HASSL: Hierarchy-Aware Self-Supervised Learning Framework for Single Cell Microscopy

Julius Riel[†1,2,3], Vishwa Mohan Singh[†1,2], Sai Anirudh Aryasomayajula[1,2], Anuun Chinbat[1], Hannes Leonhard[1,3], Moritz Ladenburger[1,3], Frederik Alexander[1,3], Vishisht Choudhary[1,3], Fabio Laredo[4], Giacomo Masserdotti[4], Thorben Prein[1,3], Carsten Marr[*5,6,7,8,9], and Amirhossein Kardoost[*5,9]

[1] TUM.ai, Munich, Germany
[2] Ludwig-Maximilian-University, Munich, Germany
[3] Technical University of Munich, Munich, Germany
[4] Institute of Stem Cell Research, Helmholtz Munich, German Research Center for Environmental Health, Neuherberg, Germany
[5] Computational Health Center, Helmholtz Munich - German Research Center for Environmental Health, Neuherberg, Germany
[6] Department of Medicine III, Ludwig-Maximilian-University Hospital, Munich, Germany
[7] Department of Physics, Ludwig-Maximilian-University, Munich, Germany
[8] German Cancer Consortium (DKTK), partner site Munich, Germany
[9] Munich Center for Machine Learning (MCML), Munich, Germany

Abstract. Hierarchical structure is common in image data, where fine-grained clusters often merge into larger, coarser semantic clusters. In biological cell images, current self-supervised learning models tend to suppress the inherent hierarchical structure, as coarse factors (such as the imaging modality) systematically obscure finer attributes (e.g., morphological details) in the latent space. We propose a novel hierarchy-aware self-supervised training framework to address this problem. Our method integrates two components: we propose (i) a distillation framework by introducing a segmentation teacher that improves the latent space's morphological awareness, and (ii) a hierarchy-aware contrastive loss based on HDBSCAN to improve decision boundaries between closely related yet distinct subtypes at each hierarchical level. Both components counteract the self-supervised learning's tendency to overemphasize coarse factors by aligning embeddings with semantic and morphological cues, yielding biologically meaningful sub-clusters driven by fine morphological detail. We train and evaluate our method on a curated corpus of 2.3M single cells aggregated from 20 microscopy datasets (labeled and unlabeled), covering 208 cell classes. Our method improves upon the counterparts and baseline, with the average top-K accuracy up by +2.8%, top-9 retrieval on the dataset with a deeper hierarchy up by +6.3%, and an F1-score on biologically relevant classification of administered drug from perturbed cell morphology downstream up by +7.8%.

[†] Equal contribution.

[*] Corresponding authors: {carsten.marr,amirhossein.kardoost}@helmholtz-munich.de.

Keywords: Self-supervised Learning · Hierarchical Representation · Single Cell Microscopy

1 Introduction

Accurately labeled microscopy datasets are scarce, motivating self-supervised learning (SSL)-based methods to encode biologically meaningful signals for downstream tasks [29,31]. Distillation-based models such as DINO [7] scale well without labels and avoid contrastive large-batch or memory-bank overhead, making them strong base models for this task [7, 9, 24, 41, 48, 58].
Biological cell imaging presents a distinct set of unsolved challenges for SSL. Unlike natural images, which exhibit easily separable class boundaries at the object level (*e.g.*, jay vs. magpie [11, 21]), cellular microscopy images rely on subtle morphological and textural cues to distinguish similar yet distinct cell types.

For instance, different cell types can appear indistinguishable under the same modality (*e.g.*, OPCs and Neurons when captured with the same modality (Fig. 1-left)). Conversely, the same cell type can appear markedly different across imaging modalities; for example, a neuron in brightfield versus the same neuron in *e.g.* fluorescence microscopy. A robust latent space should align the same cell types across modalities while separating distinct cell types within each modality. This structure reflects stronger morphology awareness and supports downstream tasks such as drug-perturbation analysis and cell classification. Examples of multi-modal cell imaging are shown in Fig. 1-left.

The imaging setup, modality, and batch effects tend to be the dominant features in cell imaging [2, 23]. This encourages models to form coarse "super-

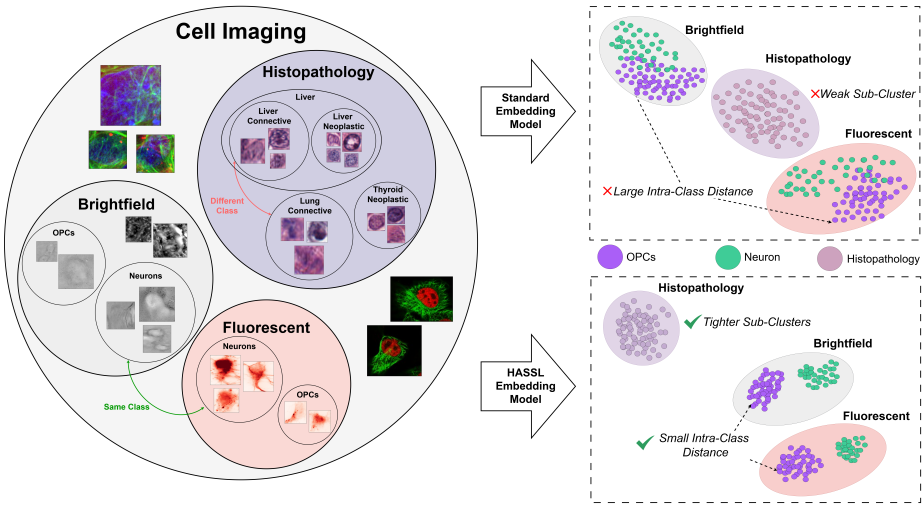

Fig. 1: HASSL learns a more hierarchical embedding space, leading to improved morphological representation and tighter subclusters. Hierarchy in cell imaging (left) and the objective with HASSL's embedding model (right).

clusters" that obscure biologically meaningful substructure needed for tasks such as phenotype classification, mechanism-of-action (MOA) inference, or cell-state discrimination [3]. Supervised class-aware methods offer better performance, although they struggle with generalization due to the scarcity of large and versatile labeled datasets in cellular imaging [26].
Therefore, we propose an SSL framework that captures hierarchical structure in a single-cell representation. Our contributions are fourfold:

- **Hierarchy-aware objective.** To improve intra-supercluster decision boundaries, we propose a novel, generalisable, label-free, hierarchy-aware objective that uses an HDBSCAN cluster tree to define subcluster centroids, pulling each individual cell towards their parent cluster's centroid while pushing them away from other clusters at each level of the hierarchy.
- **Double-teacher distillation.** To encourage morphology-aware representations, we add a second distillation teacher that uses segmentation masks as a weak prior to provide structure-aware supervision. It aims to pull each cell's embedding towards its mask view to initiate breaking up modality-driven superclusters.
- **Benchmark curation.** We consolidated scattered benchmarks into a unified single-cell, multi-modality dataset covering 208 labeled classes and 2.3 million single cells.
- **Empirical improvements.** We outperform state-of-the-art SSL cell embedding baselines across a large multi-modality corpus, with higher retrieval, clustering, and downstream metrics, reflecting better alignment with biologically meaningful subclusters.

Code is available at: `https://github.com/tum-ai/HASSL`.
Curated dataset is available at: `https://huggingface.co/datasets/tum-ai/HASSL-SingleCellBench`.

2 Related Work

In this section, we look into visual representation learning, microscopy-specific cell imaging, and hierarchical SSL, with a focus on representations that preserve multi-level structure in unlabeled data. For single-cell microscopy, this means encoding modality and fine-grained morphology across heterogeneous assays, yet standard self-supervision often overfits to global cues rather than a morphology-centric hierarchy [23, 61].

General-purpose visual representation learning has shifted from contrastive objectives to distillation and masked-image modeling. Contrastive methods such as SimCLR [9] and triplet loss [46] avoid pixel-space generation in VAEs and VQ-VAEs [30, 50], but require many comparisons, strong augmentations, and explicit negatives [9, 24]. Distillation methods such as BYOL [20] and the DINO family [7, 41, 48] remove negatives via self-distillation, but are tuned for generic semantics and coarse separation on datasets like ImageNet [11]. They do not model hierarchy, and in cell imaging they often cluster by modality rather than

capturing subtle within-modality morphology, limiting single-cell representation learning.

Cell imaging has followed a similar shift, moving from assay-specific pipelines to generalisable SSL models. CellPaint [29] benchmarks standard SSL for Cell Painting using Masked Auto Encoders (MAE) [32], DINO [7], and SimCLR [9], with DINO as the strongest baseline. Microscopy-specific designs address practical constraints. Chada-ViT [4] handles variable channel counts via channel-aware tokenisation. SubCell [22] uses MAE-style pretraining [32] to learn protein localisation and morphology from Human Protein Atlas (HPA) images [57]. OpenPhenom [32] trains channel-agnostic Vision Transformer (ViT) and MAE models on millions of RxRx [45] and Cell Painting images. scDINO [44] adapts self-distillation to fluorescent single-cell crops on top of a DINO backbone. Yet these methods are often channel- or assay-aware rather than morphology-aware because they do not link embeddings to segmentation geometry during training, leaving morphology cues weak.

The works most similar to ours are hierarchy-aware mining methods that modify pair selection so the loss encodes structure directly CoHiClust [62] learns a contrastive cluster tree, while SeedViews [59] uses multi-level views with multiple positives to keep semantically similar instances aligned. HCSC [21] is closest to our approach, dynamically constructing hierarchical prototypes and adaptively expanding positives to nearby semantics while sharpening negatives for distinct samples. However, its recursive k-NN sampling relies on hard positives and negatives that are difficult to define stably and can under-emphasise fine morphology, leading to unstable gradients and slower convergence.

Our method combines stability-weighted hierarchical prototypes with double-teacher distillation. The zero-shot segmentation masks add geometric cues that guide coarse-to-fine morphology learning, yielding a more hierarchy-aware embedding space.

3 Methodology

We design a representation-learning framework that embeds a single modality of a single-cell image. This framework enables the model to learn a hierarchy-aware latent space without any explicit supervision about the hierarchy itself. To achieve this, we approximate cellular morphology and actively encourage the formation of meaningful subclusters in the embedding space. Our method must satisfy two objectives. First, it needs to inform the embeddings about cellular structure, independent of imaging modality, by leveraging segmentation masks as weak priors. This initiates the emergence of label-free, morphology-based subclusters. Second, the loss must pull cells belonging to the same subcluster together while pushing apart cells from different subclusters across all levels of the hierarchy. We propose an annealed training strategy with an additional teacher head to instruct the embeddings about the structure, and a loss that uses hierarchical clustering to reinforce the subcluster structures in the latent space. Figure 2 illustrates our model.

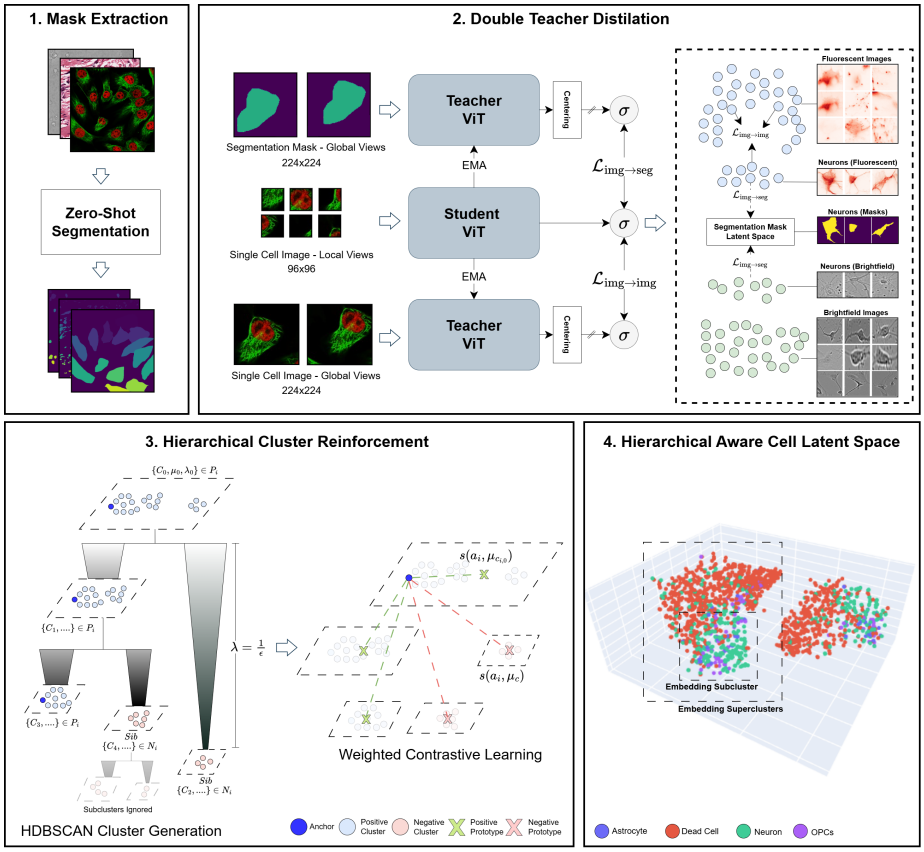

Fig. 2: Overview of the HASSL pipeline. (1) Segmentation-map generation: microscopy images are zero-shot segmented [49], which is used as a proxy for morphological features (2) Double-Teacher distillation: a student ViT is trained with a global image teacher and a segmentation teacher, promoting subcluster breakout in the latent space. The teacher's parameters are updated as EMA (Exponential Moving Average) of the student's weights (3) Hierarchical cluster reinforcement: multi-level clusters are formed with HDBSCAN [5]; cluster stability λ weights positive pairs within parents, while negative is defined as $\frac{1}{\lambda}$, yielding stability-weighted contrastive learning. (4) Resulting latent space: the learned embedding organizes cells into superclusters with clearer subclusters (illustrated for astrocytes, neurons, oligodendrocyte progenitor cells (OPCs), and dead cells), reflecting improved hierarchy awareness and structure-preserving separation.

3.1 Double-Teacher DINO Distillation

To bias representations away from modality cues and toward morphology, we guide the model's attention towards cell shape. To avoid leaking labels, we pre-compute segmentation masks using a generalist segmentation model. These mask-derived shape cues enter only through the dataloader as weak, structure-aware supervision. As a weak prior, these masks can tolerate an acceptable level of noise and error, while still providing meaningful gradients to the model.

Therefore, any zero-shot segmentation method would suffice in this case. We have used CellposeSAM [49] in our approach. With this mask, we implement a double teacher distillation framework with an additional segmentation teacher that conditions on both images and masks, supervising the student with mask-informed targets. This encourages morphology-focused representations and begins to break modality-driven superclusters into morphology-based ones.

As shown in Fig. 2 (Sec. 2), instead of the traditional single-teacher EMA setup in DINO (global teacher targets vs. cropped student views), we use two self-supervised teachers: an EMA image teacher that provides global semantic targets and a segmentation teacher that provides segmentation-aware targets. Both loss objectives $\mathcal{L}$ aim to minimize the distance between the student ViT's [14] local-view embedding and the segmentation and image global-view embeddings of the teacher ViTs. This design encourages cells of the same class but from different modalities to align with their segmentation-based embeddings, while allowing them to separate slightly from their broader modality-specific superclusters.

For the teacher, we compute sharpened, centered targets using Sinkhorn-Knopp [6], following the approach used in DINOv3 [48]. This gives us our two targets q_{img} and q_{seg} over the global view g, defined for image and the segmentation map as:

$$q^{(g)}_{\text{img}} = \text{SK}\big(z^{(g)}_{T,\text{img}}; \tau\big), \qquad q^{(g)}_{\text{seg}} = \text{SK}\big(z^{(g)}_{T,\text{seg}}; \tau\big), \tag{1}$$

where z_T are the teacher embeddings, and τ is the temperature for Sinkhorn-Knopp. To write all view-to-view distillation terms compactly, we functionalize the objective of DINOv3 [48]. For this, let V be the set of all student views (global and local) and $V_t \subset V$ the teacher views (the global crops). With $CE(q, p) = -\sum_k q_k \log p_k$, the teacher targets $q^{(v_t)}$ for $v_t \in V_t$, and the student probabilities $p^{(v_s)}$ for $v_s \in V$. The image-level objective is the pooled average over all valid teacher→student pairs:

$$\Phi\big(q, \{p\}\big) \;=\; \frac{1}{|V_t|\,(|V|-1)} \sum_{v_t \in V_t} \sum_{\substack{v_s \in V \\ v_s \neq v_t}} CE\big(q^{(v_t)},\, p^{(v_s)}\big). \tag{2}$$

With this notation, the standard image-to-image loss term of DINO is defined as:

$$\mathcal{L}_{\text{img}\to\text{img}} = \Phi\big(q_{\text{img}},\, \{p^{(v)}_{S,\text{img}}\}_{v\in V}\big). \tag{3}$$

The teacher's contribution to the segmentation branch is:

$$\mathcal{L}_{\text{seg}\to\text{seg}} = \Phi\big(q_{\text{seg}},\, \{p^{(v)}_{S,\text{seg}}\}_{v\in V}\big), \tag{4}$$

$$\mathcal{L}_{\text{img}\to\text{seg}} = \Phi\big(q_{\text{seg}},\, \{p^{(v)}_{S,\text{img}}\}_{v\in V}\big). \tag{5}$$

The $\mathcal{L}_{\text{seg}\to\text{seg}}$ stabilizes the segmentation embedding space. Otherwise, it would pull the image latent toward unmodeled/noisy segmentation targets. We combine Eqs. 3, 4, and 5 via a convex combination to get our Double-Teacher objective as follows:

$$\mathcal{L}_{\text{DoubleTeacher}} = (1-\gamma)\,\mathcal{L}_{\text{img}\to\text{img}} + \gamma\,(\mathcal{L}_{\text{seg}\to\text{seg}} + \mathcal{L}_{\text{img}\to\text{seg}})\,, \tag{6}$$

where $\gamma \in [0, 1]$.

3.2 Hierarchy Aware Contrastive Loss

We aim to strengthen the hierarchical subcluster structure in the latent space without using any labels. In a nutshell, given the batch embeddings, we run HDBSCAN [5] to obtain a minimum spanning tree (MST) that reveals each point's cluster memberships from leaf to root across the hierarchy. At each level of the tree, we compute a cluster centroid ("prototype") and, for each point, mine positive and negative prototypes within our stability-λ-weighted hinge-contrastive loss. Together, the losses draw each point toward its ancestor prototypes while repelling negatives at each hierarchical level. Negatives that share the point's parent are still drawn to that parent, but in a manner that maximizes their separation from the point while preserving latent-space integrity, resulting in clear separation between traditionally similar yet morphologically distinct cells (*e.g.*, OPCs and Neurons in brightfield as seen in Fig. 1-left).

Multi-resolution hierarchy-aware prototypes. Flat clustering methods like DBSCAN [15] force one resolution. They either merge subtypes or over-fragment into microclusters. Hierarchical clustering methods avoid this by keeping fine clusters while grouping them under broader parent clusters. Hence, we use HDBSCAN on in-memory batch embeddings to obtain the condensed cluster tree, a minimum-spanning-tree (MST) view of the latent-space hierarchy, setting `min_cluster_size`=2 to maximize depth while avoiding singleton leaves. This depth-first, label-free approach is needed because the categories and subcategories are unknown. Hence, the model discriminates on minute morphological differences at the leaf level, which supports subtle, biologically relevant downstream tasks. As illustrated in Fig. 2 (Sec. 3), each anchor is assigned, at every level of the hierarchy, to a cluster together with its member points, forming a nested hierarchy of memberships. Clusters that the anchor does not belong to serve as negatives. We denote the D dimensional, L2-normalized student embedding of sample i by $\hat{x}_i \in \mathbb{R}^D$ and define the anchor as $a_i = \hat{x}_i$. Each node c in the tree corresponds to a cluster with member set $\mathcal{C}_c$ and stability (persistence) proxy λ_c. These level-wise prototypes and stabilities define the positive and negative sets used for prototype construction and mining in our contrastive objective.

Rather than traditional instance pairs, we utilize prototypical objectives. For each positive and negative cluster in Fig. 2 (Sec. 3), we compute centroid prototypes μ_{c_i} (green and red crosses) relative to the anchor, which summarize the

clusters the anchor belongs to and those it does not, and thus define our positive and negative prototypes. This lowers gradient variance and provides consistent coarse-to-fine targets, stabilizing optimization and encouraging hierarchical structure [6, 21, 34]. For any cluster c, we define the L2 normalized centroid

$$\mu_c \;=\; \mathrm{norm}\left(\frac{1}{|\mathcal{C}_c|}\sum_{j\in\mathcal{C}_c}\hat{x}_j\right)\in\mathbb{S}^{D-1}. \tag{7}$$

Prototype mining. In our hierarchical task (refer to Fig. 2 (Sec. 3) for the visualization of the process), not every centroid provides useful information. For optimizations across memory and to not undermine the initial benefits of self-distillation learning methods like DINO [7], we decided to mine for the minimum efficacious amount of positive and negative prototypes for our contrastive objective. For positives, we use the prototype of each parent along the leaf-to-root path of the selected node. For negatives, given our MST and computed centroids, we remove prototypes whose information is already encapsulated by others further up the hierarchy. As shown for parent cluster C_4 in Fig. 2, the information of C_4's child sub-clusters is already captured at the parent level. Non-essential child clusters are therefore ignored. This allows us to select the minimum efficacious amount of negatives by removing possibly repetitive data points and therefore optimizing our memory footprint. Here is the mathematical formalization:
For each non-noise point i marked by HDBSCAN, let $\mathcal{P}(i) = \{c_{i0}, c_{i1}, \ldots, c_{iK}\}$ be its ancestor path from fine subcluster to coarse supercluster. The *positive anchors* are the centroids along this path,

$$P_i \;=\; \{\mu_{c_{ik}} : c_{ik}\in\mathcal{P}(i)\}. \tag{8}$$

For negatives, we use only extracted prototypes from clusters that directly split from the anchor at each level. These are the other direct children of the parent cluster of our anchor. We define these as the sibling set of a node c:

$$\mathrm{Sib}(c) \;=\; \mathrm{Ch}\big(\pi(c)\big)\setminus\{c\}, \tag{9}$$

Where $\pi(c)$ denotes the parent of c, and Ch(.) as its children. With this, we collect the siblings encountered along the path as follows:

$$\mathcal{S}(i) \;=\; \bigcup_{k=0}^{K-1}\mathrm{Sib}(c_{ik}) \;=\; \bigcup_{k=0}^{K-1}\Big(\mathrm{Ch}(c_{i,k+1})\setminus\{c_{ik}\}\Big). \tag{10}$$

The negative anchors are then exactly the centroids of these sibling clusters:

$$N_i \;=\; \{\mu_c : c\in\mathcal{S}(i)\}. \tag{11}$$

Stability weighting via λ. Traditional prototype-contrastive methods assume labels and assign uniform weights to positives and negatives [28]. In a label-free setting, this is ill-posed because a point's cluster-membership confidence

increases toward deeper nodes in the MST; we quantify this confidence using HDBSCAN stability (persistence) λ from the condensed tree, evaluated across density scales. With uniform weights, an instance is pushed equally toward both reliable and unreliable positives, which distorts the latent space. Likewise, treating closely related sibling subtypes as strong negatives on par with completely unrelated cell types breaks the intended hierarchy. Hence, we apply stability weighting using the λ of HDBSCAN to preserve the natural hierarchy while maximizing separation across parents.

Let $\lambda_{ik} \equiv \lambda_{c_{ik}}$. We map stabilities to normalized, clamped weights via a generic transform ϕ_ε:

$$\phi_\varepsilon(z_{ij}) = \frac{\max\big(\varepsilon, z_{ij}\big)}{\sum_u \max\big(\varepsilon, z_{iu}\big)}, \tag{12}$$

with small $\varepsilon > 0$ for numerical stability. Keeping in mind that we invert the lambdas for the negative samples, we define the positive weights (α_{ik}) and negative weights (β_{ic}) as follows:

$$\alpha_{ik} = \phi_\varepsilon\left(\frac{\lambda_{ik} - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right), \beta_{ic} = \phi_\varepsilon\left(\frac{\lambda_{\max} - \lambda_{ic}}{\lambda_{\max} - \lambda_{\min}}\right). \tag{13}$$

With cosine similarity $s(u, v) = u^\top v$ and margin $m > 0$, we aggregate positives and negatives by a stability-weighted mean:

$$s_{\text{ap}}(i) = \sum_{k=0}^{K} \alpha_{ik}\, s(a_i, \mu_{c_{ik}}),\ s_{\text{an}}(i) = \sum_{c \in \mathcal{S}(i)} \beta_{ic}\, s(a_i, \mu_c), \tag{14}$$

and define the per-anchor hinge loss as:

$$L_i \ = \ \big[m + s_{\text{an}}(i) - s_{\text{ap}}(i)\big]_+, \quad \mathcal{L}_{\text{HDBSCAN}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} L_i, \tag{15}$$

where $\mathcal{I}$ indexes anchors with at least one positive and one negative.

In the objective in Eq. 15, maximizing s_{ap} pulls a_i toward a stability-weighted barycenter of its sub-/supercluster anchors, tightening intra-subcluster spread while preserving alignment with the enclosing supercluster. Minimizing s_{an} repels a_i from the competing sibling cluster. Together, the hinge margin enforces

$$s(a_i, \text{own path}) \ \geq \ s(a_i, \text{sibling}) + m, \tag{16}$$

which (i) compresses subclusters, (ii) preserves supercluster structure, and (iii) widens gaps at ambiguous boundaries. Noise points (unclustered by HDBSCAN) are excluded from $\mathcal{I}$ and $\lambda_{\min}, \lambda_{\max}$ are computed per batch.

4 Experiments and Results

4.1 Dataset

Popular cell-biology benchmarks, such as MedMNIST/PathMNIST [60], are not truly single-cell datasets: their images often contain multiple instances and lack

per-cell identifiers or masks, preventing direct extraction of instance-level embeddings. We therefore convert instance masks into oriented bounding boxes and crop one cell per image, thereby transforming multi-cell datasets into a unified suite of single-cell crops. The labeled sources include nuclear and cellular instance datasets from histology, microscopy, iPSC, PBMC, and cell-painting settings [19] [55] [39] [38] [56] [43] [27]. Additional labeled sources include [53] [37] [10] [16] [17] [47] [25]. Unlabeled sources are incorporated from diverse segmentation and microscopy collections [12] [49] [42] [18] [52]. Further unlabeled sources include [51] [33] [35] [13], for which the dataset name is used as a pseudo-label for retrieval. The resulting collection spans both deep hierarchies ($depth > 1$) and flat label structures ($depth = 1$), supporting robust generalization across heterogeneous cell-biology datasets.

After dedup-free aggregation, we obtain 2,390,832 single-cell crops from 20 instance-segmentation benchmarks, spanning eight imaging modalities with overlapping cell types. For a detailed breakdown of the modalities and classes, refer to the supplementary material. We split 90/10 within each dataset and pool the 10% into a single test set where all datasets are represented.

4.2 Training Methodology and Ablations

Model details. The Model backbone is $\text{ViT}_\text{S}/16$ [14] with an EMA image teacher, a segmentation-teacher loss (Sec. 3.1) whose weight γ is linearly ramped from 0 to 0.2 over pre-training, and our additional HDBSCAN-based (Sec. 3.2) whose weight is ramped from 0 to 0.1 in the final 20 epochs; all remaining losses follow the default DINOv3 [48] configuration.
Training details. We follow DINOv3 [48] defaults, training on our training split (Sec. 4.1) for 100 epochs and linearly increasing the HDBSCAN components contribution to λ=1 for 20 (batch 128, multi-crop: 2 global + 8 local) on 2× NVIDIA RTX A6000 (48GB).
Baselines and Ablations. We compare our models with several cell representations and SSL approaches. These include other cellular imaging centric DINO-based models like Cellpaint-DINO [29] and scDINO [44], HCSC [21] for an alternative hierarchical objective, and models using non-DINO frameworks, such as OpenPhenom [45] and ChadaViT [4]. Some of these, which have not seen cellular datasets before, have been trained on our set with the same training parameters and resources. This is to ensure fairness in comparing the models.

Apart from comparing with other cell representation approaches and our baseline DINOv3, we also perform ablations on certain components in our approach. To validate the need for hierarchy, we evaluate DINOv3 fine-tuned with DBSCAN. Furthermore, to see the effect of stability weighting, we also evaluate an unweighted HDBSCAN fine-tuned model. Lastly, we apply weighted HDBSCAN and Double Teacher separately to measure the contribution of these components individually. All ablations use the same backbone, crops, and optimization as above.

4.3 Evaluation Method

We assess embedding quality with a k-NN retrieval protocol on the global embedding space over all datasets. For each test cell, we retrieve its top-K cosine neighbors ($K \in 1, 3, 5, 9$), exclude the query, and compute the top-K accuracy (Acc@K), precision (Prec@K), and mean average precision (mAP), treating same-type cells as positives and others as negatives. Average Precision (AP) is computed on the ranked list truncated at the largest K, with queries lacking a same-type neighbor assigned $\texttt{AP} = 0$. This evaluation emphasizes neighborhood coherence and nearest-neighbor exactness, aligning with our goal of improving subcluster consistency across modalities.

Furthermore, to gauge the quality of the latent space, we also test with cluster-specific metrics. For this, we use Adjusted and Normalized Mutual Information (AMI and NMI) [54] to identify the amount of class information captured by the latent space.

4.4 Downstream Cell-Type Classification

To further assess the discriminative quality and generalizability of the learned embeddings, we train and evaluate a multilayer perceptron (MLP) classifier on embeddings from frozen models. The MLP has two hidden layers, ReLU activations, a 0.2 dropout rate, and is optimized via a cross-entropy loss.

In addition to our curated dataset, we introduce an unseen evaluation set, the Human Protein Atlas (HPA) [57] and the Allen Institute of Cell Science Perturbation Set [1], which was excluded from the training pipeline to assess the generalization capability of HASSL. All evaluations use the same MLP architecture and hyperparameters to ensure comparability across datasets. We report classification accuracy, macro F1-score, as well as weighted F1-score, for both our dataset and HPA. Since accuracy can be misleading under class imbalance, macro F1 provides a class-balanced view by averaging per-class F1-scores uniformly, while weighted F1 accounts for class prevalence. Reporting both thus captures performance on rare classes (macro) and reflects robustness on the empirical distribution (weighted).

4.5 Results

Retrieval and Latent Space. Table 1 reports accuracy@K, precision@K, mAP, and the clustering metrics (NMI and AMI) for our variants and self-supervised baselines. Both components, double-teacher distillation and weighted HDBSCAN, consistently boost retrieval: the segmentation teacher reduces the modality-driven separation and improves top-range neighborhoods (K=5–9), while stability-weighted HDBSCAN sharpens hierarchical subclusters and improves small to mid-K retrieval. Ablations with flat DBSCAN or unweighted HDBSCAN largely remove these gains, showing that treating all hierarchy levels equally weakens boundaries and multi-resolution structure. Together, the two components are complementary and yield consistent improvements across all K

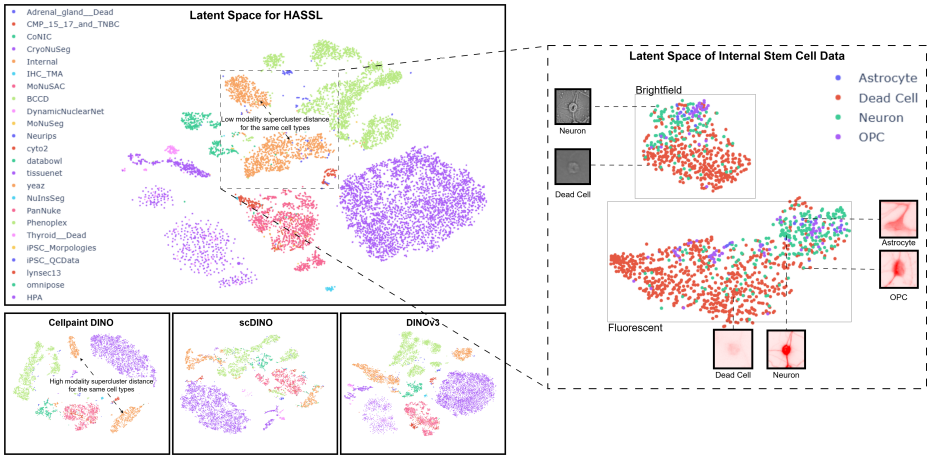

Fig. 3: Latent space visualization from our model on the test set, produced using t-SNE [36] with dimension=2. In the internal stem cell data (orange), the double teacher distillation brings superclusters of different modalities but the same cell type closer, and HDBSCAN still keeps the subclusters compact and discernible. In the global space, the Fluorescent Cluster is the 2nd closest neighbour to the Brightfield cluster for HASSL, versus 8th for DINOv3 and 22nd for scDINO.

Method	**K=1**		**K=3**		**K=5**		**K=9**		**mAP**	**NMI ↑**	**AMI ↑**
	Acc	Prec	Acc	Prec	Acc	Prec	Acc	Prec			
Cellpaint-DINO [29]	44.8	44.8	55.7	43.8	61.7	43.2	68.9	42.4	49.7	47.0	46.4
ChadaViT [4]	46.3	46.3	58.1	44.8	71.7	45.4	79.4	43.8	52.9	45.2	44.6
OpenPhenom [45]	45.3	45.3	57.1	35.1	63.9	32.6	71.9	30.5	50.8	40.3	39.5
scDINO [44]	50.4	50.4	61.9	49.2	68.0	48.5	75.3	47.5	55.4	43.9	43.3
HCSC [21]*	46.1	46.1	61.6	45.3	64.5	44.7	79.7	40.7	53.9	45.3	44.6
Baseline DINOv3 [48]*	49.4	49.4	64.4	48.5	72.1	48.0	80.9	47.4	56.5	46.8	46.2
DINOv3 + DBSCAN	49.1	49.1	61.6	47.6	68.1	46.7	75.6	45.2	55.2	47.2	46.6
DINOv3 + Unweighted HDBSCAN	48.4	48.4	60.6	46.7	67.2	45.8	74.5	44.7	54.3	47.0	46.3
HASSL (without Double Teacher)	**50.9**	**50.9**	66.6	**50.1**	74.3	**49.5**	82.8	**48.7**	**58.1**	47.6	46.9
HASSL (without HDBSCAN)	50.6	50.6	67.6	49.5	75.1	48.9	82.9	48.0	57.4	47.2	46.2
HASSL	**50.9**	**50.9**	**68.0**	50.0	**75.5**	49.4	**83.5**	48.6	57.8	**47.9**	**47.3**

Table 1: Top-k retrieval and clustering agreement results (%). We report Acc/Prec for $K \in \{1, 3, 5, 9\}$, overall mAP, and clustering agreement metrics (NMI and AMI). Relative to the DINOv3 baseline, stability-weighted HDBSCAN and Double-Teacher distillation provide complementary gains in retrieval performance, and their combination yields the strongest clustering agreement. * *retrained on our subset for fairness.*

and a latent space that preserves modality superclusters but exhibits crisper type-level groupings (Fig. 3). To further show the effect of this approach on deep hierarchies, we divide our datasets into single-level hierarchy (depth=1) and multi-level hierarchy (depth>1) subsets, and show their retrieval accuracy in Fig. 4. The plots show that HASSL improves considerably over the baseline (6.3% for $k = 9$) without compromising its performance on datasets with flat hierarchies. The combination also yields the highest clustering agreement across all metrics.

Method	Acc (%)	$F1_{macro}$ (%)	$F1_{weighted}$ (%)
HASSL (Ours)	**45.6** (0.0)	48.8 (0.3)	**43.6** (0.0)
Baseline DINOv3 [48]	44.9 (0.1)	47.6 (0.2)	42.8 (0.1)
HASSL (w/o DT)	44.8 (0.1)	46.5 (0.1)	42.9 (0.2)
HASSL (w/o HDBSCAN)	44.6 (0.1)	47.8 (0.0)	42.6 (0.2)
Cellpaint-DINO [29]	43.4 (0.3)	**49.4** (0.2)	41.4 (0.3)
ChadaViT [4]	42.6 (0.1)	46.1 (0.2)	40.6 (0.1)
HCSC [21]	41.9 (0.1)	43.1 (0.1)	41.9 (0.3)
scDINO [44]	37.6 (0.0)	37.7 (0.9)	34.4 (0.1)
OpenPhenom [45]	34.7 (0.1)	28.6 (0.1)	31.7 (0.2)

Table 2: Downstream multi-class classification on our dataset, sorted by $F1_{macro}$ (high $\rightarrow$ low). Values are mean (standard deviation) across 5 folds, reported as percentages. HASSL remove bold face achieves the best accuracy and $F1_{weighted}$, and improves over Baseline DINOv3 by +0.7 Acc, +1.2 $F1_{macro}$, and +0.8 $F1_{weighted}$, while remaining within 0.6 $F1_{macro}$ of the best Cellpaint-DINO.

Downstream cell classification on our curated dataset. The purpose of this test is to show the model's ability to recognize cell morphologies amongst the modalities and morphologies it has seen. As shown in Table 2, HASSL yields the strongest separability on frozen features. Macro F1 improves because the hierarchy-aware loss reduces confusion among small subclasses, while weighted F1 remains competitive as coarse cluster structure is preserved.

Drug identification from perturbed cell images. The main goal with this task is to measure the model's capability of understanding morphological differences in a deeply hierarchical dataset. For this, we have used immunofluorescent images from the Perturbation Dataset from the Allen Institute of Cell Science [1] containing 7 cell lines (AICS-10/12/16/22/23/24/25) perturbed with 2 drugs - paclitaxol and brefeldin. We frame the task such that a multi-layer perceptron takes in the embedding of the cell image and predicts the drug it was perturbed with (or if the image was a control). Compared to standard cell classification, this downstream gives us a more biologically relevant benchmark, as modeling the biochemistry behind drug perturbation is often challenging and very important for drug discovery. The results in Table 3 show that HASSL demonstrates

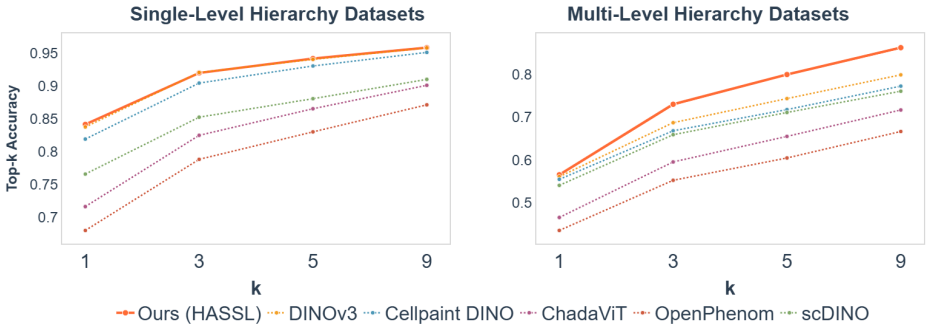

Fig. 4: Top-K retrieval accuracies for subsets containing single-level hierarchy datasets (left, depth=1) and multi-level hierarchy datasets (right, depth>1). Here, HASSL is the best-performing model, with a 6.3% improvement on multi-level hierarchy datasets.

a considerable improvement over the baselines, demonstrating its capabilities in understanding hierarchy and identifying small morphological differences within the same cell lines.

Human Protein Atlas (HPA) dataset. Although HPA [57] does not show a deep hierarchy, it has a vastly different modality with the immunofluorescent images and cell types that the model has not seen. This gives us a good insight into the model's capability to generalize to a different type of data. Table 4 shows that the model tuned to HPA-like data, CellPaint-DINO [29], achieves the best performance. This is expected: it was trained on the massive JUMP-CP [8] dataset, whose immunofluorescent dyes and organelle-targeted genetic perturbations closely match the imaging characteristics of HPA. As a result, CellPaint-DINO has effectively "seen" similar organelle structures during training, giving it a clear advantage on this task.

However, HASSL yields a performance very close to Cellpaint-DINO while outperforming every other model, including the baseline DINOv3 [48], despite never seeing HPA during training. Furthermore, we also see the importance of the double-teacher approach, where the majority of the gains come from the double teacher, as HDBSCAN lacks any hierarchy in the dataset to reinforce. This indicates that encouraging morphology-driven structure improves transferability across modalities, even when the external label differs.

Compute Overhead. Over 100 epochs, overhead is small: +43.6 min (+2.3%) wall-clock and +4.44 GB (+10.2%) VRAM vs. DINOv3 (32h01m/43.44 GB → 32h44m/47.88 GB) at identical GPU power and clocks, enabled by frozen teacher passes, $seg \rightarrow img$ reusing DINO activations, and checkpoint-bounded $seg \rightarrow seg$.

5 Conclusion and Future Work

We propose a hierarchy-aware self-supervised framework for learning representations from single-cell microscopy images. Our method adds a drop-in objective for self-distillation embedding models by constructing hierarchical prototypes via in-batch HDBSCAN and weighting their reliability using clustering confidence. Based on these prototypes, we perform hierarchy-consistent pair mining:

Method	Acc (%)	$\mathbf{F1}_{\mathbf{macro}}$ (%)	$\mathbf{F1}_{\mathbf{weighted}}$ (%)
HASSL (ours)	**92.2** (0.4)	**88.9** (0.8)	**92.0** (0.5)
HASSL (w/o HDBSCAN)	88.8 (0.3)	86.7 (0.9)	88.3 (0.5)
OpenPhenom [45]	85.7 (0.4)	75.8 (0.6)	84.2 (0.4)
HASSL (w/o DT)	83.0 (0.1)	79.0 (0.1)	82.8 (0.1)
Baseline DINOv3 [48]	81.9 (0.2)	77.1 (0.3)	81.8 (0.2)
Cellpaint-DINO [29]	84.2 (0.4)	71.8 (1.3)	81.7 (0.7)
ChadaViT [4]	79.2 (0.5)	72.5 (0.9)	78.3 (0.6)
HCSC [21]	64.4 (1.2)	56.2 (2.5)	63.5 (1.3)
scDINO [44]	65.0 (1.2)	53.8 (2.6)	63.4 (1.4)

Table 3: Downstream perturbation identification task. Values represent the mean with standard deviation in parentheses across 5 folds, reported as percentages. HASSL improves considerably over the other baselines, beating the next best model, Cellpaint-DINO, by +7.8 in $\mathrm{F1}_{\mathrm{weighted}}$.

Method	Acc (%)	$F1_{macro}$ (%)	$F1_{weighted}$ (%)
Cellpaint-DINO [29]	**54.8** (0.2)	**53.5** (0.3)	**54.8** (0.3)
HASSL (ours)	54.2 (0.3)	52.7 (0.4)	54.2 (0.4)
HASSL (w/o HDBSCAN)	53.9 (0.3)	52.2 (0.5)	53.9 (0.4)
Baseline DINOv3 [48]	51.5 (0.3)	50.0 (0.3)	51.5 (0.3)
HASSL (w/o DT)	50.1 (0.3)	48.4 (0.4)	50.0 (0.4)
HCSC [21]	39.6 (0.3)	38.0 (0.4)	39.5 (0.3)
scDINO [44]	39.5 (0.1)	37.4 (0.3)	39.3 (0.3)
ChadaViT [4]	35.7 (0.2)	34.3 (0.2)	35.6 (0.1)
OpenPhenom [45]	33.4 (0.3)	31.5 (0.4)	32.9 (0.4)

Table 4: Downstream multi-class classification on HPA [57] with frozen embedding models. Values represent the mean (standard deviation) across 5 folds, reported as percentages. The only model outperforming HASSL, Cellpaint-DINO [29], likely benefits from exposure to organelle-targeted genetic perturbations and a similar immunofluorescent dye in JUMP-CP [8]. Nevertheless, HASSL remains close to the top-performing model and substantially improves over the other baselines, without having seen the modality or the data type.

positives favor ancestors and confident siblings, while negatives are scaled by hierarchical distance to avoid over-repelling closely related subtypes. A contrastive objective aligns samples with prototypes while enforcing cross-parent separation, and a double-teacher design yields more structured embedding spaces that better capture subtle morphological differences. Extensive experiments on clustering quality and downstream tasks demonstrate the effectiveness of our approach and motivate hierarchy-aware supervision for large-scale cellular representation learning and microscopy foundation models. The scope of our work is not limited to DINO-based architectures only. Segmentation embeddings can define pseudo-positive and pseudo-negative pairs for any SSL or class-guided objective (e.g., BYOL [20], Triplet Loss [46], SimCLR [9]), while the HDBSCAN term is an additive batch-level loss, which can be used with any SSL setup. These variants, including replacing the hinge loss with InfoNCE [40], can be explored in future work.

Acknowledgement. C.M. acknowledges support from the European Research Council (ERC; Grant Nos. 866411, 101113551, and 101213822), the High-tech Agenda Bayern, and the Deutsche Forschungsgemeinschaft (DFG, TRR359, Project No. 491676693). This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE.

References

1. Allen Institute for Cell Science: Drug perturbation pilot study. [dataset] (2018), `https://www.allencell.org/drug-perturbation-pilot.html`, accessed 2026-02-24
2. Arevalo, J., Su, E., Ewald, J.D., van Dijk, R., Carpenter, A.E., Singh, S.: Evaluating batch correction methods for image-based cell profiling. Nature Communications **15**, 6516 (2024). `https://doi.org/10.1038/s41467-024-50613-5`

3. Bendidi, I., Bardes, A., Cohen, E., Lamiable, A., Bollot, G., Genovesio, A.: Exploring self-supervised learning biases for microscopy image representation. Biological Imaging **4**, e12 (Nov 2024). https://doi.org/10.1017/S2633903X2400014X
4. Bourriez, N., Bendidi, I., Cohen, E., Watkinson, G., Sanchez, M., Bollot, G., Genovesio, A.: Chada-vit : Channel adaptive attention for joint representation learning of heterogeneous microscopy images (2024), https://arxiv.org/abs/2311.15264
5. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) Advances in Knowledge Discovery and Data Mining. pp. 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 9912–9924. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf
7. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
8. Chandrasekaran, S.N., Ackerman, J., Alix, E., Ando, D.M., Arevalo, J., Bennion, M., Boisseau, N., Borowa, A., Boyd, J.D., Brino, L., et al.: Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. bioRxiv (2023). https://doi.org/10.1101/2023.03.23.534023, https://www.biorxiv.org/content/early/2023/03/24/2023.03.23.534023
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
10. Cutler, K.J., Stringer, C., Lo, T.W., Rappez, L., Stroustrup, N., Brook Peterson, S., Wiggins, P.A., Mougous, J.D.: Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. Nature Methods **19**(11), 1438–1448 (Nov 2022). https://doi.org/10.1038/s41592-022-01639-4, https://doi.org/10.1038/s41592-022-01639-4
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
12. Depto, D.S., Rahman, S., Hosen, M.M., Akter, M.S., Reme, T.R., Rahman, A., Zunair, H., Rahman, M.S., Mahdy, M.: Automatic segmentation of blood cells from microscopic slides: A comparative analysis. Tissue and Cell **73**, 101653 (2021). https://doi.org/https://doi.org/10.1016/j.tice.2021.101653, https://www.sciencedirect.com/science/article/pii/S0040816621001695
13. Dietler, N., Minder, M., Gligorovski, V., Economou, A.M., Joly, D.A.H.L., Sadeghi, A., Chan, C.H.M., Koziński, M., Weigert, M., Bitbol, A.F., Rahi, S.J.: A convolutional neural network segments yeast microscopy images with high accuracy. Nature Communications **11**(1), 5723 (Nov 2020). https://doi.org/10.1038/s41467-020-19557-4, https://doi.org/10.1038/s41467-020-19557-4
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second

International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
16. Gamper, J., Koohbanani, N.A., Benes, K., Khuram, A., Rajpoot, N.: Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European Congress on Digital Pathology. pp. 11–19. Springer (2019)
17. Gamper, J., Koohbanani, N.A., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N.: Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778 (2020)
18. Goodman, A., Carpenter, A., Park, E., jlefman nvidia, BoozAllen, J., Kyle, Maggie, Nilofer, Sedivec, P., Cukierski, W.: 2018 data science bowl (2018), kaggle
19. Graham, S., Jahanifar, M., Vu, Q.D., Hadjigeorghiou, G., Leech, T., Snead, D., Raza, S.E.A., Minhas, F., Rajpoot, N.: Conic: Colon nuclei identification and counting challenge 2022 (2021), https://arxiv.org/abs/2111.14485
20. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
21. Guo, Y., Xu, M., Li, J., Ni, B., Zhu, X., Sun, Z., Xu, Y.: Hcsc: Hierarchical contrastive selective coding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), https://openaccess.thecvf.com/content/CVPR2022/papers/Guo_HCSC_Hierarchical_Contrastive_Selective_Coding_CVPR_2022_paper.pdf
22. Gupta, A., Wefers, Z., Kahnert, K., Hansen, J.N., Leineweber, W., Cesnik, A., Lu, D., Axelsson, U., Ballllosera Navarro, F., Karaletsos, T., et al.: Subcell: Vision foundation models for microscopy capture single-cell biology. bioRxiv pp. 2024–12 (2024)
23. Haslum, J.F., Matsoukas, C., Leuchowius, K.J., Müllers, E., Smith, K.: Metadata-guided consistency learning for high content images. In: Medical Imaging with Deep Learning (MIDL), Proceedings of Machine Learning Research. vol. 227, pp. 918–936 (2024), https://proceedings.mlr.press/v227/haslum24a.html
24. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), https://arxiv.org/abs/1911.05722
25. Howard, A., Chow, A., CorporateResearchSartorius, Ca, M., Culliton, P., Jackson, T.: Sartorius - cell instance segmentation (2021), kaggle
26. Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Medicine **6**(1), 74 (Apr 2023). https://doi.org/10.1038/s41746-023-00811-0
27. Hussein, N., Reinhard, B., Adrian, S., Marie-Lisa, E., Philipp, L., Katarzyna, B.: Lynsec: Lymphoma nuclear segmentation and classification (Jun 2023). https://doi.org/10.5281/zenodo.8065174, https://doi.org/10.5281/zenodo.8065174
28. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS (2020)
29. Kim, V., Adaloglou, N., Osterland, M., Morelli, F.M., Halawa, M., König, T., Gnutt, D., Marin Zapata, P.A.: Self-supervision advances morphological profiling by unlocking powerful image representations. Scientific Reports **15**(1), 4876 (2025)
30. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)

31. Kobayashi, H., Cheveralls, K.C., Leonetti, M.D., Royer, L.A.: Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Nature Methods **19**(8), 995–1003 (Aug 2022). https://doi.org/10.1038/s41592-022-01541-z
32. Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al.: Masked autoencoders for microscopy are scalable learners of cellular biology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11757–11768 (2024)
33. Kumar, N., Verma, R., Anand, D., et al.: A multi-organ nucleus segmentation challenge. IEEE Transactions on Medical Imaging **39**(5), 1380–1391 (2020). https://doi.org/10.1109/TMI.2019.2947628
34. Li, J., Zhou, P., Xiong, C., Hoi, S.C.H.: Prototypical contrastive learning of unsupervised representations. In: 9th International Conference on Learning Representations (ICLR) (2021)
35. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., Lou, W., Li, H., Upschulte, E., Dickscheid, T., de Almeida, J.G., Wang, Y., Han, L., Yang, X., Labagnara, M., Gligorovski, V., Scheder, M., Rahi, S.J., Kempster, C., Pollitt, A., Espinosa, L., Mignot, T., Middeke, J.M., Eckardt, J.N., Li, W., Li, Z., Cai, X., Bai, B., Greenwald, N.F., Valen, D.V., Weisbart, E., Cimini, B.A., Cheung, T., Brück, O., Bader, G.D., Wang, B.: The multimodality cell segmentation challenge: Towards universal solutions. Nature Methods **21**, 1103–1113 (2024). https://doi.org/https://doi.org/10.1038/s41592-024-02233-6
36. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
37. Mahbod, A., Polak, C., Feldmann, K., Khan, R., Gelles, K., Dorffner, G., Woitek, R., Hatamikia, S., Ellinger, I.: Nuinsseg: A fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. arXiv preprint arXiv:2308.01760 (2023)
38. Mahbod, A., Schaefer, G., Bancher, B., Löw, C., Dorffner, G., Ecker, R., Ellinger, I.: Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. Computers in Biology and Medicine **132**, 104349 (2021). https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104349, https://www.sciencedirect.com/science/article/pii/S0010482521001438
39. Naylor, P., Laé, M., Reyal, F., Walter, T.: Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Transactions on Medical Imaging **38**(2), 448–459 (Feb 2019). https://doi.org/10.1109/TMI.2018.2865709, https://doi.org/10.1109/TMI.2018.2865709
40. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
42. Pachitariu, M., Rariden, M., Stringer, C.: Cellpose-sam: superhuman generalization for cellular segmentation. bioRxiv (2025). https://doi.org/10.1101/2025.04.28.651001, https://www.biorxiv.org/content/early/2025/05/01/2025.04.28.651001
43. Pfaendler, R.: Morphologically annotated single-cell images of human induced pluripotent stem cells for deep learning (Nov 2022). https://doi.org/10.3929/

ethz-b-000581447, https://www.research-collection.ethz.ch/entities/researchdata/b48aa89f-0c90-44e9-ad67-325ae37e3e89, data Collection; Creative Commons Attribution-ShareAlike 4.0 International
44. Pfaendler, R., Hanimann, J., Lee, S., Snijder, B.: Self-supervised vision transformers accurately decode cellular state heterogeneity. Biorxiv pp. 2023–01 (2023)
45. Recursion Pharmaceuticals: OpenPhenom: Groundbreaking publicly accessible foundation models for microscopy data (2025), accessed October 31, 2025
46. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015). https://doi.org/10.1109/CVPR.2015.7298682
47. Severin, Y.: Deep phenotyping reveals the molecular and health determinants of human immune cell morphology (Jun 2021). https://doi.org/10.3929/ethz-b-000343106, https://www.research-collection.ethz.ch/entities/researchdata/d0feef10-453f-4b8d-b8e1-582ea2976f91, dataset; Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
48. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv preprint arXiv:2508.10104 (2025). https://doi.org/10.48550/arXiv.2508.10104
49. Stringer, C., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. bioRxiv (2020). https://doi.org/10.1101/2020.02.02.931238, https://www.biorxiv.org/content/early/2020/02/03/2020.02.02.931238
50. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems (2017)
51. Van Valen Lab (Caltech): Tissuenet: training dataset for nuclear and whole-cell segmentation (Apr 2022), deepCell datasets documentation; modified Apache license (non-commercial academic use). Accessed 2025-10-27.
52. Van Valen Lab (Caltech): Dynamicnuclearnet segmentation (v1.0) (2023), https://deepcell.readthedocs.io/en/master/data-gallery/dynamicnuclearnet.html, deepCell Datasets; non-commercial academic use (modified Apache license); accessed 2025-10-27
53. Verma, R., Kumar, N., et al.: Monusac2020: A multi-organ nuclei segmentation and classification challenge. IEEE Transactions on Medical Imaging **40**(12), 3413–3423 (2021). https://doi.org/10.1109/TMI.2021.3085712
54. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (Dec 2010)
55. Vu, Q.D., Graham, S., Kurc, T., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., Gupta, R., Kwak, J.T., Rajpoot, N., Saltz, J., Farahani, K.: Methods for segmentation and classification of digital microscopy tissue images. Frontiers in Bioengineering and Biotechnology **7**, 53 (Apr 2019). https://doi.org/10.3389/fbioe.2019.00053, https://www.frontiersin.org/articles/10.3389/fbioe.2019.00053/full
56. Wang, R., Qiu, Y., Hao, X., Jin, S., Gao, J., Qi, H., Xu, Q., Zhang, Y., Xu, H.: Simultaneously segmenting and classifying cell nuclei by using multi-task learning in multiplex immunohistochemical tissue microarray sections. Biomedical Signal Processing and Control **93**, 106143 (2024)

57. Winsnes, C., Lundberg, E., Maggie, Culliton, P., Le, T., UAxelsson, Ouyang, W.: Human protein atlas - single cell classification (2021), kaggle
58. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018), https://arxiv.org/abs/1805.01978
59. Xu, H., Zhang, X., Li, H., Xie, L., Dai, W., Xiong, H., Tian, Q.: Seed the views: Hierarchical semantic alignment for contrastive representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3753–3767 (2023). https://doi.org/10.1109/TPAMI.2022.3176690, https://arxiv.org/abs/2012.02733
60. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (Jan 2023). https://doi.org/10.1038/s41597-022-01721-8, https://doi.org/10.1038/s41597-022-01721-8
61. Yao, H., Hanslovsky, P., Huetter, J.C., Hoeckendorf, B., Richmond, D.: Weakly supervised set-consistency learning improves morphological profiling of single-cell images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6978–6987 (2024)
62. Znaleźniak, M., Rola, P., Kaszuba, P., Tabor, J., Śmieja, M.: Contrastive hierarchical clustering. arXiv preprint arXiv:2303.03389 (2023), https://arxiv.org/abs/2303.03389