A. Experimentation Details

A.1. Dataset statistics

We perform our experiments on an internal stimulated Raman histology (SRH) dataset and light microscopy images from the publicly available TCGA dataset. Both datasets are split randomly into training and evaluation sets at the patient-level.

SRH dataset. The SRH dataset is collected using a commercially available NIO microscope (Invenio Imaging, inc, CA), following the protocol and descriptions in [41]. Our SRH dataset consists of 852K patches, 3560 whole-slide images from 896 patients in 7 different classes. The detailed training and evaluation set breakdown is listed in 4.

Tumor class	# Train Patients	# Train Slides	# Train Patches	# Eval Patients	# Eval Slides	# Eval Patches
HGG	149	541	139K	36	132	30K
LGG	93	313	139K	24	107	22K
Mening	154	533	130K	47	204	22K
Met	93	331	64K	24	114	17K
Pit	145	527	137K	46	194	30K
Schwan	15	47	10K	5	22	5K
Normal	99	343	82K	27	152	30K

Table 4. SRH dataset number of patients, slides, and patches breakdown of each class. HGG, high grade glioma; LGG, low grade glioma; mening, meningioma; met, metastasis; pit, pituitary adenoma; schwan, schwannoma; normal, normal brain tissue.

TCGA dataset. TCGA is a large-scale, multicenter consortium that includes biospecimens from 33 cancer types. We focus on brain tumor specimens from the TCGA-GBM and TCGA-LGG studies to evaluate HiDisc for diffuse glioma molecular genetic classification. The brain specimens in the TCGA dataset contain over 10 million patches from 29 institutions. Detailed dataset breakdown for each class is shown in Table 5.

IDH status	# Train Patients	# Train Slides	# Train Patches	# Eval Patients	# Eval Slides	# Eval Patches
Wt	367	732	4.7M	92	191	1.2M
Mut	336	626	4.4M	84	154	1.1M

Table 5. TCGA dataset number of patients, slides, and patches breakdown of each class. IDH, isocitrate dehydrogenase (IDH-1/2); wt, wildtype; mut, mutant.

A.2. Implementation

Augmentations. The strong augmentations we use in HiDisc training follow [9], and include the following augmentations applied sequentially, with a probability of 0.3 for each augmentation. Default PyTorch parameters are used, unless otherwise specified.

- Random horizontal and vertical flip;
- Gaussian noise;
- Color jittering;
- Random autocontrast;
- Random solarize with threshold 0.2;
- Random adjust sharpness with sharpness factor 2;
- Gaussian blur with kernel size 5 and sigma 1;
- Random erasing;
- Random affine transformation with max 10 degrees rotation and 10-30% image translation;
- Random resized crop.

Figure 6 demonstrates the random strong augmentations for the SRH and the TCGA datasets, respectively.

Filtering and Preprocessing. Following prior work, we divide all whole-slide images into 300×300 patches. We use a previously trained tumor segmentation model [18] to filter out blank and non-diagnostic patches from the SRH dataset. For the TCGA dataset, a heuristic algorithm based on the standard deviation of pixel values is used. TCGA patches are then normalized using the Macenko algorithm [40].

B. Extended Experimentation Metrics

We present our main results in Table 2. In addition to patch and patient-level evaluation, we compute slide-level metrics by aggregating the prediction on each whole slide image. For both SRH and TCGA experiments, we add area under the precision-recall curve (AUPRC), and for TCGA experiments, we also include sensitivity and specificity. The extended main results are in Tables 6 and 7 for SRH and TCGA, respectively. The metrics reported in these tables are consistent with Table 2, with slide and patient discrimination in HiDisc outperforming existing contrastive learning baselines across multiple different metrics and different levels. Confusion matrices are included in Figure 9.

C. Additional Ablation Studies

C.1. Weak augmentations

We also report the same additional metrics for our experiments with weak augmentation in Table 3. The extended results are in Tables 8 and 9 for SRH and TCGA datasets, respectively, and the confusion matrices are reported in Figure 10. While we observe a slight reduction in accuracy metrics at all levels for HiDisc, models trained using only instance discrimination, like SimCLR, collapse as expected because it fails to provide a meaningful pretext task to learn a good representation with weak augmentation. This can also be observed in the confusion matrices, where predic-

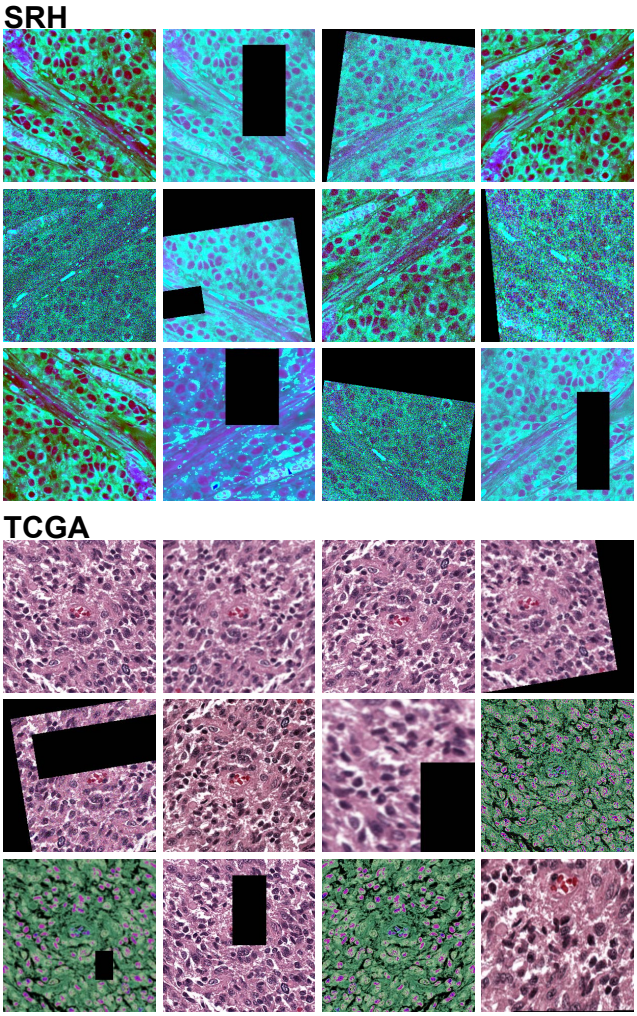

Figure 6. **SRH and TCGA augmentation panel.** Panels demonstrate examples of augmented image patches. Top left in each panel is the original, and the other patches are generated by randomly sampling from our set of strong augmentations. While these augmentations can help to regularize training and improve generalizability, they have been shown to decrease performance on some histopathology classification tasks.

tions from SimCLR and HiDisc-Patch are overwhelmingly in the majority classes.

C.2. Weighting factor lambdas

We conduct ablation on the effect of weighted factor λ at different level of discrimination. All experiments use the HiDisc-Patient with strong augmentations. HiDisc-Patient utilizes all levels of discrimination but is weighted by different λ values. For SRH experiments, we set one of the λ values to 0 or 5, and for TCGA experiments, we set one of the λ values to 0, 2, or 5, as well as setting two of the λ values to 0. The metrics are reported in Table 10 and 11 for SRH and TCGA, respectively. HiDisc is relatively robust to changes in λ_{Patient} and λ_{Slide} values, as slide- and patient-level discrimination are complementary to each other. As expected, when $\lambda_{\text{Patient}} = \lambda_{\text{Slide}} = 0$, we observe a significant reduction in model performance because only patch discrimination is used to supervise model training. Interestingly, we can see a slight performance drop when λ_{Patch} is amplified, and removing patch discrimination slightly boosted performance in SRH.

C.3. Learning rates

We evaluate model performance with different learning rates, and the model performances for SRH and TCGA datasets are reported in Table 12 and 13, respectively. The HiDisc performance on the SRH dataset vs learning rate is also summarized in Figure 7. We can observe that HiDisc training is robust to variation in learning rate, achieving good performance on the SRH dataset from 10^{-1} to 10^{-5}.

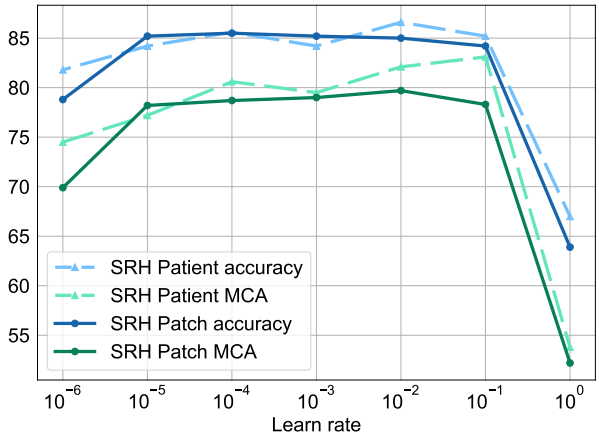

Figure 7. **Learning rate ablation.** HiDisc-Patch models are trained with batch size 512. We choose a wide range of learning rates and it shows HiDisc performs robustly from 10^{-1} to 10^{-5}. MCA, mean class accuracy.

C.4. Batch sizes

Literature for early contrastive learning algorithms such as SimCLR [9] shows benefits from training with larger batch size. We perform ablation studies to investigate the effect of batch size on HiDisc training. Due to the computation resources limit, we are only able to ablate batch size on 512 and 1024 for SRH dataset. The batch size is defined here as the total number of images including augmentation in one batch. As shown in Table 14, we do not observe a significant benefit of using a larger batch size.

C.5. Iterations

We present the training curve of HiDisc-Patient on the SRH dataset as an example in Figure 8 (showing validation

set metrics). This experiment uses batch size 512, learning rate as 10^{-3} and strong augmentation. We can observe HiDisc does not need long training time and achieve a performance plateau after 40K iterations of training.

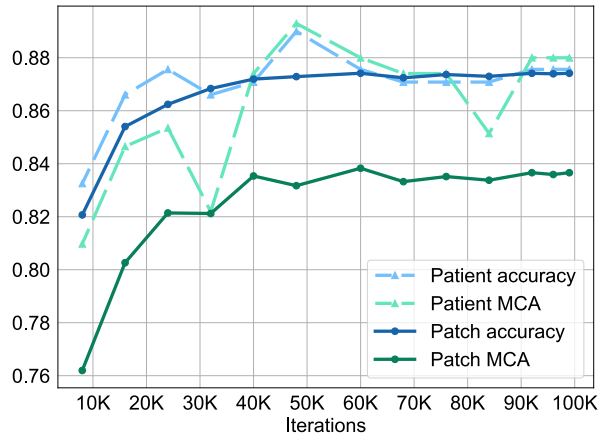

Figure 8. **Iterations ablation.** We empirically show the convergence of HiDisc until 100K iterations of training. HiDisc achieves a performance plateau after 40K iterations. MCA, mean class accuracy.

Method	**Patch** Level Metrics			**Slide** Level Metrics			**Patient** Level Metrics		
	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC
SimCLR	81.0 (0.1)	73.9 (0.2)	81.5 (0.2)	82.1 (0.3)	76.1 (0.3)	87.8 (0.2)	83.1 (0.7)	78.4 (0.6)	87.8 (0.2)
SimSiam	80.3 (1.9)	73.6 (2.7)	79.5 (3.7)	81.4 (1.8)	75.1 (2.6)	86.0 (3.5)	82.3 (1.7)	77.0 (4.0)	85.9 (3.2)
BYOL	83.5 (0.1)	78.2 (0.2)	84.8 (0.5)	84.3 (0.4)	79.9 (0.8)	90.5 (0.3)	84.8 (1.0)	82.7 (1.0)	90.8 (0.3)
VICReg	82.1 (0.3)	76.0 (0.4)	80.7 (0.7)	83.4 (0.8)	77.8 (1.0)	87.4 (0.5)	82.1 (0.7)	78.7 (1.9)	88.0 (0.4)
HiDisc-Patch	80.8 (0.0)	73.5 (0.1)	81.9 (0.0)	82.3 (0.2)	76.4 (0.2)	88.3 (0.2)	82.6 (0.3)	77.9 (0.3)	88.6 (0.2)
HiDisc-Slide	86.9 (0.2)	**83.2 (0.2)**	87.4 (0.6)	**88.1 (0.5)**	85.5 (0.3)	91.9 (0.6)	**87.6 (0.5)**	**87.0 (1.4)**	**90.4 (1.4)**
HiDisc-Patient	**87.4 (0.1)**	**83.5 (0.2)**	**88.7 (0.2)**	**88.5 (0.2)**	**86.2 (0.2)**	**92.8 (0.2)**	**87.9 (0.5)**	**86.4 (0.6)**	**92.3 (1.1)**
Supervised	88.9 (0.3)	86.3 (0.3)	90.8 (0.3)	89.0 (0.5)	88.8 (0.6)	93.9 (0.3)	88.5 (0.5)	89.1 (0.5)	93.6 (0.2)

Table 6. **Extended Main SRH Results.** Complete patch-, slide-, and patient-level metrics are shown. Slide-level metrics are aggregated using average pooling, similar to patient-level evaluation. Slide-level results are consistent with the patient-level metrics, showing HiDisc-Patient outperforms all other self-supervised learning baselines. We repeat experiments across three different random seeds, and standard deviations are reported in parentheses. MCA, mean class accuracy, AUPRC, area under the precision-recall curve.

	Method	Accuracy	MCA	Sensitivity	Specificity	AUROC	AUPRC
Patch level metrics	SimCLR	77.8 (0.0)	77.5 (0.0)	74.5 (0.2)	80.5 (0.2)	85.2 (0.1)	79.6 (0.3)
	SimSiam	68.4 (0.3)	68.0 (0.3)	64.3 (0.6)	71.7 (0.2)	74.1 (0.3)	65.4 (0.3)
	BYOL	80.0 (0.1)	79.8 (0.1)	77.7 (0.4)	81.9 (0.3)	87.5 (0.1)	82.0 (0.3)
	VICReg	75.5 (0.1)	75.2 (0.1)	72.8 (0.4)	77.6 (0.3)	82.9 (0.1)	75.5 (0.1)
	HiDisc-Patch	77.2 (0.1)	76.7 (0.1)	72.6 (0.1)	80.9 (0.1)	84.7 (0.1)	78.8 (0.2)
	HiDisc-Slide	82.7 (0.2)	82.5 (0.2)	80.5 (0.2)	**84.4 (0.2)**	89.3 (0.2)	85.0 (0.2)
	HiDisc-Patient	**83.1 (0.1)**	**83.0 (0.1)**	**81.9 (0.3)**	**84.2 (0.2)**	**90.1 (0.1)**	**86.2 (0.2)**
	Supervised	85.1 (0.3)	85.0 (0.3)	83.7 (0.6)	86.3 (0.1)	91.7 (0.2)	89.1 (0.2)
Slide level metrics	SimCLR	83.0 (0.3)	82.7 (0.3)	80.1 (0.6)	**85.3 (0.3)**	90.3 (0.2)	86.5 (0.4)
	SimSiam	77.2 (0.7)	76.6 (0.8)	71.3 (1.2)	81.9 (0.5)	82.9 (0.4)	73.4 (0.4)
	BYOL	84.1 (0.3)	84.0 (0.3)	83.5 (0.6)	**84.6 (0.5)**	91.6 (0.2)	88.2 (0.4)
	VICReg	80.8 (0.3)	80.6 (0.3)	78.3 (0.9)	82.8 (0.6)	88.7 (0.1)	83.5 (0.3)
	HiDisc-Patch	82.7 (0.2)	82.4 (0.2)	79.4 (0.3)	**85.3 (0.3)**	89.8 (0.2)	85.2 (0.4)
	HiDisc-Slide	**85.5 (0.4)**	**85.6 (0.4)**	**86.9 (0.9)**	84.4 (0.3)	**93.9 (0.1)**	**91.7 (0.6)**
	HiDisc-Patient	**85.1 (0.2)**	**85.2 (0.2)**	**86.4 (0.3)**	84.1 (0.3)	**93.8 (0.3)**	**91.7 (0.7)**
	Supervised	88.3 (1.3)	88.3 (1.3)	89.0 (1.7)	87.7 (1.3)	95.4 (0.1)	94.5 (0.2)
Patient level metrics	SimCLR	80.7 (0.6)	80.7 (0.6)	80.2 (1.0)	**81.3 (0.7)**	88.9 (0.3)	86.1 (0.4)
	SimSiam	76.6 (0.6)	76.5 (0.6)	73.9 (1.1)	79.0 (0.8)	82.4 (0.3)	73.3 (0.5)
	BYOL	83.1 (0.6)	83.3 (0.6)	86.6 (1.0)	80.0 (1.1)	89.8 (0.2)	86.2 (0.6)
	VICReg	77.0 (0.5)	77.0 (0.5)	77.6 (1.0)	76.3 (1.2)	86.0 (0.3)	79.9 (0.5)
	HiDisc-Patch	81.0 (0.5)	80.9 (0.5)	79.6 (0.4)	**82.2 (0.9)**	88.1 (0.2)	84.1 (0.5)
	HiDisc-Slide	**84.3 (0.3)**	**84.5 (0.3)**	**88.4 (0.8)**	80.6 (0.7)	**92.3 (0.3)**	**89.4 (0.8)**
	HiDisc-Patient	**83.6 (0.3)**	83.8 (0.3)	**87.6 (0.6)**	80.0 (0.6)	**91.8 (0.2)**	**89.2 (0.8)**
	Supervised	88.3 (0.4)	88.4 (0.4)	92.6 (0.8)	84.3 (0.6)	95.2 (0.2)	94.2 (0.9)

Table 7. **Extended Main TCGA Results.** For the binary molecular genetic classification task on TCGA dataset, we provide additional performance metrics, including sensitivity and specificity, as well as metrics at the whole-slide level. HiDisc maintains superior performance on all metrics across different levels compared to SSL baselines. We repeat experiments across three different random seeds, and randomly sampled 400 patches from each whole slide for nearest neighbor evaluation across three different random seeds. Standard deviations across nine evaluations are reported in parentheses. MCA, mean class accuracy, AUROC, area under the receiver operating characteristic curve, AUPRC, area under the precision-recall curve.

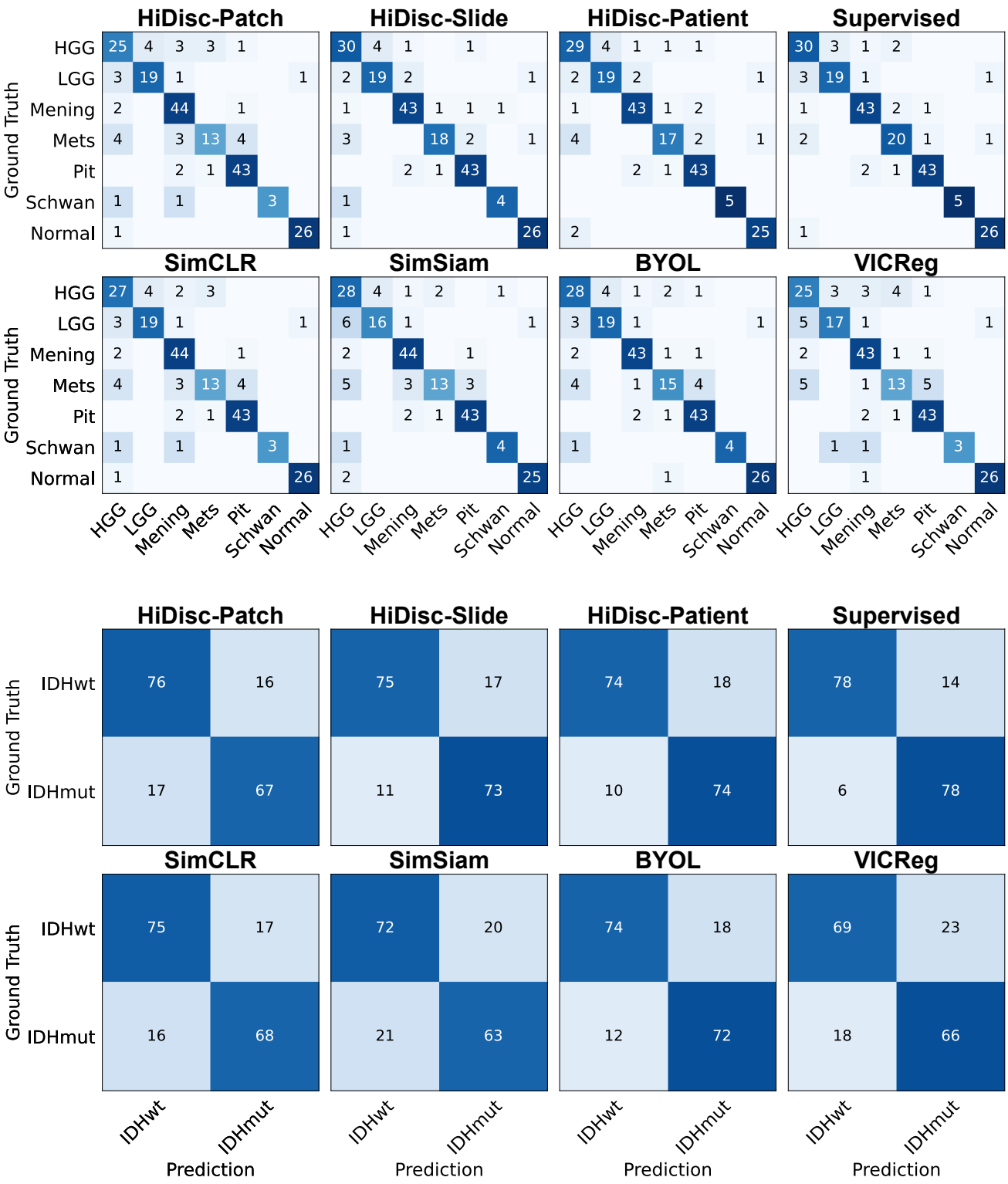

Figure 9. **Patient-level confusion matrices for main experiments.** These confusion matrices correspond to experiments reported in table 2. We can observe that HiDisc-Slide and HiDisc-Patient achieve superior performance compared to existing methods. HGG, high grade glioma, LGG, low grade glioma, mening, meningioma, mets, metastasis, pit, pituitary adenoma, schwan, schwannoma, normal, normal brain tissue, IDHwt, IDH wildtype, IDHmut, IDH mutant.

Method	**Patch** Level Metrics			**Slide** Level Metrics			**Patient** Level Metrics		
	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC
SimCLR	31.5 (2.3)	23.1 (1.9)	25.0 (2.3)	36.6 (4.1)	28.5 (2.7)	46.8 (4.3)	40.2 (6.9)	28.9 (4.5)	48.4 (3.8)
HiDisc-Patch	31.3 (0.6)	22.2 (0.5)	24.8 (1.1)	43.0 (1.6)	32.5 (1.1)	50.9 (3.8)	47.4 (2.1)	33.1 (1.6)	51.9 (2.3)
HiDisc-Slide	82.8 (0.2)	77.4 (0.3)	77.6 (0.3)	85.5 (0.3)	**81.1 (0.5)**	88.1 (0.2)	**84.2 (0.5)**	**82.3 (0.4)**	88.6 (0.6)
HiDisc-Patient	**84.9 (0.2)**	**78.9 (0.1)**	**81.9 (0.3)**	**86.6 (0.2)**	**81.7 (0.7)**	**89.8 (0.1)**	**84.7 (0.5)**	**80.9 (1.4)**	**90.3 (0.2)**
Supervised	90.0 (0.2)	87.4 (0.3)	85.4 (0.2)	91.0 (0.3)	90.7 (0.6)	93.5 (0.8)	90.0 (0.5)	90.3 (0.4)	93.2 (0.4)

Table 8. **Extended SRH Results with Weak Augmentations.** The weak augmentations here only use random vertical and horizontal flip. Without strong augmentation, the instance discrimination at patch level fails to provide a meaningful pretext task to learn meaningful representation. We only observe a 1-3 points drop as compared to strong augmentation on HiDisc-Slide and HiDisc-Patient, showing hierarchical discrimination reduces the reliance on the augmentation. We repeat experiments across three different random seeds, and standard deviations are reported in parentheses. MCA, mean class accuracy, AUPRC, area under the precision-recall curve.

	Method	Accuracy	MCA	Sensitivity	Specificity	AUROC	AUPRC
Patch level metrics	SimCLR	57.1 (1.1)	54.6 (1.0)	31.2 (4.6)	77.9 (4.7)	58.4 (2.2)	51.6 (1.9)
	HiDisc-Patch	59.0 (0.8)	57.2 (0.8)	40.0 (3.7)	74.4 (3.2)	61.5 (1.3)	54.3 (1.0)
	HiDisc-Slide	79.6 (0.1)	79.6 (0.1)	79.0 (0.2)	80.1 (0.1)	86.3 (0.2)	79.6 (0.4)
	HiDisc-Patient	**82.9 (0.2)**	**82.7 (0.2)**	**81.3 (0.2)**	**84.1 (0.2)**	**89.6 (0.2)**	**85.1 (0.4)**
	SupCon	85.4 (0.4)	85.3 (0.4)	83.9 (0.7)	86.7 (0.2)	92.0 (0.2)	89.6 (0.5)
Slide level metrics	SimCLR	58.7 (0.9)	54.3 (1.0)	12.8 (2.9)	95.7 (2.0)	70.2 (4.7)	62.5 (2.7)
	HiDisc-Patch	63.7 (2.5)	60.4 (2.9)	29.7 (6.7)	**91.2 (2.1)**	75.6 (3.4)	67.9 (2.5)
	HiDisc-Slide	81.0 (0.3)	81.0 (0.3)	81.4 (0.7)	80.6 (0.0)	88.6 (0.2)	83.4 (0.6)
	HiDisc-Patient	**85.0 (0.4)**	**85.2 (0.4)**	**86.9 (0.8)**	83.5 (0.5)	**92.6 (0.1)**	**89.4 (0.7)**
	SupCon	89.0 (0.7)	89.1 (0.8)	90.0 (1.2)	88.1 (0.6)	95.6 (0.2)	94.8 (0.4)
Patient level metrics	SimCLR	58.1 (1.1)	56.2 (1.1)	14.9 (2.5)	97.5 (1.2)	72.8 (3.2)	71.7 (2.0)
	HiDisc-Patch	61.2 (4.0)	59.7 (4.2)	25.3 (10.8)	**94.1 (3.4)**	75.8 (2.5)	71.8 (2.4)
	HiDisc-Slide	77.7 (0.4)	77.7 (0.4)	82.9 (0.8)	72.8 (0.0)	85.3 (0.3)	79.5 (0.7)
	HiDisc-Patient	**82.3 (0.3)**	**82.5 (0.3)**	**86.6 (0.5)**	78.4 (0.8)	**90.3 (0.3)**	**85.5 (1.3)**
	SupCon	88.4 (0.8)	88.6 (0.8)	92.7 (0.9)	84.5 (0.7)	95.2 (0.4)	94.3 (0.7)

Table 9. **Extended TCGA Results with Weak Augmentations.** The same weak augmentation experiment is conducted on TCGA dataset. Since this is a binary classification task, a random guess will have an accuracy of 50%. We observe SimCLR and HiDisc-Patch have a low accuracy close to random guessing, suggesting they collapse without learning meaningful representation. We repeat experiments across three different random seeds, and randomly sampled 400 patches from each whole slide for nearest neighbor evaluation across three different random seeds. Standard deviations across nine evaluations are reported in parentheses. MCA, mean class accuracy, AUROC, area under the receiver operating characteristic curve, AUPRC, area under the precision-recall curve.

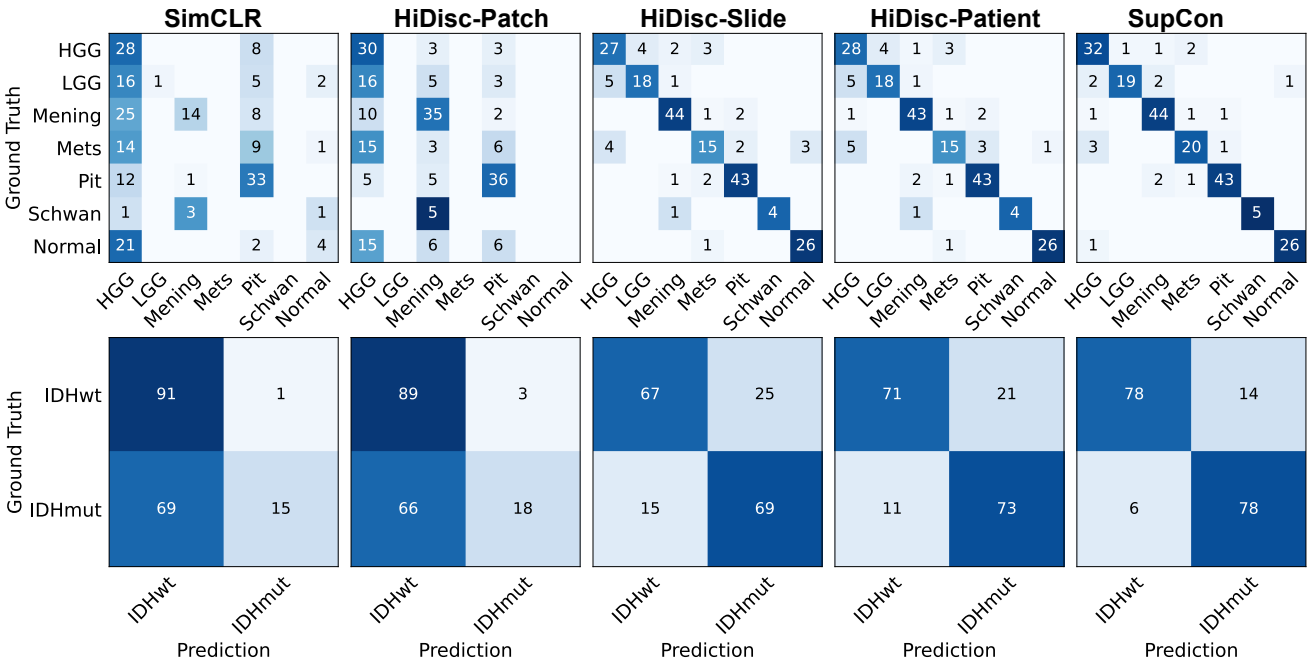

Figure 10. **Patient-level confusion matrices for experiments with weak augmentations.** These confusion matrices correspond to experiments reported in table 3. We can observe that HiDisc-Slide and HiDisc-Patient achieve superior performance compared to existing methods. As expected, patch discrimination methods, such as SimCLR and HiDisc-Patch, collapse because they fail to provide a meaningful pretext task to learn a good representation with weak augmentation. HGG, high grade glioma, LGG, low grade glioma, mening, meningioma, mets, metastasis, pit, pituitary adenoma, schwan, schwannoma, normal, normal brain tissue, IDHwt, IDH wildtype, IDHmut, IDH mutant.

			Patch Level Metrics			**Slide** Level Metrics			**Patient** Level Metrics		
λ_{Patient}	λ_{Slide}	λ_{Patch}	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC
1	1	0	**87.1**	**84.2**	87.1	**89.3**	**88.6**	**92.9**	**89.5**	**89.9**	**92.3**
1	0	1	86.7	82.4	88.0	88.3	85.8	92.5	86.6	87.0	91.9
0	1	1	86.7	81.6	**88.1**	88.0	84.4	92.4	87.6	83.0	92.2
1	1	5	86.8	81.8	**88.7**	87.1	82.4	92.2	86.6	82.1	91.9
1	5	1	**87.0**	82.6	87.4	**88.8**	85.8	**92.4**	**87.6**	85.5	91.2
5	1	1	86.8	**83.7**	86.6	88.4	**87.6**	**92.4**	86.1	**86.4**	**92.4**

Table 10. **Ablation study on λ weighting factor for the SRH dataset.** In these experiments, we changed one of the λ coefficients to 0 or 5. HiDisc is relatively robust to changes in λ_{Patient} and λ_{Slide} values, as slide and patient level discrimination are complementary to each other. Interestingly, we can observe a slight performance drop when λ_{Patch} is amplified, and removing patch discrimination slightly boosted performance. MCA, mean class accuracy, AUPRC, area under the precision-recall curve.

	λ_{Patient}	λ_{Slide}	λ_{Patch}	Accuracy	MCA	Sensitivity	Specificity	AUROC	AUPRC
Patch level metrics	1	1	0	**83.0 (0.1)**	**82.7 (0.1)**	80.6 (0.1)	**84.9 (0.1)**	89.3 (0.1)	84.7 (0.2)
	1	0	1	82.8 (0.1)	**82.7 (0.1)**	**81.4 (0.0)**	84.0 (0.1)	**90.2 (0.1)**	**86.6 (0.1)**
	0	1	1	82.4 (0.1)	82.3 (0.1)	80.8 (0.0)	83.7 (0.2)	**90.2 (0.1)**	**86.2 (0.3)**
	1	1	2	82.8 (0.1)	82.7 (0.1)	81.3 (0.0)	84.1 (0.2)	**90.2 (0.1)**	**86.2 (0.2)**
	1	2	1	83.1 (0.1)	**82.9 (0.1)**	81.1 (0.1)	**84.7 (0.1)**	**90.1 (0.1)**	**86.1 (0.2)**
	2	1	1	**83.3 (0.0)**	83.2 (0.0)	**82.1 (0.0)**	84.3 (0.1)	**90.1 (0.1)**	**86.1 (0.2)**
	1	1	5	82.2 (0.1)	82.0 (0.1)	80.6 (0.1)	83.4 (0.2)	89.6 (0.1)	85.5 (0.2)
	1	5	1	**83.0 (0.1)**	**82.8 (0.1)**	**81.2 (0.0)**	**84.4 (0.1)**	**90.0 (0.1)**	**86.2 (0.3)**
	5	1	1	**82.9 (0.0)**	**82.7 (0.0)**	80.8 (0.1)	**84.6 (0.1)**	89.5 (0.1)	85.2 (0.2)
	0	0	1	75.7 (0.1)	75.1 (0.1)	69.5 (0.2)	80.8 (0.0)	83.3 (0.1)	76.8 (0.1)
	0	1	0	**83.0 (0.0)**	**82.8 (0.0)**	**80.6 (0.1)**	**84.9 (0.1)**	**89.7 (0.1)**	**85.7 (0.1)**
	1	0	0	82.8 (0.1)	82.6 (0.1)	**80.5 (0.2)**	84.7 (0.0)	89.1 (0.1)	84.1 (0.1)
Slide level metrics	1	1	0	84.2 (0.2)	84.3 (0.2)	85.3 (0.4)	83.2 (0.0)	**93.7 (0.1)**	**91.0 (0.9)**
	1	0	1	**85.5 (0.0)**	**85.7 (0.0)**	**87.0 (0.0)**	84.3 (0.0)	**93.8 (0.1)**	**91.6 (0.4)**
	0	1	1	84.6 (0.0)	84.6 (0.0)	84.4 (0.0)	**84.8 (0.0)**	**93.7 (0.2)**	**91.7 (0.4)**
	1	1	2	**85.1 (0.4)**	**85.2 (0.4)**	**86.1 (0.4)**	**84.3 (0.5)**	93.6 (0.2)	**91.6 (0.5)**
	1	2	1	**85.1 (0.2)**	**85.2 (0.2)**	**86.1 (0.4)**	**84.3 (0.0)**	**93.7 (0.1)**	**91.4 (0.7)**
	2	1	1	84.6 (0.0)	84.8 (0.0)	**86.4 (0.0)**	83.2 (0.0)	**93.8 (0.1)**	**91.4 (0.8)**
	1	1	5	83.9 (0.2)	83.9 (0.2)	83.8 (0.0)	**83.9 (0.3)**	93.1 (0.1)	90.8 (0.4)
	1	5	1	**85.0 (0.2)**	**85.2 (0.2)**	**86.4 (0.0)**	**83.9 (0.3)**	**93.7 (0.1)**	**91.8 (0.4)**
	5	1	1	**84.9 (0.0)**	**85.1 (0.0)**	**86.4 (0.0)**	**83.8 (0.0)**	**93.8 (0.1)**	**91.6 (0.3)**
	0	0	1	81.7 (0.8)	81.2 (0.8)	76.6 (1.1)	**85.9 (0.5)**	89.0 (0.1)	82.6 (0.2)
	0	1	0	**85.4 (0.2)**	**85.5 (0.2)**	**86.8 (0.4)**	84.3 (0.0)	**93.8 (0.1)**	**91.6 (0.7)**
	1	0	0	**85.6 (0.2)**	**85.7 (0.2)**	**86.8 (0.4)**	84.6 (0.3)	93.6 (0.0)	**90.1 (0.6)**
Patient level metrics	1	1	0	81.4 (0.3)	81.6 (0.3)	84.9 (0.7)	78.3 (0.0)	91.6 (0.0)	**87.9 (0.9)**
	1	0	1	**84.1 (0.0)**	**84.3 (0.0)**	**88.1 (0.0)**	80.4 (0.0)	**91.9 (0.2)**	**89.3 (0.7)**
	0	1	1	83.7 (0.3)	83.8 (0.3)	86.1 (0.7)	**81.5 (0.0)**	**91.5 (0.3)**	**89.3 (0.9)**
	1	1	2	**83.3 (0.9)**	**83.5 (0.9)**	86.5 (0.7)	**80.4 (1.1)**	91.4 (0.3)	**88.9 (0.8)**
	1	2	1	**83.9 (0.3)**	**84.1 (0.3)**	**87.7 (0.7)**	**80.4 (0.0)**	91.4 (0.1)	**88.2 (0.5)**
	2	1	1	83.0 (0.0)	83.2 (0.0)	**88.1 (0.0)**	78.3 (0.0)	**92.0 (0.1)**	**88.6 (0.8)**
	1	1	5	82.0 (0.3)	82.1 (0.3)	84.5 (0.0)	**79.7 (0.6)**	90.9 (0.2)	**88.4 (0.5)**
	1	5	1	83.1 (0.3)	83.3 (0.3)	86.9 (0.0)	**79.7 (0.6)**	**92.0 (0.1)**	**89.2 (0.5)**
	5	1	1	**83.5 (0.0)**	**83.7 (0.0)**	**88.1 (0.0)**	**79.3 (0.0)**	91.6 (0.2)	**88.5 (0.5)**
	0	0	1	80.7 (1.7)	80.5 (1.7)	77.4 (2.4)	**83.7 (1.1)**	87.4 (0.2)	81.2 (0.4)
	0	1	0	**83.5 (0.0)**	**83.7 (0.0)**	**86.9 (0.0)**	80.4 (0.0)	**92.0 (0.1)**	**88.9 (0.7)**
	1	0	0	**83.7 (0.3)**	**83.8 (0.3)**	**86.5 (0.7)**	81.2 (0.6)	91.2 (0.0)	85.9 (0.3)

Table 11. **Ablation study on λ weighting factor for the TCGA dataset.** In these experiments, we changed one of the λ coefficients to 0, 2or 5, as well as changing two of the λ coefficients to 0. We randomly sample 400 patches from each whole slide for nearest neighbor evaluation across three different random seeds, and standard deviations are reported in parentheses. We can observe that HiDisc is relatively robust to changes in λ_{Patient} and λ_{Slide} values. As expected, when $\lambda_{\text{Patient}} = \lambda_{\text{Slide}} = 0$, we observe a reduction in model performance because only patch discrimination is used to supervise model training. MCA, mean class accuracy, AUROC, area under the receiver operating characteristic curve, AUPRC, area under the precision-recall curve.

	Patch Level Metrics			**Slide** Level Metrics			**Patient** Level Metrics		
LR	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC
1	63.9	52.2	57.2	68.5	55.8	67.4	67.0	53.8	72.8
1E-1	84.2	78.3	84.0	85.4	80.6	88.6	85.2	**83.1**	89.3
1E-2	85.0	**79.7**	82.5	**86.9**	**82.3**	89.5	**86.6**	82.1	88.9
1E-3	85.2	79.0	83.8	86.4	80.4	**90.3**	84.2	79.5	**90.7**
1E-4	**85.5**	78.7	**84.5**	86.2	80.6	89.6	85.6	80.6	89.8
1E-5	85.2	78.2	84.0	85.4	77.4	89.4	84.2	77.2	89.9
1E-6	78.8	69.9	76.4	80.5	72.1	82.9	81.8	74.5	85.9
1E-7	68.5	58.5	64.8	75.0	65.9	73.1	74.2	65.0	74.8

Table 12. **SRH Learn rate ablations.** We can observe that HiDisc training is robust to learning rate variations, achieving good performance from 10^{-1} to 10^{-5} in SRH. These experiments are performed with weak augmentations. MCA, mean class accuracy, AUPRC, area under the precision-recall curve.

	LR	Accuracy	MCA	Sensitivity	Specificity	AUROC	AUPRC
Patch level metrics	0.01	**83.1 (0.1)**	**82.9 (0.1)**	81.3 (0.1)	**84.6 (0.1)**	**89.9 (0.1)**	85.5 (0.1)
	0.001	**83.1 (0.1)**	**82.9 (0.1)**	**81.9 (0.1)**	84.0 (0.1)	**90.0 (0.1)**	**86.1 (0.1)**
	0.0001	82.0 (0.2)	81.9 (0.1)	81.1 (0.1)	82.7 (0.2)	89.5 (0.1)	85.4 (0.2)
Slide level metrics	0.01	**85.8 (0.0)**	**85.9 (0.0)**	**87.0 (0.0)**	**84.8 (0.0)**	**93.6 (0.1)**	**91.4 (0.4)**
	0.001	85.0 (0.2)	85.2 (0.2)	86.4 (0.0)	83.9 (0.3)	**93.5 (0.1)**	**91.2 (0.8)**
	0.0001	**85.2 (0.8)**	**85.3 (0.8)**	85.9 (1.4)	**84.6 (0.3)**	93.2 (0.2)	**90.6 (0.4)**
Patient level metrics	0.01	**84.1 (0.0)**	84.2 (0.0)	86.9 (0.0)	81.5 (0.0)	**91.5 (0.1)**	**88.4 (0.8)**
	0.001	83.7 (0.3)	83.9 (0.3)	**88.1 (0.0)**	79.7 (0.6)	**91.6 (0.2)**	**88.6 (0.5)**
	0.0001	**84.8 (0.7)**	**85.0 (0.7)**	87.7 (0.7)	**82.2 (0.6)**	90.8 (0.2)	87.9 (0.5)

Table 13. **TCGA learn rate ablations**. We can observe that HiDisc training is robust to learning rate variations, achieving good performance from 10^{-4} to 10^{-2} in TCGA. These TCGA experiments are performed with strong augmentations, and 400 patches are sampled randomly from each whole slide for nearest neighbor evaluation across three different random seeds, and standard deviations are reported in parentheses. MCA, mean class accuracy, AUROC, area under the receiver operating characteristic curve, AUPRC, area under the precision-recall curve.

	Patch Level Metrics			**Slide** Level Metrics			**Patient** Level Metrics		
Effective batch size	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC	Accuracy	MCA	AUPRC
512	**85.8**	**81.3**	**87.6**	**88.1**	**85.1**	**92.6**	**88.0**	**85.7**	**92.7**
1024	**85.8**	80.4	87.4	87.1	83.4	92.2	**88.0**	85.4	92.3

Table 14. **Batch size ablations.** We perform ablation studies to investigate the effect of batch size on HiDisc training. Due to the computation resources limit, we are only able to ablate batch size on 512 and 1024 for SRH dataset. We can observe that HiDisc training does not benefit from a larger batch size. Experiments in the table are performed without sync batch norm. MCA, mean class accuracy, AUPRC, area under the precision-recall curve.