Hierarchical discriminative learning improves visual representations of biomedical microscopy

Cheng Jiang\textsuperscript{1}\textsuperscript{*} Xinhai Hou\textsuperscript{1}\textsuperscript{*} Akhil Kondepudi\textsuperscript{1} Asadur Chowdury\textsuperscript{1} Christian W. Freudiger\textsuperscript{2} Daniel A. Orringer\textsuperscript{3} Honglak Lee\textsuperscript{1,4} Todd C. Hollon\textsuperscript{1}

\textsuperscript{1}University of Michigan \textsuperscript{2}Invenio Imaging \textsuperscript{3}New York University \textsuperscript{4}LG AI Research \textsuperscript{*}Equal Contribution

\{chengjia, xinhaih, tocho\}@umich.edu https://hidisc.mlins.org/

Abstract

Learning high-quality, self-supervised, visual representations is essential to advance the role of computer vision in biomedical microscopy and clinical medicine. Previous work has focused on self-supervised representation learning (SSL) methods developed for instance discrimination and applied them directly to image patches, or fields-of-view, sampled from gigapixel whole-slide images (WSIs) used for cancer diagnosis. However, this strategy is limited because it (1) assumes patches from the same patient are independent, (2) neglects the patient-slide-patch hierarchy of clinical biomedical microscopy, and (3) requires strong data augmentations that can degrade downstream performance. Importantly, sampled patches from WSIs of a patient’s tumor are a diverse set of image examples that capture the same underlying cancer diagnosis. This motivated HiDisc, a data-driven method that leverages the inherent patient-slide-patch hierarchy of clinical biomedical microscopy to define a hierarchical discriminative learning task that implicitly learns features of the underlying diagnosis. HiDisc uses a self-supervised contrastive learning framework in which positive patch pairs are defined based on a common ancestry in the data hierarchy, and a unified patch, slide, and patient discriminative learning objective is used for visual SSL. We benchmark HiDisc visual representations on two vision tasks using two biomedical microscopy datasets, and demonstrate that (1) HiDisc pretraining outperforms current state-of-the-art self-supervised pretraining methods for cancer diagnosis and genetic mutation prediction, and (2) HiDisc learns high-quality visual representations using natural patch diversity without strong data augmentations.

1. Introduction

Biomedical microscopy is an essential imaging method and diagnostic modality in biomedical research and clinical medicine. The rise of digital pathology and whole-slide images (WSIs) has increased the role of computer vision and machine learning-based approaches for analyzing microscopy data [51]. Improving the quality of visual representation learning of biomedical microscopy is critical to
introducing decision support systems and automated diagnostic tools into clinical and laboratory medicine.

Biomedical microscopy and WSIs present several unique computer vision challenges, including that image resolutions can be large (10K×10K pixels) and annotations are often limited to weak slide-level or patient-level labels. Moreover, even weak annotations are challenging to obtain in order to protect patient health information and ensure patient privacy [54]. Additionally, data that predates newly developed or future clinical testing methods, such as genomic or methylation assays, also lack associated weak annotations. Because of large WSI sizes and weak annotations, the majority of computer vision research in biomedical microscopy has focused on WSI classification using a weakly supervised, patch-based, multiple instance learning (MIL) framework [2, 7, 20, 37, 38, 48]. Patches are arbitrarily defined fields-of-view (e.g., 256×256 pixels) that can be used for model input. The classification tasks include identifying the presence of cancerous tissue, such as breast cancer metastases in lymph node biopsies [13], differentiating specific cancer types [7, 11, 18], predicting genetic mutations [11, 26, 32], and patient prognosis [8, 29]. A limitation of end-to-end MIL frameworks for WSI classification is the reliance on weak annotations to train a patch feature extractor and achieve high-quality patch-level representation learning. This limitation, combined with the challenge of obtaining fully annotated, high-quality WSIs, necessitates better methods for self-supervised representation learning (SSL) of biomedical microscopy.

To date, research into improving the quality and efficiency of patch-level representation learning without annotations has been limited. Previous studies have focused on using known SSL methods, such as contrastive learning [35, 47, 50], and applying them directly to WSI patches for visual pretraining. These SSL methods are not optimal because the majority use instance (i.e., patch) discrimination as the pretext learning task [5, 9, 10, 15, 55]. Patches belonging to the same slide or patient are correlated, which can decrease the learning efficiency. Instance discrimination alone does not account for patches from a common slide or patient being different and diverse views of the same underlying pathology. Moreover, previous SSL methods neglect the inherent patient-slide-patch data hierarchy of clinical biomedical microscopy as shown in Figure 1. This hierarchical data structure is not used to improve representation learning when training via a standard SSL objective. Lastly, most SSL methods require strong data augmentations for instance discrimination tasks [9]. However, strong and domain-agnostic augmentations can worsen representation learning in microscopy images by corrupting semantically important and discriminative features [21, 50].

Here, we introduce a method that leverages the inherent patient-slide-patch hierarchy of clinical biomedical microscopy to define a self-supervised hierarchical discriminative learning task, called HiDisc. HiDisc uses a self-supervised contrastive learning framework such that positive patch pairs are defined based on a common ancestry in the data hierarchy, and a combined patch, slide, and patient discriminative learning objective is used for visual SSL. By sampling patches across the data hierarchy, we introduce increased diversity between the positive examples, allowing for better visual representation learning and bypassing the need for strong, out-of-domain data augmentations. While we examine the HiDisc learning objective in the context of contrastive learning, it can be generalized to any siamese representation learning method [10].

We benchmark HiDisc self-supervised pretraining on two computer vision tasks using two diverse biomedical microscopy datasets: (1) multiclass histopathologic cancer diagnosis using stimulated Raman scattering microscopy [41] and (2) molecular genetic mutation prediction using light microscopy of hematoxylin and eosin (H&E)-stained cancer specimens [30]. These tasks are selected because of their clinical importance and they represent examples of how deep learning-based computer vision methods can push the limits of what is achievable through biomedical microscopy [18, 24, 26, 31]. We benchmark HiDisc in comparison to several state-of-the-art SSL methods, including SimCLR [9], BYOL [15], and VICReg [1]. We demonstrate that HiDisc has superior performance compared to other SSL methods across both datasets and computer vision tasks. Our results demonstrate how hierarchical discriminative learning can improve self-supervised visual representations of biomedical microscopy.

2. Related Work

Biomedical microscopy and computational pathology Biomedical microscopy refers to a diverse set of microscopy methods used in both clinical medicine and biomedical research. The most common clinical use of biomedical microscopy is light microscopy combined with H&E staining of clinical tissue specimens, such as tissue biopsies for cancer diagnosis [34]. The introduction of fast and efficient whole-slide digitization resulted in a rapid increase in the availability of WSIs and accelerated the field of computational pathology [42, 51]. Computational pathology aims to discover and characterize histopathologic features within biomedical microscopy data for cancer diagnosis, prognostication, and to estimate response to treatment.

The introduction of deep learning to WSI has resulted in clinical-grade computational pathology with diagnostic performance on par with board-certified pathologists [2, 18, 20]. See [43] for a comprehensive survey of deep learning-based methods in computational pathology. Ilse et al. presented MIL framework using an attention-based global pooling operation for slide-level inference [22]. Campanella et al.
Extended the strategy of a trainable aggregation operation using a recurrent neural network for gigapixel WSI classification tasks [2]. Lu et al. updated the attention-based MIL method to allow for better interpretability and efficient weakly supervised training using transformers [38]. HiDisc is complementary to any MIL framework and can be used as an effective self-supervised pretraining strategy.

Other biomedical microscopy methods have an increasing role in clinical medicine. Electron [39], fluorescence [44, 46, 53], and stimulated Raman scattering microscopy [14, 41] are a few examples of imaging methods that generate microscopy images used for patient diagnosis. Several studies have applied deep learning-based methods to these modalities for image analysis [18, 19, 23, 24, 44, 53].

Self-supervised pretraining in biomedical microscopy
Self-supervised pretraining has been used in computational pathology to improve patch-level representation learning [7, 28, 43, 45, 57]. Generally, a two-stage approach is used where first an SSL method is applied for patch-level feature learning using instance discrimination, and then the patch-level features are aggregated for slide- or patient-level diagnosis. SSL patch pretraining can reduce the amount of data needed compared to end-to-end MIL training [43]. Contrastive predictive coding [36], SimCLR [28], MoCo [45], VQ-VAE [6], and VAE-GAN [57] are examples of deep self-supervised visual representation learning methods applied to biomedical microscopy images [43].

Chen et al. presented a study using vision transformers and self-supervised pretraining at different image scales within individual WSIs [7]. They aim to represent the hierarchical structure of visual histopathologic tokens (e.g., cellular features, supracellular structures, etc.) at varying image resolutions using a transformer pyramid, resulting in a single slide-level representation. Knowledge distillation was used for SSL at each image resolution [5]. HiDisc is complementary to their method and can be used for SSL at any image resolution or, more generally, field-of-view pretraining for any MIL method for slide-level representations.

3. Methods
3.1. The Patient-Slide-Patch Hierarchy

The motivation for HiDisc is that fields-of-view from clinical WSIs, sampled from within a patient’s tumor, are a diverse set of image examples that capture the same underlying cancer diagnosis. Our approach focuses on how to use these diverse fields-of-view in the context of the known clinical patient-slide-patch hierarchical structure to improve visual representation learning. Most patients included in
public cancer histopathology datasets, including The Cancer Genome Atlas (TCGA) [4] and OpenSRH [24], contain multiple WSIs as part of their clinical cancer diagnosis. These WSIs may be sampled from different locations in the patient’s tumor, or different regions within the same tumor specimen. Both histopathologic and molecular heterogeneity has been well described within human cancers, encouraging clinicians to obtain multiple specimens/samples/views of the patient’s tumor [49]. To leverage the hierarchy shown in Figure 1, we create positive pairs at the patch-, slide-, and patient-level to define different discriminative learning tasks with a corresponding increase in visual feature diversity:

- **Patch discrimination**: Positive pairs are created from different random augmentations of the same patch. This strategy is similar to existing work on SSL via instance discrimination [1,9,10].

- **Slide discrimination**: Positive pairs are created from different augmented patches sampled from the same WSI. These pairs capture local feature diversity within the same specimen. Regional differences in cytologic and histoarchitectural features can be captured at this hierarchical level.

- **Patient discrimination**: Positive pairs are created from different WSIs from the same patient. Patches from different WSIs have the same underlying cancer diagnosis, but can have the greatest degree of feature diversity due to spatially separated tumor specimens. Additionally, diversity in specimen quality, processing, and staining, etc., is captured at this level.

An overview of the hierarchical discrimination tasks is shown in Figure 2.

### 3.2. Hierarchical Discrimination (HiDisc)

The formulation of the HiDisc loss function is based on NT-Xent [9] and inspired by [27, 56] for the purpose of multiple positive pairs. The fundamental difference is that no class labels are used during training with a HiDisc loss.

<table>
<thead>
<tr>
<th>Discrimination level</th>
<th>Number of samples treated as independent</th>
<th>Number of Positive pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patch</td>
<td>$n \cdot n_s \cdot n_p$</td>
<td>$n_a$</td>
</tr>
<tr>
<td>Slide</td>
<td>$n \cdot n_s$</td>
<td>$n_p \cdot n_a$</td>
</tr>
<tr>
<td>Patient</td>
<td>$n$</td>
<td>$n_s \cdot n_p \cdot n_a$</td>
</tr>
</tbody>
</table>

Table 1. **Batch composition at all discrimination levels.** The number of samples in the batch treated as independent and the number of positive pairs at each discrimination level. $n$, number of patients in the batch, $n_s$, number of slides sampled per patient, $n_p$, number of patches sampled per slide, $n_a$, number of augmentations performed on each patch.

HiDisc utilizes the natural hierarchy inherent to biomedical microscopy images to improve visual SSL.

We randomly sample a minibatch of $n$ patients, $n_s$ slides from each patient, $n_p$ patches from each slide, and we augment each patch $n_a$ times. For patients with less than $n_s$ slides, the slides are repeated and sampled. We assume $n_p <<$ numbers of patches available for WSI sampling. Note that if a patient has only one WSI, then patient discrimination degenerates to slide discrimination.

The HiDisc loss consists of the sum of three losses, each of which corresponds to a discrimination task at a different level of the patch-slide-patient hierarchy. Similar to the supervised contrastive learning loss [27], the component HiDisc losses are designed to fit more than one pair of positives within each level of the hierarchy. We define the HiDisc loss at the level $\ell$ to be:

$$L^\ell_{\text{HiDisc}} = \sum_{i \in I} \sum_{p \in \mathcal{P}_\ell(i)} \frac{1}{|\mathcal{P}_\ell(i)|} \log \frac{\exp (z_i \cdot z_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp (z_i \cdot z_a / \tau)},$$

(1)

where $\ell \in \{\text{Patch}, \text{Slide}, \text{Patient}\}$ is the level of discrimination, and $I$ is the set of all images in the minibatch. $\mathcal{A}(i)$ is the set of all images in $I$ except for the anchor image $i$,

$$\mathcal{A}(i) = I \setminus \{i\},$$

(2)

and $\mathcal{P}_\ell(i)$ is a set of images that are positive pairs of $i$ at the $\ell$-level,

$$\mathcal{P}_\ell(i) = \{ p \in \mathcal{A}(i) : \text{ancestry}_\ell(p) = \text{ancestry}_\ell(i) \},$$

(3)
where ancestry \(\ell\cdot\) is the \(\ell\)-level ancestry for an augmented patch in the batch. For example, patches \(x_i\) and \(x_j\) from the same patient would have the same patient ancestry, i.e., \(\text{ancestry}_{\text{Patient}}(x_i) = \text{ancestry}_{\text{Patient}}(x_j)\).

Patch-, slide-, and patient-level HiDisc losses share the same overall contrastive objective, but capture positive pairs at different levels in the hierarchy. Each loss has a different number of positive pairs within a minibatch. Details about this relationship are shown in Table 1.

Finally, the complete HiDisc loss is the sum of the patch-, slide-, and patient-level discrimination losses defined above:

\[
L_{\text{HiDisc}} = \sum_{\ell \in \{\text{Patch}, \text{Slide}, \text{Patient}\}} \lambda_{\ell} L_{\text{HiDisc}}^{\ell},
\]

where \(\lambda_{\ell}\) is a weighting hyperparameter for level \(\ell\) in the total loss. The pseudocode in PyTorch-style detailing the training process is shown in Algorithm 1.

### 4. Experiments

We evaluate HiDisc using two different computational histopathology tasks: multiclass histopathologic cancer diagnosis and molecular genetic mutation prediction. We present quantitative classification performance metrics, as well as qualitative evaluation with tSNE visualizations [52] of the learned patch-level representations.

#### 4.1. Datasets

**Stimulated Raman histology (SRH)** We validate HiDisc on a multiclass histopathological cancer diagnosis task using an SRH dataset. Stimulated Raman histology is an optical microscopy method that provides rapid, label-free, sub-micron resolution images of unprocessed biological tissues [14, 41]. The SRH dataset includes specimens from patients who underwent brain tumor biopsy or tumor resection. Patients were consecutively and prospectively enrolled at the University of Michigan for intraoperative SRH imaging, and this study was approved by the Institutional Review Board (HUM00083059). Informed consent was obtained for each patient prior to SRH imaging and approved the use of tumor specimens for research and development. The SRH dataset consists of 852K patches from 3560 slide images from 896 patients with classes consisting of normal brain tissue and 6 different brain tumor diagnoses: high-grade glioma (HGG), low-grade glioma (LGG), meningioma, pituitary adenoma, schwannoma, and metastatic tumor. All slides are divided into 300×300 patches, and they are preprocessed to exclude the empty or non-diagnostic regions using a segmentation model [18].

**TCGA diffuse gliomas** We additionally validate HiDisc using WSIs from The Cancer Genome Atlas (TCGA) dataset. We focus on WSIs from brain tumor patients diagnosed with diffuse gliomas, the most common and deadly primary brain tumor [34]. Molecular genetic mutation classification is used as the evaluation task. The most important genetic mutation that defines lower grade versus high grade diffuse gliomas is isocitrate dehydrogenase-1/2 (IDH) mutational status [3]. IDH-mutant tumors are known to have a better prognosis and overall survival (median survival ~10 years) compared to IDH-wildtype tumors (~1.5 years). The classification task is to predict IDH mutational status using formalin-fixed, paraffin-embedded H&E-stained WSI images at 20× magnification from the TCGA dataset. Predicting IDH mutational status from WSIs is currently not feasible for board-certified neuropathologists [3, 12]; genetic mutation prediction from WSIs could avoid time-consuming and expensive laboratory techniques like genetic sequencing. WSIs are divided into 300×300 pixel fields-of-view, and blank patches are excluded. The rest of the patches are stain normalized using the Macenko algorithm [40]. The TCGA data we included consists of a total of 879 patients and 1703 slides, and 11.3M patches.

#### 4.2. Implementation details

We train HiDisc using ResNet-50 [17] as the backbone feature extractor and a one-layer MLP projection head to project the embedding to 128-dimensional latent space. We use an AdamW [33] optimizer with a learning rate of 0.001 and a cosine decay scheduler after warmup in the first 10% of the iterations. For a fair comparison, we control the total number of images in one minibatch \(n \cdot n_{a} \cdot n_{s} \cdot n_{p}\) as 512 and 448 for SRH and TCGA data, respectively. We train HiDisc with patient discrimination by setting \(n_{a} = n_{s} = n_{p} = 2\), slide discrimination by setting \(n_{a} = n_{s} = 2, n_{p} = 1\), and patch discrimination by setting \(n_{a} = 2, n_{s} = n_{p} = 1\). The number of patients \(n\) sampled from each batch is adjusted accordingly. \(\lambda_{\ell}\) for each level of discriminating loss is set to 1, and temperature \(\tau\) is set as 0.7 for all experiments. We define weak augmentation as random horizontal and vertical flipping. The strong augmentations are similar as [9], including random erasing, color jittering, and affine transformation (for details, see Appendix A). We train HiDisc till convergence for both datasets (100K and 60K iterations for SRH and TCGA, respectively) with three random seeds. Training details for all baselines, including SimCLR [9], SimSiam [10], BYOL [15], VICReg [1], and SupCon [27], are similar to HiDisc.

#### 4.3. Evaluation protocols

**kNN evaluation** Standard protocols to evaluate self-supervised representation learning include linear and fine-tuning evaluation. However, both methods are sensitive to hyperparameters, such as learning rate [5]. Therefore, we use the \(k\) nearest neighbor (kNN) classifier for quantitative
<table>
<thead>
<tr>
<th></th>
<th>SRH - Patch</th>
<th></th>
<th>SRH - Patient</th>
<th></th>
<th>TCGA - Patch</th>
<th></th>
<th>TCGA - Patient</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>MCA</td>
<td>Accuracy</td>
<td>MCA</td>
<td>Accuracy</td>
<td>AUROC</td>
<td>Accuracy</td>
<td>AUROC</td>
</tr>
<tr>
<td>SimCLR [9]</td>
<td>81.0 (0.1)</td>
<td>73.9 (0.2)</td>
<td>83.1 (0.7)</td>
<td>78.4 (0.6)</td>
<td>77.8 (0.0)</td>
<td>85.2 (0.1)</td>
<td>80.7 (0.6)</td>
<td>88.9 (0.3)</td>
</tr>
<tr>
<td>SimSiam [10]</td>
<td>80.3 (1.9)</td>
<td>73.6 (2.7)</td>
<td>82.3 (1.7)</td>
<td>77.0 (4.0)</td>
<td>68.4 (0.3)</td>
<td>74.1 (0.3)</td>
<td>76.6 (0.6)</td>
<td>82.4 (0.3)</td>
</tr>
<tr>
<td>BYOL [15]</td>
<td>83.5 (0.1)</td>
<td>78.2 (0.2)</td>
<td>84.8 (1.0)</td>
<td>82.7 (1.0)</td>
<td>80.0 (0.1)</td>
<td>87.5 (0.1)</td>
<td>83.1 (0.6)</td>
<td>89.8 (0.2)</td>
</tr>
<tr>
<td>VICReg [1]</td>
<td>82.1 (0.3)</td>
<td>76.0 (0.4)</td>
<td>82.1 (0.7)</td>
<td>78.7 (1.9)</td>
<td>75.5 (0.1)</td>
<td>82.9 (0.1)</td>
<td>77.0 (0.5)</td>
<td>86.0 (0.3)</td>
</tr>
<tr>
<td>HiDisc-Patch</td>
<td>80.8 (0.0)</td>
<td>73.5 (0.1)</td>
<td>82.6 (0.3)</td>
<td>77.9 (0.3)</td>
<td>77.2 (0.1)</td>
<td>84.7 (0.1)</td>
<td>81.0 (0.5)</td>
<td>88.1 (0.2)</td>
</tr>
<tr>
<td>HiDisc-Slide</td>
<td>86.9 (0.2)</td>
<td><strong>83.2 (0.2)</strong></td>
<td><strong>87.6 (0.5)</strong></td>
<td><strong>87.0 (1.4)</strong></td>
<td>82.7 (0.2)</td>
<td>89.3 (0.2)</td>
<td><strong>84.3 (0.3)</strong></td>
<td><strong>92.3 (0.3)</strong></td>
</tr>
<tr>
<td>HiDisc-Patient</td>
<td><strong>87.4 (0.1)</strong></td>
<td><strong>83.5 (0.2)</strong></td>
<td><strong>87.9 (0.5)</strong></td>
<td><strong>86.4 (0.6)</strong></td>
<td><strong>83.1 (0.1)</strong></td>
<td><strong>90.1 (0.1)</strong></td>
<td><strong>83.6 (0.3)</strong></td>
<td><strong>91.8 (0.2)</strong></td>
</tr>
<tr>
<td>Supervised</td>
<td>88.9 (0.3)</td>
<td>86.3 (0.3)</td>
<td>88.5 (0.5)</td>
<td>89.1 (0.5)</td>
<td>85.1 (0.3)</td>
<td>91.7 (0.2)</td>
<td>88.3 (0.4)</td>
<td>95.2 (0.2)</td>
</tr>
</tbody>
</table>

Table 2. **Main results.** We compare HiDisc to state-of-the-art SSL. Supervised refers to models trained with supervision from class labels. Standard deviations are reported in parentheses. HiDisc-Slide and HiDisc-Patient outperform all baseline methods in all metrics for both tasks. Our HiDisc benchmark outperforms previously reported fully *supervised* baselines from the existing literature on the same genetic mutation classification task on TCGA [30]. MCA, mean class accuracy, AUROC, area under the receiver operating characteristic curve.

Figure 3. **Visualization of learned SRH representations using SimCLR and HiDisc.** Top. Randomly sampled patch representations are visualized after SimCLR versus HiDisc pretraining using tSNE [52]. Representations are colored based on brain tumor *diagnosis*. HiDisc qualitatively achieves higher quality feature learning and class separation compared to SimCLR. Expectedly, HiDisc shows within-diagnosis clustering that corresponds to patient discrimination. **Bottom.** Magnified cropped regions of the above visualizations show subclusters that correspond to individual patients. Patch representations in magnified crops are colored according to patient membership. We see patient discrimination within the different tumor diagnoses. Importantly, we do not see patient discrimination within normal brain tissue because there are minimal-to-no differentiating microscopic features between patients. This demonstrates that in the absence of discriminative features at the slide- or patient-level, HiDisc can achieve good feature learning using patch discrimination without overfitting the other discrimination tasks. HGG, high grade glioma; LGG, low grade glioma; Normal, normal brain tissue.
evaluation. We freeze the pretrained backbone to compute the representation vectors for both training and testing data, and the nearest neighbor classifier is used to match each patch in the testing dataset to the most similar $k$ patches in the training set based on cosine similarity. It also outputs a prediction score measured by the cosine similarity between each test image and its $k$ nearest neighbors. Due to the size of the TCGA dataset, we randomly sample 400 patches from each slide for evaluation across three different random seeds. Using the $k$NN classifier, we can compute accuracy (ACC), mean class accuracy (MCA), and area under receiver operating characteristic curve (AUROC) for patch metrics. We use MCA for SRH dataset because it is a multiclass classification problem and the classes are imbalanced. The AUROC is used for the TCGA dataset since it is a balanced binary classification task.

**Slide and patient metrics** In contrast to patch-level evaluation, slide and patient predictions are more practical for cancer diagnosis and other clinical uses [2]. After getting patch prediction scores by $k$NN evaluation, we use average pooling over the scores within each WSI or patient to obtain an aggregated prediction score. Compared to most MIL methods, this non-parametric method directly evaluates representation learning without additional training.

## 5. Results

### 5.1. Quantitative metrics

In this section, we evaluate the representations learned by self-supervised HiDisc pretraining using the training and evaluation protocols described in Section 4. We compare HiDisc and SSL baselines on the testing set of both datasets in Table 2. Since HiDisc-Patch is most similar to SimCLR, we observe similar performances. After we add slide and patient discrimination in HiDisc-Slide and HiDisc-Patient, we observe a significant boost in patch accuracy (+6.1 and +6.6 on SRH, +5.5 and +5.9 on TCGA), and a similar increase in other performance metrics as well. Among all methods, HiDisc-Patient has the best performance and outperforms the best baseline, BYOL, with an improvement of 3.9% and 3.1% in classification accuracy for SRH and TCGA, respectively. The GPU wall time needed to train BYOL is roughly 1.5x longer than HiDisc because it requires updating the weights of the target network using exponential moving average. Appendix B shows additional model evaluation metrics.

### 5.2. Qualitative evaluation

We also qualitatively evaluate the learned patch representations with $t$SNE [52] colored by class and patient label for both SimCLR and HiDisc. Figures 3 and 4 show the learned representations for the SRH and TCGA datasets, respectively. Here, we randomly sample patches from the validation set and plot them by class membership (tumor or molecular subtype). We observe that HiDisc learns better representations for both classification tasks, with more discernible clusters for each class. We also observe better patient clusters within each tumor class in the representations learned by HiDisc. Furthermore, patient clusters are not observed in normal brain tissues for both SimCLR and HiDisc-Patch, as there are minimal differentiating microscopic features between patients.

### 5.3. Ablation Studies

**Weak Augmentation** We demonstrate that HiDisc is capable of preserving excellent performance without the use of strong, domain-agnostic data augmentations as shown in Table 3. SimCLR suffers from dimensional collapse with weak augmentations [25]. Similar to SimCLR, HiDisc-Patch collapses because it is limited to the diversity from augmentations alone. HiDisc-Slide and HiDisc-Patient performance remain high across both datasets and tasks. HiDisc-Patient outperforms HiDisc-Slide, especially when evaluating at the patch level. We hypothesize that this is a result of additional diversity between positive pairs contributed by patient-level discrimination. We also provide supervised contrastive learning (SupCon) baselines [27] as an upper performance bound. Figure 5 shows SimCLR fails to learn semantically meaningful features while HiDisc achieves high-quality representations. Overall, we observe that HiDisc, especially HiDisc-Patient, performs well regardless of whether strong augmentations are used.

**Figure 4. Visualization of learned TCGA representations using SimCLR and HiDisc.** We randomly sample patches from the validation set, and visualize these representations using $t$SNE [52]. Representations on the plots are colored by IDH mutational status. Qualitatively, we can observe that HiDisc forms better representations compared to SimCLR, with clusters within each mutation that corresponds to patient membership. This observation is consistent with the visualizations for the SRH dataset in Figure 3.
Hierarchical self-supervised discriminative learning without strong data augmentations. Randomly sampled patch representations are shown after SimCLR and HiDisc-Patient pre-training without the use of strong, domain-agnostic data augmentations on SRH dataset. HiDisc achieves high-quality representation learning, while SimCLR is unable to learn semantically meaningful features via instance discrimination alone. HGG, high grade glioma; LGG, low grade glioma; Normal, normal brain tissue.

<table>
<thead>
<tr>
<th></th>
<th>SRH - Patch</th>
<th></th>
<th>SRH - Patient</th>
<th></th>
<th>TCGA - Patch</th>
<th></th>
<th>TCGA - Patient</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>MCA</td>
<td>Accuracy</td>
<td>MCA</td>
<td>Accuracy</td>
<td>AUROC</td>
<td>Accuracy</td>
<td>AUROC</td>
</tr>
<tr>
<td>SimCLR [9]</td>
<td>31.5 (2.3)</td>
<td>23.1 (1.9)</td>
<td>40.2 (6.9)</td>
<td>28.9 (4.5)</td>
<td>57.1 (1.1)</td>
<td>58.4 (2.2)</td>
<td>58.1 (1.1)</td>
<td>72.8 (3.2)</td>
</tr>
<tr>
<td>HiDisc-Patch</td>
<td>31.3 (0.6)</td>
<td>22.2 (0.5)</td>
<td>47.4 (2.1)</td>
<td>33.1 (1.6)</td>
<td>59.0 (0.8)</td>
<td>61.5 (1.3)</td>
<td>61.2 (4.2)</td>
<td>75.8 (2.5)</td>
</tr>
<tr>
<td>HiDisc-Slide</td>
<td>82.8 (0.2)</td>
<td>77.4 (0.3)</td>
<td>84.2 (0.5)</td>
<td>82.3 (0.4)</td>
<td>79.6 (0.1)</td>
<td>86.3 (0.2)</td>
<td>77.7 (0.4)</td>
<td>85.3 (0.3)</td>
</tr>
<tr>
<td>HiDisc-Patient</td>
<td>84.9 (0.2)</td>
<td>78.9 (0.1)</td>
<td>84.7 (0.5)</td>
<td>80.9 (1.4)</td>
<td>82.9 (0.2)</td>
<td>89.6 (0.2)</td>
<td>82.3 (0.3)</td>
<td>90.3 (0.3)</td>
</tr>
<tr>
<td>SupCon [27]</td>
<td>90.0 (0.2)</td>
<td>87.4 (0.3)</td>
<td>90.0 (0.5)</td>
<td>90.3 (0.4)</td>
<td>85.4 (0.4)</td>
<td>92.0 (0.2)</td>
<td>88.4 (0.8)</td>
<td>95.2 (0.4)</td>
</tr>
</tbody>
</table>

Table 3. Results for weak augmentation. Comparison of HiDisc with the SimCLR and SupCon on SRH and TCGA datasets with weak augmentation (random flipping). Both SimCLR and HiDisc-Patch patch collapse since weak augmentation will make the pretext task trivial. SupCon, on the other hand, is not sensitive to data augmentations since its positive pairs are defined by class labels. HiDisc-Slide and HiDisc-Patient are only slightly affected by the augmentation and achieve performance close to the supervised method.

Limitations Like other WSI classification methods, HiDisc representation learning is currently limited to single-resolution fields-of-view that are arbitrarily defined. Expanding HiDisc to include multiple image resolutions could improve its ability to capture multiscale image features of the patient’s underlying diagnosis. Also, we have limited the evaluation to a contrastive learning framework and HiDisc can also be evaluated using other siamese learning frameworks [1,5,10].

Broader Impacts We have limited this investigation to biomedical microscopy and WSIs. However, many imaging modalities, such as magnetic resonance imaging and fundoscopy [16], have a clinical hierarchical data structure that could benefit from a similar hierarchical representation learning framework. We hope that hierarchical discriminative learning will extend beyond microscopy to other medical and non-medical imaging domains.

Acknowledgements and Competing Interests

We would like to thank Karen Eddy, Lin Wang, Andrea Marshall, Eric Landgraf, and Alexander Gedeon for their support in data collection and programming.

This work is supported in part by grants NIH/NIGMS T32GM141746, NIH K12 NS080223, NIH R01CA226527, Cook Family Brain Tumor Research Fund, Mark Trauner Brain Research Fund: Zenkel Family Foundation, Ian’s Friends Foundation, and the Investigators Awards grant program of Precision Health at the University of Michigan.

C.W.F. is an employee and shareholder of Invenio Imaging, Inc., a company developing SRH microscopy systems. D.A.O. is an advisor and shareholder of Invenio Imaging, Inc, and T.C.H. is a shareholder of Invenio Imaging, Inc.
References


[36] Ming Yu Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-Supervised histology classifi-


