graphics*, 99:102093, 2022. 1

[11] 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*, 81:102559, 2022. 1

[12] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 1

[13] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502. IEEE, 2022. 2