Medical Image Analysis 102 (2025) 103479

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# TriDeNT Ψ: Triple deep network training for privileged knowledge distillation in histopathology

Lucas Farndale [a,b,c,d,*], Robert Insall [a,b,e], Ke Yuan [a,b,c,*]

[a] School of Cancer Sciences, University of Glasgow, Scotland, UK
[b] Cancer Research UK Scotland Institute, Scotland, UK
[c] School of Computing Science, University of Glasgow, Scotland, UK
[d] School of Mathematics and Statistics, University of Glasgow, Scotland, UK
[e] Division of Biosciences, University College London, England, UK

ARTICLE INFO

*Keywords:*
Multi-modality
Self-supervised representation learning
Immunohistochemistry
Spatial transcriptomics
Cancer
Amyotrophic lateral sclerosis

ABSTRACT

Computational pathology models rarely utilise data that will not be available for inference. This means most models cannot learn from highly informative data such as additional immunohistochemical (IHC) stains and spatial transcriptomics. We present TriDeNT Ψ, a novel self-supervised method for utilising privileged data that is not available during inference to improve performance. We demonstrate the efficacy of this method for a range of different paired data including immunohistochemistry, spatial transcriptomics and expert nuclei annotations. In all settings, TriDeNT Ψ outperforms other state-of-the-art methods in downstream tasks, with observed improvements of up to 101%. Furthermore, we provide qualitative and quantitative measurements of the features learned by these models and how they differ from baselines. TriDeNT Ψ offers a novel method to distil knowledge from scarce or costly data during training, to create significantly better models for routine inputs.

## 1. Introduction

Humans are able to easily transfer knowledge gained from studying one imaging technique into another. A clinician working with a rare type of histological staining who discovers a morphological change indicating disease will easily leverage this knowledge when they see the same change in routine stains from new patients, such as H&E (Haematoxylin and Eosin)[1] staining. For a deep learning model, this information would be useless, as it would have to be retrained from scratch for the new type of staining. There exist methods to enable deep learning algorithms to shift domains, however, the generality typically comes at the cost of performance on the primary domain (Rusu et al., 2016).

Deep learning approaches are quickly becoming pre-eminent in computational pathology, as they are able to make fast and accurate predictions at scale. Furthermore, research interest is beginning to turn to methods which do not require manual labelling of data, finding features in data without supervision. Despite this, deep learning models for pathology images are often only suitable for the task on which they were trained, and cannot meaningfully transfer to new domains without significant and costly re-training.

Histology has long been the focus of a large amount of research attention in deep learning, and as a result there exist large datasets, such as TCGA (The Cancer Genome Atlas) (Weinstein et al., 2013) and HTAN (Rozenblatt-Rosen et al., 2020) containing data from routine examinations, such as H&E stains, CT scans and X-rays. This has enabled very powerful models to be trained for these modalities, as these datasets are large, well-curated, and often cover many different demographics. What these datasets typically lack, however, is strong labels for most features present in these images. This means supervised models can only be trained on these large training datasets to predict slide-level labels such as survival, rather than clinically relevant features that require more extensive annotation.

Despite not usually having strong labels, many datasets contain data from multiple sources and modalities, ranging from common techniques such as immunohistochemistry (IHC) (Liu et al., 2022) to cutting-edge technologies such as super-resolution microscopy (Qiao et al., 2021), spatial transcriptomics (Maniatis et al., 2019), and multiplex IHC (Ghahremani et al., 2023). For example, studies using spatial transcriptomics typically also obtain H&E stains alongside the genetic data.

* Correspondence to: CRUK Scotland Institute, Switchback Rd, Bearsden, Glasgow G61 1BD, Scotland, UK.
*E-mail addresses:* lucas.farndale@glasgow.ac.uk (L. Farndale), ke.yuan@glasgow.ac.uk (K. Yuan).

[1] See Figure S1 for a full list of abbreviations used in the text.

https://doi.org/10.1016/j.media.2025.103479
Received 8 February 2024; Received in revised form 13 January 2025; Accepted 21 January 2025
Available online 18 March 2025

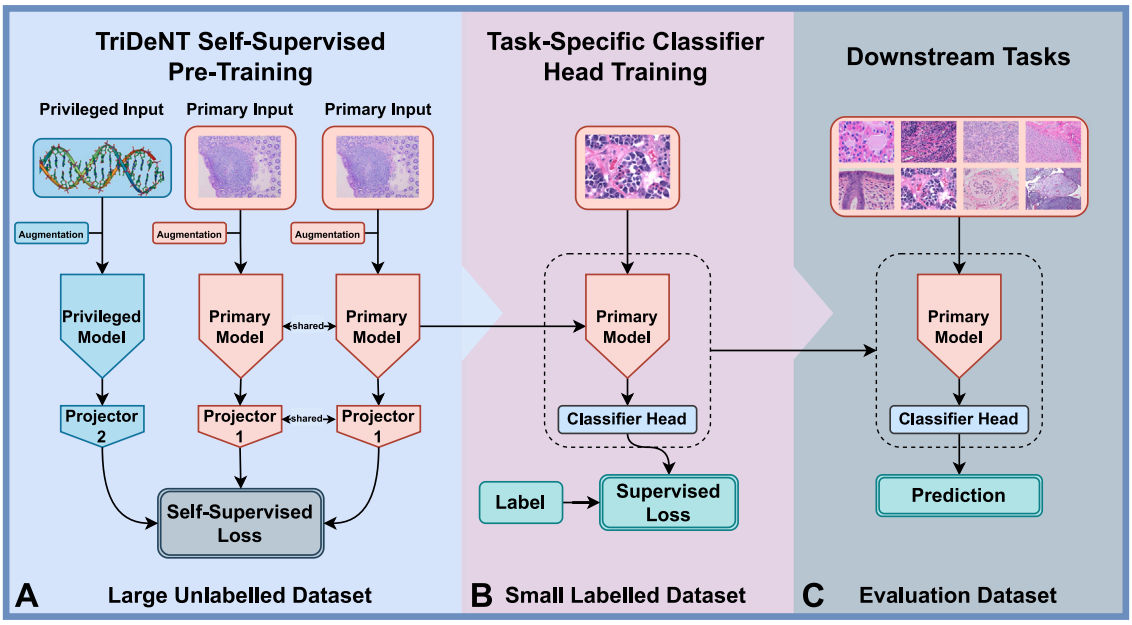



**Fig. 1.** A: TriDeNT Ψ architecture. TriDeNT Ψ incorporates information from privileged input data to complement a primary data source. There are two encoder/projector pairs, one for the primary input (e.g. H&E patches), and one for the secondary input (e.g. transcriptomics). The primary patches are augmented and passed to the primary encoder, followed by the projector, to output a representation. The privileged data are similarly passed to the privileged data encoder and projector. All representations are then used to calculate the self-supervised loss, which enforces invariance between representations. B: Classifier head training. Following this pre-training, the primary encoder is then used as a backbone for a downstream task, with a small classifier head appended. This is then trained in a supervised manner, requiring only a small amount of data. C: Use for downstream tasks. Finally, this trained model with a classifier head can be rolled out for use.

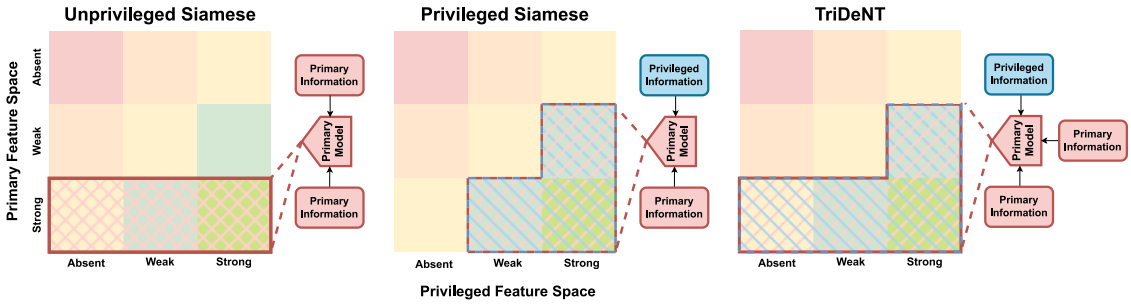
While models utilising multiple sources of data have been shown to be highly effective (Song et al., 2021; Arevalo et al., 2017; Kiela and Bottou, 2014), the abounding issue with these approaches is that obtaining additional data sources in practice is extremely difficult. State-of-the-art techniques are typically prohibitively expensive or impractical to be routinely used until long after their invention, and consequently their use is limited to a few research activities in exceptionally well-resourced labs. Even some more routine techniques, such as many immunohistochemical (IHC) stains, are arduous and expensive to obtain, register and align with existing data, and the available data will vary between samples or patients. It is therefore an important research direction to find methods which can use these additional privileged data sources during training to build better models of routine data, as these can be collected at scale and with fewer resources.

*Learning Using Privileged Information* (LUPI) methods (Vapnik and Vashist, 2009), which seek to improve performance by utilising additional data during training that is not available during inference, could be a framework to achieve this goal. By training models to learn from *privileged data*, we can develop models which make use of multiple sources to better analyse routine medical imaging without supervision. Furthermore, if they are available, manual annotations can be used as an additional input source for models to learn from during training without being restricted to only learning to output these annotations, as in supervised learning.

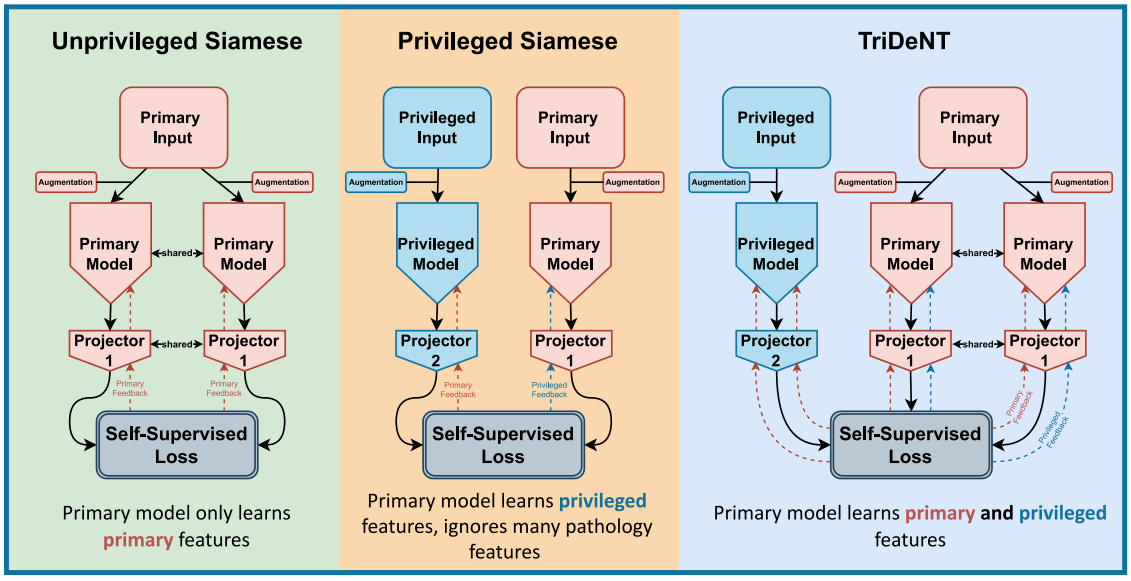
Primarily motivated by text/image retrieval tasks, there have been many LUPI methods developed. In general, these have been in supervised settings, however recently several unsupervised and self-supervised approaches have been developed. In some cases, existing unimodal self-supervised architectures have been shown to be amenable to LUPI, for example SimCLR (Chen et al., 2020), Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2021), while others have been explicitly designed to cater to this problem setting, notably CLIP (Radford et al., 2021), DeCLIP (Li et al., 2021) and ALIGN (Jia et al., 2021), which use a contrastive objective similar to SimCLR to train models to predict the correct image/text pairings, and VSE++ (Faghri et al., 2017) which uses hard-negative mining to improve representations.

Self-supervised LUPI training has been shown to be a highly effective method of improving the performance of models whose privileged data contains more task-relevant information than the primary data (Farndale et al., 2023; Girdhar et al., 2023). These methods are designed to minimise the difference between the representations of each input by mapping all embeddings into a shared latent space (*joint embedding*) (LeCun, 2022). However, in the case where we are only interested in the output of one branch, this can be restrictive. For example, if a feature is not shared between both inputs, these methods will neglect it, leading to worse performance (Farndale et al., 2023) (Fig. 2(a), see Section 2.3). This was apparent in Girdhar et al. (2023), where, despite impressive retrieval performance, the proposed joint embedding model significantly underperformed supervised models on classification tasks, implying that important features were neglected in the primary domain. In this work we present TriDeNT Ψ (Fig. 1, Section 2.4), a new method designed to enable features which are only present in the primary data to be learned in addition to those shared between inputs. The main contributions of this work are:

- We develop a new three-branch self-supervised model architecture, TriDeNT Ψ, which utilises privileged information without compromising the features learned from the primary data;
- Using standard computational pathology tasks, we find that the previous state of the art – standard Siamese self-supervised joint embedding architectures (e.g. Girdhar et al. 2023) – embeds only information shared between views, meaning performance is reduced where not all task-relevant information is present in the privileged input;
- We show that TriDeNT Ψ can incorporate features from additional stains, spatial transcriptomics, or nuclei annotations, for unprivileged downstream data analysis, and learns considerably more biologically relevant information from H&E images.




(a)




(b)

**Fig. 2.** (a) Abstract description of the features which will be learned by different types of self-supervised models. The colour of the lines reflects the information being leveraged by privileged and unprivileged primary models. Features are either strongly present, weakly present, or absent in the primary and privileged data. Unprivileged Siamese models learn only features strongly present in the primary input, and are unlikely to learn any features which are only weakly present. Privileged models are likely to only learn features strongly or weakly present in both primary or privileged inputs. TriDeNT Ψ combines the benefits of both methods to learn all features strongly present in the primary data, even those absent in the privileged data, while also learning features weakly present in the primary data that are strongly present in the privileged data. (b) Schematic for the learning process of these models. Black arrows indicate the forward flow of information through the network, and dashed lines indicate the signals which are received during backpropagation. Each branch effectively acts as a supervisory signal for the other branches, backpropagating feedback on the best features to learn. The primary model in the unprivileged Siamese setting only receives supervisory feedback from the primary data, so only learn primary features. Primary models in the privileged Siamese setting only receive supervisory feedback from the privileged data, so neglect many primary features. With TriDeNT Ψ, primary models receive feedback from both data types, leading to features from both inputs being learned.

## 2. Methodology

### 2.1. Self-supervised learning

In contrast to supervised learning which requires labels, and unsupervised learning which utilises task agnostic methods to find structure in data, *self-supervised learning* (SSL) seeks to extract supervisory signals from data that it can leverage to produce meaningful representations of its inputs. These methods differ from unsupervised methods as they require manually engineered architectures for the specific data of interest, such as choosing appropriate data augmentations.

There are two types of SSL: (i) generative, e.g., imputation of missing data, where the missing data provides the supervisory signal, such as a word masked from a sentence, and (ii) discriminative, e.g., *Siamese* models which map of multiple inputs into the same latent space, using the representation of each input as a supervisory signal for the other source.

In the typical supervised setting, training consists of passing input/label pairs $(\boldsymbol{x}, \boldsymbol{y})$ to a model $\boldsymbol{y} = \psi(\boldsymbol{x})$ and optimising $\psi$ for some loss function comparing $y$ with a ground truth. Siamese self-supervised models instead take as inputs pairs $(\boldsymbol{x}, \boldsymbol{x}')$, and use models $\boldsymbol{z} = \phi(\boldsymbol{x})$, $\boldsymbol{z}' = \phi'(\boldsymbol{x}')$ with the aim of minimising the difference $d(\boldsymbol{z}, \boldsymbol{z}')$ between $\boldsymbol{z}$ and $\boldsymbol{z}'$. Typically this is implemented with $\phi = \phi'$, with $\boldsymbol{x}$ and $\boldsymbol{x}'$ both augmentations of the same input.

This method is amenable to a trivial constant solution where $\boldsymbol{z} = \boldsymbol{c}$ for some constant $c$ for all inputs. Therefore, Siamese methods require some regularisation to avoid such collapse. There are two approaches: contrastive and non-contrastive. Contrastive methods such as Chen et al. (2020), Radford et al. (2021), Caron et al. (2020) use both *positive* (matching) and *negative* (non-matching) pairs, and seek to pull together positive pairs while pushing apart negative pairs, either in embedding space (Chen et al., 2020; Radford et al., 2021) or through cluster assignment (Caron et al., 2020). Non-contrastive methods such as Zbontar et al. (2021), Bardes et al. (2021), Chen and He (2021), Grill et al. (2020) instead use only positive pairs, and regularise the representations to avoid collapse by using architectural constraints such as momentum encoders (Grill et al., 2020), stop gradients (Chen and He, 2021) and covariance constraints (Zbontar et al., 2021; Bardes et al., 2021).

### 2.2. Knowledge distillation

Knowledge distillation is the transfer of knowledge from a *teacher* model to a *student* model (Hinton et al., 2015), usually with the goal

of either teaching a smaller student model to emulate the performance or learn some desirable property of the teacher model. This is usually achieved by passing the same input to both models, and minimising the difference between their outputs, using either their representations or possibly a projection head. The student model is usually smaller, and the teacher model may be pre-trained. The objective is to teach the student model to produce the same or a related output to the teacher, possibly using fewer resources.

In this work, we utilise this approach not to distil knowledge on the level of the model, but on the level of the data. We consider settings where there are multiple sources of data about the same input, such as different histological stains from the same sample, and we wish to train models to classify only the primary data. By mapping both inputs into a joint embedding space, the models' objective is to produce the same representation for each, with the goal of improving the quality of representations of the primary data.

### 2.3. Privileged information

Privileged information is information which is available during training but not during inference. In a supervised setting this is defined, using the notation of Section 2.1, as having data $(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{y})$ during training, and optimising a model $\boldsymbol{y} = \psi(\boldsymbol{x})$, which will then be used in inference without the privileged information $\boldsymbol{x}^*$. Most existing work on LUPI is focused on understanding supervised learning dynamics using support vector machines (SVMs), however, much of this has been extended to neural networks (Vapnik and Izmailov, 2017), unsupervised learning (Feyereisl and Aickelin, 2012; Karaletsos et al., 2015), and knowledge distillation (Lopez-Paz et al., 2015). The original framework (Vapnik and Vashist, 2009) was defined for SVMs, with the privileged information being used to estimate the slack values. The slack values can be understood equivalently for neural networks as loss values. This was leveraged by Yang et al. (2017) to use privileged information to estimate loss values for a neural network for multiple instance learning, integrating privileged information at both the instance level and the bag level. Rather than directly using privileged information to inform predictions, Lambert et al. (2018) use privileged information to determine dropout variance during training, leading to greater sample efficiency.

Despite the apparent advantage of providing privileged information to a model, it has been shown that training with privileged information does not satisfy a *no-harm guarantee* (Lambert et al., 2018). This can be due to a variety of factors, such as because estimating properties of the privileged information can be more difficult than estimating the same properties of the primary data. It was shown in Farndale et al. (2023) that Siamese LUPI leads to improved performance on tasks where the privileged input contains more task-relevant information than the primary input, e.g., a low-resolution image paired with a high-resolution image. However, it is also observed that if the privileged input contains less task-relevant information, it can reduce performance. This is because mapping both inputs into the same latent space causes task-relevant information in the primary input to be lost if it is not shared between branches, as is visualised in Fig. 2(a). Consequently, non-LUPI learning can lead to better performance in these scenarios, despite the loss of additional task-relevant information which could be gained from an privileged input.

Extending the supervised setting to the Siamese self-supervised setting, we have inputs pairs $(\boldsymbol{x}, \boldsymbol{x}^*)$, and use models $\boldsymbol{z} = \phi(\boldsymbol{x})$, $\boldsymbol{z}^* = \phi^*(\boldsymbol{x}^*)$ with only the model $\phi$ being used for inference. For example, we may have a set of H&E images $\boldsymbol{x}$ and privileged paired IHC images $\boldsymbol{x}^*$ which are only available during training.

As Siamese joint-embedding models minimise the difference between representations in the shared embedding space, any features which are not shared between branches will be neglected. There is no way to predict a feature in the privileged input from the primary input if no information exists about that feature in the primary input. On the other hand, features which are weakly present in the primary input but strongly present in the privileged input may be learned, as there is a strong supervisory signal from the privileged data. In the non-LUPI setting (Siamese learning without privileged inputs), such features are unlikely to be learned due to the absence of the strong signal from the privileged input. Formally, following Jing et al. (2021), we consider features which have variance that is nonzero but lower than the augmentation regime to be weakly present, and those with greater variance than the augmentation regime to be strongly present.

### 2.4. TriDeNT Ψ

The goal of TriDeNT Ψ method is to combine the benefits of both LUPI and non-LUPI methods in such a way that the primary encoder can make best use of signals from all inputs. We use a three-branched approach, with two branches acting on the primary input and a third acting on the privileged input. Our method can be considered a generalisation of the standard Siamese self-supervised architecture. We take inputs $\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}^*) \in (\mathcal{X}, \mathcal{X}^*)$, where we assume each input contains some information about their shared source. The inputs could represent any type of input array, such as images, -omics data, or patient information. We assume $\boldsymbol{x}^*$ contains some mutual information with $\boldsymbol{x}$. We aim to obtain representations $\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$, such that the $\boldsymbol{z}$ is a *sufficient* representation of $\boldsymbol{x}$ for some task $T$, that is to say we have mutual information $I(\boldsymbol{z}; T) = I(\boldsymbol{x}; T)$. Note that, in contrast to comparable approaches, we are only interested in optimising $\boldsymbol{z}^*$ insofar as this benefits $\boldsymbol{z}$, as only $\boldsymbol{z}$ is to be used for inference.

Inputs $\boldsymbol{x}, \boldsymbol{x}^*$ are augmented by stochastic operators

$$a : \mathcal{X} \to \mathcal{X}, \quad a^* : \mathcal{X}^* \to \mathcal{X}^* \tag{1}$$

respectively, and mapped to representations $\boldsymbol{z}^i \in \mathcal{Z}$ by encoders $f^i : \mathcal{X}^i \to \mathcal{Z}$ according to the rule

$$\boldsymbol{z}^i = f^i(\hat{a}(\boldsymbol{x})), \quad i = 1, 2, *\,. \tag{2}$$

We have defined $\hat{a}$ to be $a$ if its input is $\boldsymbol{x}$ and $a^*$ if its input is $\boldsymbol{x}^*$, as in general there is no reason for augmentations to be the same for primary and privileged data. This yields three representations, $\boldsymbol{z}^1, \boldsymbol{z}^2$, and $\boldsymbol{z}^*$, where $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ are representations of each augmentation of the primary data, and $\boldsymbol{z}^*$ is the representation of the privileged data. Representations are then mapped to embeddings $\boldsymbol{e}^i \in \mathcal{E}$ by a projector $g^i : \mathcal{Z} \to \mathcal{E}$ with the rule $\boldsymbol{e}^i = g^i(\boldsymbol{z}^i)$. In general we will have primary encoder $f = f^1 = f^2$ and projector $g = g^1 = g^2$.

Note that the spaces $\mathcal{Z}$ and $\mathcal{E}$ are not dependent on $i$, as these are shared latent spaces. Projections into an embedding space are used in keeping with existing approaches (Zbontar et al., 2021; Bardes et al., 2021; Chen et al., 2020), as this has been shown to improve generalisation and feature learning. For inference, augmentations are not applied, so $\hat{a}$ is set to $\hat{a}(x) = x$. In general, we use the same encoder for both branches taking $\boldsymbol{x}$ as input, as sharing weights has been shown to improve performance on unprivileged tasks (Farndale et al., 2023). Typically, we will have $\mathcal{Z} = \mathbb{R}^{n \times d}$ where $n$ is the batch size and $d$ is the dimension of the representation. For pseudocode, see Algorithm S1.

### 2.5. Objective function

Consider a setting where $N$ is the number of branches with the primary input and $M$ is the number of branches with the privileged input. We generalise a two-branch self-supervised loss $\mathcal{L}_2(\boldsymbol{z}^i, \boldsymbol{z}^j)$ to $N + M$ branches by summing over the losses between representations such that the $N + M$ branch loss is defined as

$$\mathcal{L}_{N,M}(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^N, \boldsymbol{z}^{*1}, \ldots, \boldsymbol{z}^{*M}) := \sum_{i \neq j}^{N+M} \mathcal{L}_2(\boldsymbol{z}^i, \boldsymbol{z}^j). \tag{3}$$

We investigate the case where $N = 2$ and $M = 1$ (giving three branches overall) however the method could easily be generalised to

**Table 1**
Datasets and tasks.

| | |
|---|---|
| Training Datasets | |
| Name: **SegPath**<br>Reference: Komura et al. (2023)<br>Tissue: Pan-Cancer<br>Split: See Table B.5<br>Task: None | Contains eight subsets of H&E slides, each with a different paired IF stain (see Table B.5). Features 1,583 patients and 18 different tissue types, with no overlap between the H&E images in each subset. IF images are only released as binarised images using a threshold value determined in the original study (Komura et al., 2023). Only used for training. |
| Name: **BCI**<br>Reference: Liu et al. (2022)<br>Tissue: Breast<br>Split: 62,336/15,632<br>Task: HER2 Status Prediction | HER2 (Human Epidermal growth factor Receptor 2) is a protein which has been found to be prognostic for breast cancer.<br>It is tested for using IHC staining, and classified into 4 grades (0, 1+, 2+, 3+).<br>BCI contains paired H&E/IHC patches from 51 breast cancer patients, which have been registered for precise correspondence between the H&E/IHC patches. |
| Name: **PanNuke**<br>Reference: Gamper et al. (2019)<br>Tissue: Pan-Cancer<br>Split: 4295/2283<br>Task: Neoplastic Cell Detection | PanNuke contains H&E patches paired with exhaustive nuclei segmentations from 19 tissue types. The associated task is<br>neoplastic cell detection, following Huang et al. (2023), where the model must determine whether a patch<br>contains an abnormal, excessive growth of tissue, whether benign or malignant. |
| Name: **ALS-ST**<br>Reference: Maniatis et al. (2019)<br>Tissue: Mouse/Human Spinal Cord<br>Split: See Appendix B.12<br>Task: Genotype Prediction & White/Grey Matter Classification | Dataset containing 80 human and 331 mouse spinal cord sections from 7 humans and 67 mice who have ALS, a neurodegenerative disease affecting the motor neurons. All samples feature a H&E slide with matched and aligned spatial transcriptomics.<br>The tasks are to predict the mouse SOD1 genotype from SOD1-G93A (ALS),SOD1-WT (Wildtype), and Knockout, and to classify white matter and grey matter. |
| Evaluation Datasets | |
| Name: **NCT**<br>Reference: Kather et al. (2018)<br>Tissue: Colorectal<br>Split: 100,000/7,177<br>Task: Tissue Classification | Manually annotated patches of nine tissue types: adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). Patches extracted from H&E slides from 86/50 patients in the train/test sets respectively.<br>This task assesses the models' ability to differentiate features which are primarily determined by the H&E image, but can be enhanced by paired information, such as presence of immune cells helping classify lymphocytes. |
| Name: **Camelyon**<br>Reference: Bandi et al. (2018)<br>Tissue: Lymph Node<br>Split: 179,394/146,722<br>Task: Out-of Distribution Metastasis Detection | There is a large degree of variation between different scanners, staining protocols and sample collection methods, so H&E images can look very different depending on how, when, and where they were collected. The WILDS distribution of Camelyon features 1399 breast lymph node whole slide images from 5 different hospitals, with centres 1,2, and 3 comprising the train set, 4 being the validation set, and 5 being the test set. There is a large difference between sets, so Camelyon assesses models' generalisation ability. |
| Name: **MHIST**<br>Reference: Wei et al. (2021)<br>Tissue: Colorectal Polyps<br>Split: 2175/977<br>Task: Polyp Classification | Features patches from 328 whole slide images of colorectal growths, *polyps*, which can become cancerous.<br>MHIST contains two classes: *serrated* polyps, which can become cancerous, and *hyperplastic* polyps, which are typically benign.<br>Note that these images are at 8× magnification, so this task assesses the models' generalisation performance across magnification scales. |
| Name: **Singapore**<br>Reference: Oner et al. (2022)<br>Tissue: Prostate<br>Split: 3843/4261<br>Task: Prostate Gland Malignancy Classification | Classification dataset with samples from 46 patients who underwent a prostate core needle biopsy. Patches are centred on a prostate gland labelled as benign or malignant. This task assesses the models' ability to make classifications based on a specific biological feature which was uncommon or absent during training. |
| Name: **TIL**<br>Reference: Kaczmarzyk et al. (2022)<br>Abousamra et al. (2022)<br>Saltz et al. (2018)<br>Tissue: Pan-Cancer<br>Split: 209,221/56,275<br>Task: Tumour Infiltrating Lymphocyte Detection | Features patches from 7983 whole slide images from 23 cancer types. The task is to detect *tumour infiltrating lymphocytes* (TILs),<br>which are an important biomarker for cancer prognosis, with increased TIL density being associated with positive clinical outcomes.<br>Assesses models' ability to detect features which co-occur with more prominent labels such as tumour. Also assesses relative performance of different paired data, as immune-related paired data is far more relevant to performance than other stains. |
| Name: **PANDA**<br>Reference: Bulten et al. (2022)<br>Tissue: Prostate<br>Split: 7962/2654<br>Task: Prostate Biopsy ISUP Grading | Large dataset featuring 10616 whole-slide images of prostate biopsies, weakly labelled with ISUP grades from 1 to 5. Results are reported as Cohen's $\kappa$.<br>Assesses performance of models' representations for aggregated slide level predictions on a difficult,<br>clinically relevant task. |

(*continued on next page*)

**Table 1** (*continued*).

| | |
|---|---|
| Name: **IMP 1K/4K**<br>Reference: Oliveira et al. (2021)<br>Neto et al. (2022)<br>Neto et al. (2024)<br>Tissue: Colorectal<br>Split: 1132(1K)/4433(4K)/900<br>Task: Colorectal Dysplasia Detection | Dataset with 1132 whole slide images (1K) from colorectal biopsies and polypectomies, with an extension to 4433 slides (4K). Labels are: non-neoplastic, low-grade lesions (conventional adenomas with low-grade dysplasia), and high-grade lesions (conventional adenomas with high-grade dysplasia, intra-mucosal carcinomas and invasive adenocarcinomas). |
| Name: **IMP Cervix**<br>Reference: Oliveira et al. (2023)<br>Tissue: Cervix<br>Split: 480/120<br>Task: Cervical Dysplasia Detection | Features 600 whole slide images of cervical Loop Electrosurgical Excision Procedure (LEEP) samples, with 4 classes: non-neoplastic, low-grade, and high-grade squamous intraepithelial lesion. |

more branches. Siamese unprivileged learning is the case with $N = 2$ and $M = 0$, and Siamese privileged learning is the case with $N = M = 1$. A short discussion of the use of additional privileged branches is presented in Appendix E, where we see that using more than one privileged branch could deteriorate performance. In the present work we use both contrastive and non-contrastive choices of $\mathcal{L}_2$ to demonstrate that TriDeNT Ψ is robust to the choice of self-supervised loss. We illustrate this using the VICReg (Variance Invariance Covariance Regularisation) (Bardes et al., 2021) objective and the InfoNCE ([Mutual Information] Noise Contrastive Estimation) objective (Oord et al., 2018), which have both been used extensively in self-supervised architectures (e.g. Lee et al. 2022, Chen et al. 2020, Radford et al. 2021, Girdhar et al. 2023).

For brevity we focus only on architectures which can be summed in this way, although this setting can be easily extended to architectures requiring more complex structuring of their loss function. For example, self-predictive architectures such as BYOL (Grill et al., 2020) and SimSiam (Chen and He, 2021) would require designation of *online* and *target* branches and pairings between them.

#### 2.5.1. VICReg

The VICReg objective is defined as

$$\mathcal{L}_{VICReg}(\boldsymbol{z}^1, \boldsymbol{z}^2) := \lambda s(\boldsymbol{z}^1, \boldsymbol{z}^2) + \sum_{i=1}^{2} \left[\mu v(\boldsymbol{z}^i) + \nu c(\boldsymbol{z}^i)\right] \tag{4}$$

where $s(\cdot,\cdot)$ is an invariance regularisation term, $v(\cdot)$ is a variance regularisation term, and $c(\cdot)$ a covariance regularisation term, with $\boldsymbol{z}^i$ being the embedding of branch $i$, and $\lambda, \mu, \nu$ weighting coefficients. These functions are

$$s(\boldsymbol{z}^i, \boldsymbol{z}^j) := \frac{1}{n} \sum_{k=1}^{n} \|\boldsymbol{z}^i_k - \boldsymbol{z}^j_k\|_2^2, \tag{5}$$

$$v(\boldsymbol{z}^i) := \frac{1}{D} \sum_{d=1}^{D} \max\left(0, \gamma - \sqrt{\mathrm{Var}([\boldsymbol{z}^i]_d) + \epsilon}\right), \tag{6}$$

$$c(\boldsymbol{z}^i) := \frac{1}{D} \sum_{d \neq \delta} \left[C(\boldsymbol{z}^i)\right]^2_{d,\delta}, \tag{7}$$

where

$$C(\boldsymbol{z}^i) := \frac{1}{n-1} \sum_{k=1}^{n} (\boldsymbol{z}^i_k - \bar{\boldsymbol{z}}^i)(\boldsymbol{z}^i_k - \bar{\boldsymbol{z}}^i)^{\mathrm{T}}, \quad \bar{\boldsymbol{z}}^i := \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{z}^i_k, \tag{8}$$

$n$ is the batch size with $[z^i_a]_j \in \boldsymbol{z}^i$ being dimension $j$ of element $a$ in the batch of representations $\boldsymbol{z}^i$, $\gamma$ is a term determining the desired variance of the representations, $D$ is the dimension of the representation, and $\epsilon$ is a small constant to ensure numerical stability. Unlike many Siamese networks (e.g. Chen and He 2021, Grill et al. 2020) VICReg can admit distinct inputs and architectures on each branch. This is because both branches are regularised separately by the covariance term, and consequently has been shown to work better than VSE++ (Faghri et al., 2017) and Barlow Twins (Zbontar et al., 2021) for multi-modal data (Bardes et al., 2021).

Both the variance and covariance functions $v$ and $c$ are applied to each branch independently, meaning that the invariance between branches is achieved simply through the distance function $s$. In the original description, these functions were implemented with all parameters shared between both branches, but this is not a necessary restriction.

#### 2.5.2. InfoNCE

We use the variant of the InfoNCE/NT-Xent/N-pairs losses used in SimCLR (Chen et al., 2020), ImageBind (Girdhar et al., 2023), etc., which is defined as

$$\mathcal{L}_{InfoNCE}(\boldsymbol{z}^1, \boldsymbol{z}^2) := \frac{1}{2n} \sum_{i=1}^{n} \left(l(\boldsymbol{z}^1_i, \boldsymbol{z}^2_i) + l_i(\boldsymbol{z}^2_i, \boldsymbol{z}^1_i)\right) \tag{9}$$

with

$$l(\boldsymbol{z}^a_i, \boldsymbol{z}^b_i) := -\log \frac{\exp\left(\mathrm{sim}(\boldsymbol{z}^a_i, \boldsymbol{z}^b_i)/\tau\right)}{\sum_{k=1}^{n} \exp\left(\mathrm{sim}(\boldsymbol{z}^a_i, \boldsymbol{z}^b_k)/\tau\right)}, \tag{10}$$

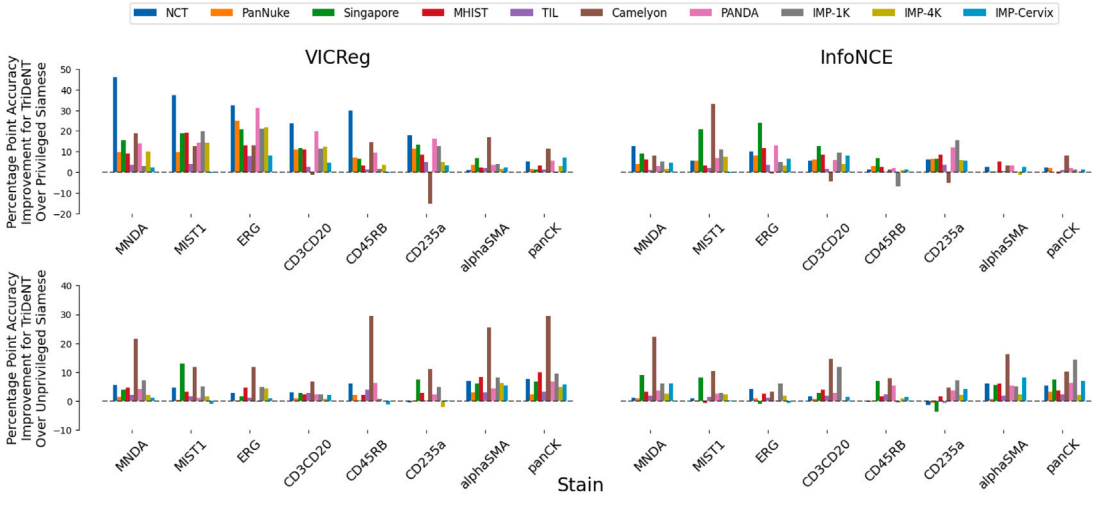
where $\mathrm{sim}(\cdot)$ is the cosine similarity, $\tau$ is a temperature parameter, and $n$ is the batch size.

### 2.6. Primary and privileged features

For an intuitive understanding of the method, it is helpful to consider the representation of each branch as a supervisory signal for the others. Our model can therefore be considered a multi-objective setting, where the primary encoder $f$ aims to balance the information extracted from each augmentation of $\boldsymbol{x}$ which is shared with $\boldsymbol{x}^*$, and that which is shared with the other augmentation of $\boldsymbol{x}$. In turn, the supervisory signals for $\boldsymbol{x}^*$ are $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$, and consequently they will only extract features which can also be found in $\boldsymbol{x}$. In our typical setting, this corresponds to balancing information which is only weakly present in the primary input $\boldsymbol{x}$, but strongly present in the privileged input $\boldsymbol{x}^*$, with information which is strongly present in primary input $\boldsymbol{x}$. The result of this trade-off is that privileged features with a strong supervisory signal from $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ are learned, but primary features with a strong supervisory signal from $\boldsymbol{z}^*$ are also learned. This is in contrast to the dichotomy between only learning strong features or only learning shared features presented by the standard 2-branch approaches.

### 2.7. Datasets and tasks

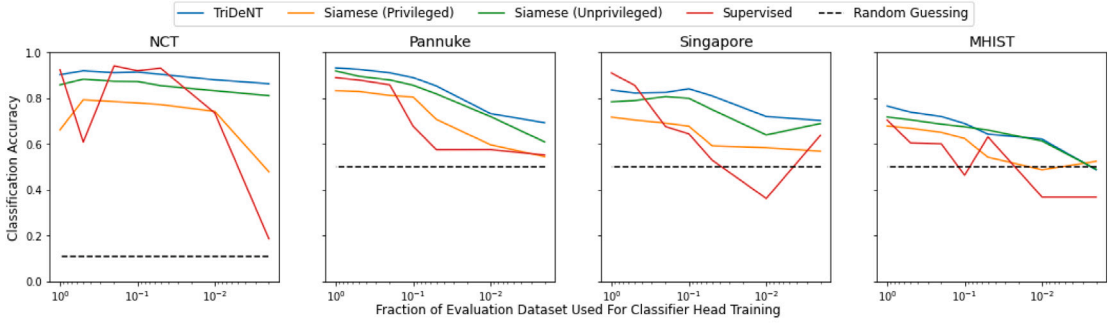
While H&E staining is the routine protocol for tissue analysis, pathologists usually rely on IHC or IF staining to obtain information about the locations of individual proteins, which may aid further investigation or confirm their diagnoses. IHC and IF stains contain highly specific information about a particular protein, so add useful information beyond that which can be readily identified with H&E staining. While this is necessary for human pathologists to identify features which cannot be identified by eye in the generic H&E stains, it has been shown that neural networks can accurately reproduce many of these stains from H&E images (Xu et al., 2019), although in other cases this is not possible, such as stains for different immune cell subtypes.




(a)

| Task | Loss: / Method: / Privileged: | VICReg TriDeNT Ψ ✓ | VICReg Siamese ✓ | VICReg Siamese ✗ | InfoNCE TriDeNT Ψ ✓ | InfoNCE Siamese ✓ | InfoNCE Siamese ✗ | Supervised Baseline |
|---|---|---|---|---|---|---|---|---|
| NCT | | **0.9031** | 0.6618 | 0.8582 | **0.8979** | 0.8405 | 0.8762 | 0.9245 |
| PanNuke | | **0.9324** | 0.8332 | 0.9190 | **0.9372** | 0.8926 | 0.9299 | 0.8901 |
| Singapore | | **0.8359** | 0.7173 | 0.7839 | **0.8488** | 0.7486 | 0.8044 | 0.9103 |
| MHIST | | **0.7656** | 0.6786 | 0.7179 | **0.7802** | 0.7236 | 0.7534 | 0.7042 |
| TIL | | **0.9270** | 0.8934 | 0.9048 | **0.9300** | 0.9127 | 0.9154 | 0.9216 |
| Camelyon | | **0.7067** | 0.6179 | 0.5229 | **0.6689** | 0.6147 | 0.5570 | 0.8440† |
| PANDA | | **0.7099** | 0.5675 | 0.6763 | **0.7243** | 0.6648 | 0.6873 | - |
| IMP 1K | | **0.7186** | 0.6415 | 0.6389 | **0.7471** | 0.6850 | 0.6832 | - |
| IMP 4k | | **0.8554** | 0.7657 | 0.8332 | **0.8632** | 0.8356 | 0.8454 | - |
| IMP Cervix | | **0.7439** | 0.7289 | 0.7301 | **0.7622** | 0.7247 | 0.7346 | - |

(b)




(c)

**Fig. 3.** (a) Difference in accuracy between TriDeNT Ψ and privileged/unprivileged Siamese training on SegPath. Values greater than zero (above the dashed line) indicate a higher accuracy for TriDeNT Ψ. For example, if TriDeNT Ψ has an accuracy of 90% and the Siamese method has an accuracy of 70%, this will give a 20 percentage point improvement. (b) Results for ten evaluation tasks averaged across all eight stains. Supervised baseline is provided for comparison, bold indicates best performance for the given self-supervised loss function. Supervised comparisons are only given for patch-level tasks, as train-time patch aggregation for slide-level tasks cannot be comparably achieved. Higher values indicate better performance. For full results see Table S4. Value marked † from Farndale et al. (2024). (c) Classification training dataset size performance comparison. Models were all pretrained on SegPath and evaluated using the full test set of each dataset. Training was carried out on 100%, 50%, 20%, 10%, 5%, 1%, and 0.2% of each classifier training dataset, and averaged over all SegPath stains. Supervised comparisons are trained in the same fashion, but not averaged.

It may be tempting to conclude, then, that learning representations of H&E patches which contain information relevant to IHC stains is a solved problem, and researchers should simply use these models to produce representations of H&E patches which contain features relevant to IHC or IF stains. Unfortunately, this has been shown to perform poorly (Farndale et al., 2023; Balestriero and LeCun, 2024), as image to image translation models are restricted to learning very fine grained information, such as the exact locations of nuclei, at the expense of the types of low-redundancy coarse grained features which are learned by self-supervised *Siamese* networks.

Nevertheless, the results of these works on image translation imply that much or all of the useful information in IHC and IF stains can be predicted from H&Es. We investigate how representation are affected by distilling information from IHC and IF stains into models of H&E stains, evaluating both brightfield IHC images and thresholded IF images. We also investigate whether distilling manual nuclei segmentations can improve performance, as this is a painstaking process for pathologists to perform, but there are now large datasets of (semi-)manually generated segmentations, with accurate models able to perform segmentation from H&E stains. We finally investigate the use of spatial transcriptomics as privileged information, as this contains rich complementary information to H&E stains, and is an example of how cutting-edge data could be distilled into models of routine data.

**Table 2**
Results for models trained on the BCI dataset containing H&E patches paired with privileged brightfield IHC.

| Loss | Method | Privileged | BCI | NCT | PanNuke |
|---|---|---|---|---|---|
| VICReg | TriDeNT Ψ | ✓ | **0.8559** | **0.8347** | 0.8966 |
| | Siamese | ✓ | 0.8552 | 0.8019 | **0.9071** |
| | | ✗ | 0.6863 | 0.8103 | 0.8506 |
| InfoNCE | TriDeNT Ψ | ✓ | **0.8800** | **0.8267** | **0.9115** |
| | Siamese | ✓ | 0.8319 | 0.7961 | 0.8677 |
| | | ✗ | 0.7034 | 0.8045 | 0.9023 |
| CrossEntropy | Supervised | – | 0.6331 | 0.9245 | 0.8901 |

In Table 1 we provide a short overview of the datasets used in this work. For a full description, please see Appendix B.

## 3. Results

### 3.1. Embedding knowledge from privileged image modalities

We first demonstrate that TriDeNT Ψ is highly effective for improving the quality of representations in the primary encoder by distilling privileged information from immunofluorescence (IF) images to H&E stained images (Fig. 3 and Table S4). Models are trained on the SegPath dataset (Komura et al., 2023), which consists of eight subsets of H&E images paired with an image derived from the IF stain of a consecutive slice for one of eight antibodies. Evaluation is performed on four standard computational pathology tasks (see Appendix B for full details). We find that the model significantly increases performance by up to 101% compared to a privileged baseline model. TriDeNT Ψ retains not only the useful features shared between inputs, but also the features which are only present in the primary data, leading to better performance on all evaluated tasks. Even in cases where the privileged data does not appear to significantly improve performance, TriDeNT Ψ still achieves comparable performance, as it obtains a strong supervisory signal from the additional H&E branch. This is in contrast with the privileged Siamese setting, where it is clear that the pairing can cause a seismic drop in classification accuracy if the privileged data is not informative for the task being evaluated (see Section 3.6 for a more detailed analysis).

We see that there are significant performance gains of up to 101% (0.4566 to 0.9169, see Table S4) in the NCT tissue type classification task for TriDeNT Ψ against the baseline privileged method. The improved performance from using TriDeNT Ψ is seen across the board, with average performance improvements on all tasks, and only a handful of cases where individual stains underperform baseline models. We generally observe that when TriDeNT Ψ performs worse than a baseline, it is only marginal, however, there are many cases where the performance difference between TriDeNT Ψ and a baseline is enormous. For example, on Camelyon, performance is improved from effectively random guessing at 51% up to 81% with pan-CK as privileged information.

Perhaps unsurprisingly given the diagnostic importance of cytokeratin stains for detecting tumours, the greatest increases in performance against the unprivileged baseline method were generally achieved for the pan-CK model, with similar gains for $\alpha$SMA. Notably, for the TIL task the immune-related stains CD3CD20 and CD45RB achieved the best performance, as this privileged information was more task relevant than others. Compared to the baseline unprivileged method, there was less benefit for pairing CD235a or ERG, perhaps because red blood cells (stained by CD235a) and the endothelium (stained by ERG) were less relevant to the tasks being assessed. Still, compared to the baseline privileged method, performance on CD235a and ERG was significantly improved.

We see that in Siamese models, some stains help improve prediction accuracy while others hinder it. For example, in the PanNuke neoplastic cell detection task, privileged Siamese training is considerably less accurate for MIST1 and ERG stains, which stain plasma and endothelial cells respectively, while it is more accurate for $\alpha$SMA and pan-CK, which stain smooth muscle cells/myofibroblasts and epithelial cells respectively. We show in Figure C.7 that this difference in performance is associated with differences in the proportion of empty space in the privileged information (see Section 3.6 for further discussion). While all stainings provide valuable information about their specific cell types, some have very few features which can be learned (see Appendix D for more details). This causes the primary encoder's representations to collapse and perform poorly on downstream tasks, as Siamese models can only learn features which are shared between branches. TriDeNT Ψ, in contrast, can retain the features from both primary and privileged information, leading to improved performance over unprivileged models. In Section 3.6 we will further discuss how TriDeNT Ψ can mitigate the effects of harmful or uninformative privileged information, compared to Siamese methods. We also note that there are some small differences in the distributions and sample sizes of the tissue samples used for different stains. For example, in the most extreme case the unprivileged models have a range of 0.1399 for their accuracies on the Singapore gland malignancy detection task.

The performance improvements observed in patch-level tasks are found to carry over into slide-level tasks, which are shown in Table S5. These tasks are generally considered to be of more clinical relevance than patch-level tasks, as they involve making predictions on the level of the patient. We consider here the detection and grading of dysplasia, and the ISUP grading of tumours. We find that a simple aggregation of the representations from models trained with TriDeNT Ψ is an effective predictor on these tasks, with strong performance across all tasks.

#### 3.1.1. TriDeNT Ψ pretrained models outperform supervised models on small datasets

In Fig. 3(c) we demonstrate that TriDeNT Ψ consistently retains a higher level of performance as less classifier training data is used. Notably, there are dataset sizes where supervised and privileged Siamese models collapse to a trivial solution, while TriDeNT Ψ continues to perform well. The ability to learn well from tiny, few-shot classification datasets is evidence of the utility of models trained with TriDeNT Ψ for a variety of downstream applications, as in many biomedical settings there are very few samples available for a given topic of interest. TriDeNT Ψ can allow researchers and clinicians to make use of these few-shot datasets to enable the study of previously unworkable datasets.

### 3.2. Embedding knowledge from additional brightfield images

To demonstrate the generality of the method, we train models on the BCI dataset of paired H&E and brightfield IHC patches. We only perform evaluation on the BCI, NCT and PanNuke datasets, as the BCI dataset is a breast cancer dataset, while the Singapore and MHIST datasets are prostate and colorectal polyp specimens respectively, which are far out of the training distribution. We include the NCT dataset, despite comprising only colorectal tissue, as these patches are well curated into different tissue type classes which mostly bear a strong resemblance to those in breast cancer samples. Strikingly, TriDeNT Ψ outperforms the supervised baseline by a large margin on the BCI task. We propose that there may be features weakly present in H&E stains which are highly predictive of HER2 status, and the pairing with IHC stains which contain those features very strongly results in this improved performance.

As Table 2 shows, we find that TriDeNT Ψ is also highly effective on all tasks compared to the unprivileged Siamese baseline. As there is more information in the privileged paired data, the privileged baseline is considerably higher for this task. Despite this, TriDeNT Ψ still outperforms both comparable baselines on all tasks but one, achieving improvements of up to 25.1% compared to the unprivileged baseline and up to 5.8% compared to the privileged baseline. There

**Table 3**
Results for models trained on the PanNuke dataset containing H&E patches paired with nuclear segmentation masks.

| Loss | Method | Privileged | NCT | PanNuke | Singapore | MHIST |
|---|---|---|---|---|---|---|
| VICReg | TriDeNT Ψ | ✓ | **0.7337** | **0.9106** | **0.7975** | **0.7523** |
| | Siamese | ✓ | 0.6000 | 0.8274 | 0.7106 | 0.6530 |
| | | ✗ | 0.7301 | 0.8682 | 0.7754 | 0.7421 |
| InfoNCE | TriDeNT Ψ | ✓ | **0.7530** | **0.9115** | **0.8226** | 0.7369 |
| | Siamese | ✓ | 0.5289 | 0.7403 | 0.6951 | 0.6264 |
| | | ✗ | 0.7199 | 0.8668 | 0.8015 | **0.7451** |
| CrossEntropy | Supervised | – | 0.9245 | 0.8901 | 0.9103 | 0.7042 |

is only a single task where TriDeNT Ψ does not improve performance: evaluation on the PanNuke dataset of a model trained with the VICReg loss, performing 1.6% less than the privileged baseline. The brightfield IHC stains contain considerably more task-relevant information for cell segmentation, so this is unsurprising. This effect can be understood visually as the IHC 'weak' quadrant in Fig. 2(a) being very narrow and containing very few features. Most task-relevant features are strongly present in the IHC and therefore there is less to be gained by adding the few missing weak features.

### 3.3. Image annotations are an effective source of privileged information

We find that TriDeNT Ψ is effective not only for integrating additional sources of data, but also for manually determining the most useful aspects of the data which should be learned, where the user has some prior knowledge to incorporate into the dataset. This is intuitively the opposite of traditional machine learning approaches, where the user has to handcraft inputs to be passed to the model, and the model only learns from those features. With our approach, the user can manually handcraft inputs, such as the segmentation masks in this example, while still giving the model the flexibility to learn other features not known a priori to the user. The results in Table 3 demonstrate that TriDeNT Ψ is able to train encoders which retain the features of both the nuclei and the background/connective tissue. We see performance improvements of up to 42.4% compared to the privileged baseline, and up to 5.2% compared to the unprivileged baseline.

These results also suggest that, in the privileged Siamese case, the features that are learned are those relating to the shape of the nuclei, rather than any sub-nuclear features or features relating to the connective tissue which would enable better identification of tissue and cell types.

### 3.4. Vision models with privileged spatial transcriptomics data learn more biologically relevant features

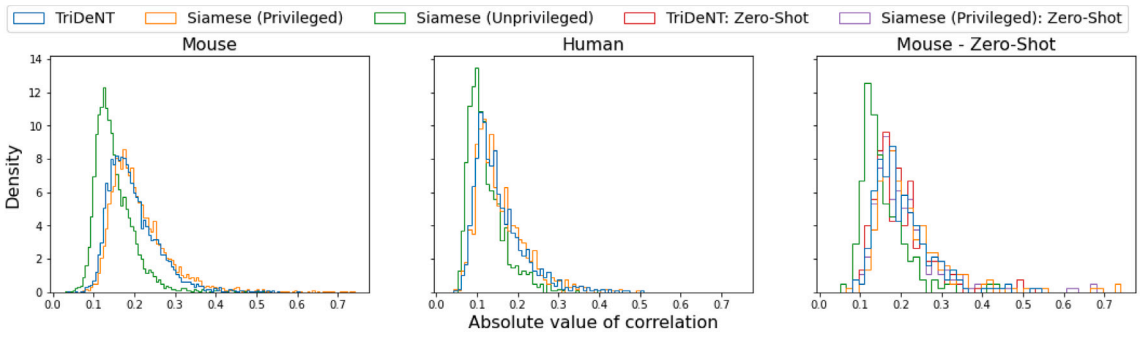
A key application of TriDeNT Ψ is the distillation of information from privileged sources beyond images. As TriDeNT Ψ does not require the architecture of each branch to be the same, it is possible to utilise any input type on any branch. We investigate the use of spatial transcriptomics (gene expression counts from an array of spatial points on a slide) as privileged data to train models for H&E inputs. These data have been shown to be highly informative and enable the study of the relationship between gene expression and tissue morphology, however, they are very expensive to generate, and as such are far from routine use. The difficulty of this task is compounded by the established poor performance of deep learning methods on tabular data (Shwartz-Ziv and Armon, 2022; Borisov et al., 2022).

Despite this, we see consistent improvements of up to 4.4% for TriDeNT Ψ over other methods for the spatial transcriptomics white matter/grey matter classification task, as shown in Fig. 4(b). It is likely the case that there are some mislabelled examples due to the processes involved in alignment, so higher accuracy on this task may simply not be possible, which could explain the saturation of performance around 89% in the mouse example. We observe a similar improvement for VICReg on the genotype prediction task, with an improvement of up to 2.2%. InfoNCE shows a similar performance for TriDeNT Ψ and unprivileged Siamese models, which both outperform privileged Siamese.

To assess the level of information shared between the transcriptomic results and the representations of the H&E patches, we investigate the cross-correlation between elements of the representations and the gene counts for each matching patch. We calculate the cross-correlation across the validation set between each element in the representations and the count for each gene, and for each gene take the correlation of the corresponding element with the maximum correlation or minimum anti-correlation, whichever has the greater absolute value. This maximum/minimum is chosen because the vast majority of elements will not correlate with any given gene, and the absolute value is taken because the sign of the element is arbitrary, so correlation and anti-correlation are equivalent. We use the absolute value of the correlation for the element selected for each gene, and use these to generate the histograms in Fig. 4(a). It is clear that privileged training obtains representations which are far more correlated to the gene counts than unprivileged training, with minimal differences in the correlations between TriDeNT Ψ and Siamese approaches. This implies that the models have learned equivalently informative representations about the coarse-grained features of the genes. Fig. 4(c) demonstrates that the correlation strength is significantly greater for TriDeNT Ψ compared to an unprivileged Siamese model, and Figures S1 and S2 show the relationships between the gene correlations of representations from TriDeNT Ψ, Siamese methods, and supervised learning. Figures S3 and S4 show the geneset enrichment for each method, demonstrating that TriDeNT Ψ captures more meaningful interrelationships that are more informative about the relationship between tissue morphology and gene expression than unsupervised Siamese models. This is especially important for scientific discovery, as these analyses are used to generate hypotheses for further research. Figure S5 shows UMAP projections of the representation space coloured by genotype and gene, to illustrate that TriDeNT Ψ identifies distinct morphological clusters which are not found by unprivileged Siamese models. Fig. 4(a) also shows that the findings are robust to human and mouse datasets, indicating the generality of the method.

We also demonstrate the correlations of the model's representations with genes unseen during the self-supervised phase of training (Fig. 4(a)). We find that the model is not simply overfitting on the given genes, as these genes are not present in the training data, yet the privileged models still demonstrate greater correlation with their counts than an unprivileged approach.

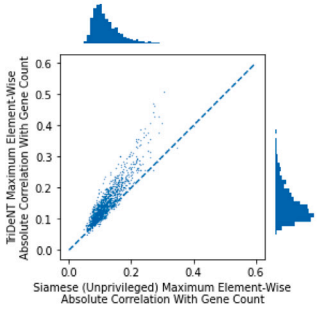
Existing approaches for integrating spatial gene expression data with tissue morphology (He et al., 2020) have focused on directly generating the transcriptomics from the H&E patches. While effective for predicting the expression of a given gene, this is highly ineffective for learning useful representations of the coarse-grained tissue features. Generative models have been shown to produce representations which are contain less semantic information than joint-embedding architectures and do not perform as well on downstream tasks (Assran et al., 2022, 2023). This has also been shown specifically for pathology images (Farndale et al., 2023). We confirm this by directly predicting the gene counts from the H&E patches, and show in Table 4(b) (Supervised (Transfer)) that transfer task performance is inferior to both the TriDeNT Ψ and Siamese approaches. We note that this leads to representations which do not generalise well, as is also true for other supervised methods and image-to-image translation methods.




(a)

| Loss | Method | Privileged | Human White/Grey | Mouse White/Grey | Mouse Genotype |
|---|---|---|---|---|---|
| VICReg | TriDeNT Ψ | ✓ | **0.8395** | **0.8861** | **0.5634** |
| | Siamese | ✓ | 0.8072 | 0.8833 | 0.5518 |
| | | ✗ | 0.8179 | 0.8741 | 0.5513 |
| InfoNCE | TriDeNT Ψ | ✓ | **0.8401** | **0.8870** | 0.5614 |
| | Siamese | ✓ | 0.8109 | 0.8856 | 0.5383 |
| | | ✗ | 0.8045 | 0.8782 | **0.5662** |
| MSE | Direct Gene Prediction | - | 0.7527 | 0.8669 | 0.5426 |
| CrossEntropy | Supervised | - | 0.8590 | 0.8720 | 0.4933 |

(b)




(c)

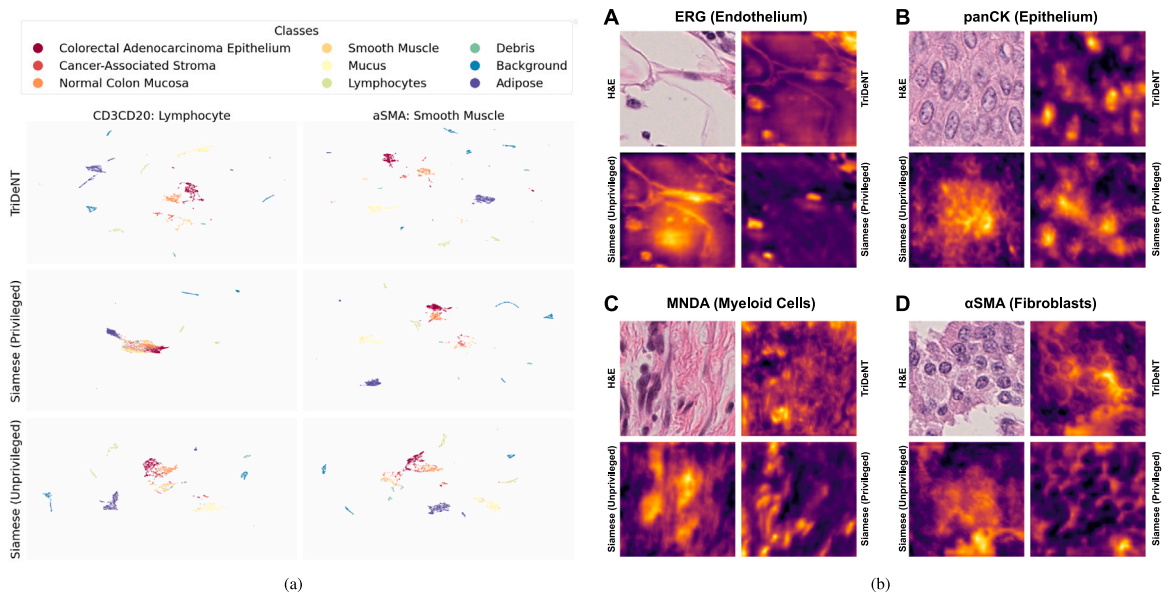
**Fig. 4.** (a) Correlation histograms between representations and gene count arrays for mouse and human ALS-ST data. Bins are chosen using the maximum of the Sturges (Sturges, 1926) and Freedman-Diaconis (Freedman and Diaconis, 1981) estimators. In the third histogram, zero-shot models are evaluated on genes which were not seen during training, while other models which did see those genes in training are evaluated on the same genes for comparison (of course, unprivileged models never see any genes). Comparison with models which saw these genes during training. (b) Spatial transcriptomics results for white/grey matter classification with both VICReg and InfoNCE losses. Baselines provided are 'Direct Gene Prediction', where a supervised model is trained to predict the gene counts for that patch directly and the representation is then fine-tuned on the white/grey classification task, and a standard supervised model. (c) Greater correlation strengths between gene counts and representations of TriDeNT Ψ models than unprivileged Siamese models. For each gene, the maximum absolute correlation between the TriDeNT Ψ representations for each patch and the corresponding gene counts are plotted against those for unprivileged Siamese representations, with TriDeNT Ψ almost always achieving greater correlation strength. Dashed line is the identity. Appended histograms show distribution of data. Mouse data only, see Figure S2 for human data, which shows a similar pattern, and Figure S1 for extended comparisons of mouse data, also including privileged Siamese and supervised results.

### *3.5. TriDeNT Ψ identifies features of both primary and privileged inputs from primary input alone*

To further analyse the learned representations, we produce UMAP projections of the latent space labelled with the tissue types for the NCT tissue type classification task, as shown for CD3CD20 and $\alpha$SMA in Fig. 5(a), and for all SegPath stains in Figures S6 and S7. These figures make the reasons for the varying performance of the privileged Siamese model more apparent. For stains with better performing privileged Siamese models, such as $\alpha$SMA, the UMAPs are very similar between Siamese methods and TriDeNT, with well-differentiated tissue type clusters. In those with worse performance, such as ERG, the tissue types are poorly differentiated, often with only adipose and background forming distinct clusters from the other classes. On closer inspection, it is notable in these projections that TriDeNT Ψ produces more well-defined and separated clusters in general than Siamese networks. This is further evidenced in Figure S5, where TriDeNT Ψ is shown to identify clusters with overexpression of a given gene significantly more effectively than an unprivileged Siamese model. Interestingly, we find that the privileged Siamese model for CD3CD20 forms distinct clusters for the *lymphocytes* class, which corresponds well to the privileged information. In contrast, ERG and MNDA appear similarly but without the presence of this cluster, suggesting that the privileged information impacts the presence of certain clusters.

We also analyse the activation maps for each model using GradCAM as described in Appendix C. This offers more insight into the areas of the image which are contributing most heavily to the models' representations. In Fig. 5(b) we present some representative examples, however, a larger selection which was chosen at random is presented in Figures S8 to S23. The larger selection makes it easier to see the emergent patterns. We see that unprivileged Siamese models tend to focus primarily on image features such as textures and colour, particularly when the image contains white background. Privileged Siamese models tend to focus on regions associated with their privileged information, primarily cell nuclei for panels A, B, and C, and smooth muscle for D. TriDeNT Ψ occupies an intermediate position, incorporating both features specific to the privileged data and more the general features associated with unprivileged Siamese networks.

We can see in Figures S10 and S18 that for ERG, the privileged Siamese model focuses almost exclusively on nuclei. As there are very few endothelial cells in the dataset, it could be an effective strategy to identify anything that could potentially be an endothelial cell to minimise the difference between the representations of the H&E model and the IF mask model (see Appendix D for more details). In the corresponding unprivileged Siamese image, we see that the model identifies some of these nuclei, albeit less strongly, but also focuses heavily on the other tissue and even the background, while strongly fixating on two spots of debris in the centre of the image. This model has less 'incentive' to learn the weak features related to endothelial cells as these occur rarely and are not easy to detect, while more generic strong features such as the presence of connective tissue and the prevalence of background are more common and predictable from augmented images. We see that the TriDeNT Ψ ERG model also largely ignores nuclei, primarily focusing on the connective tissue, supporting the argument that TriDeNT Ψ learns to ignore the privileged information when it is not useful. We note that no VICReg model appears to focus on the endothelial cells in the images we have tested, however the InfoNCE TriDeNT Ψ and privileged Siamese models do successfully identify this cell (e.g. row 2, column 4 in Figures S10 and S18).




(a) (b)

**Fig. 5.** (a) Sample UMAP projections from 2048 dimensions into 2 for models trained on the SegPath CD3CD20 and $\alpha$SMA subsets, evaluated on the NCT test dataset. Points are coloured by tissue type. Note that accuracies for these tasks were i) TriDeNT Ψ: CD3CD20 0.8982, $\alpha$SMA 0.9273; Siamese (Privileged): CD3CD20 0.6625, $\alpha$SMA 0.9186; Siamese (Unprivileged): CD3CD20 0.8694, $\alpha$SMA 0.8570. (b) GradCAM heatmaps for selected images from the SegPath dataset. Evaluated with VICReg loss. Brighter colours represent greater activation strengths. For a larger selection, including heatmaps for InfoNCE models, see Figures S8 to S23.

In panel C we see a similar pattern, with the privileged Siamese model fixating solely on the nuclei, while the TriDeNT Ψ model takes a more balanced approach. The unprivileged Siamese model appears to focus on a single cluster of nuclei while neglecting others, and similarly identifies an area of fibroblasts with its distinctive pattern but does not others.

In contrast to panels A and C which represent models with poor privileged Siamese results, panels B and D represent models whose privileged Siamese results were comparable to both TriDeNT Ψ and even the supervised baseline. It is therefore interesting to note that there are far more similarities between the privileged Siamese and TriDeNT Ψ models in both cases. Particularly in panel B, TriDeNT Ψ and the privileged Siamese model return virtually identical heatmaps, with both strongly identifying epithelial nuclei and neglecting the same areas of connective tissue. The unprivileged model in this case appears to focus solely on the centre of the image, giving a significantly different heatmap to the other panels.

Panel D again shows the previous pattern, with the privileged Siamese model identifying the features strongly present in the privileged data – fibroblasts – while neglecting the nuclei present. TriDeNT Ψ also strongly identifies the connective tissue, but, unlike the privileged Siamese model, does not completely neglect the nuclei. The unprivileged Siamese model primarily identifies background, and does not appear to identify the nuclei in this example.

### 3.6. *TriDeNT Ψ mitigates harmful and uninformative privileged information*

TriDeNT Ψ is designed to integrate privileged information, and the normal assumption is that this privileged information is useful. However, the types of information found in real medical data are highly heterogeneous and can contain both information that highly useful and information that is completely irrelevant. This is studied with two scenarios: blank privileged information, and randomly shuffled privileged information. Blank privileged information provides no information and can only be detrimental to performance, and randomly shuffled privileged information contains information that has no correspondence to the primary H&E patch it is paired with. The objective is to assess the ability of TriDeNT Ψ, and comparable baselines, to mitigate the influence of this irrelevant or harmful privileged information.

Table 4 shows that TriDeNT Ψ achieves comparable performance to unprivileged models, implying that TriDeNT Ψ is able to ignore the irrelevant and harmful privileged information, and only learn features from the primary input. In fact, in some cases with randomly shuffled patches TriDeNT Ψ marginally improves performance compared to unprivileged training. This could be circumstantial to the dataset and requires further research to assess its validity, although an intuitive explanation for the improvement could be that the irrelevant information still contains some common features between patches, such as the shapes of nuclei, which encourage the primary encoder to learn to detect similar round shapes in the H&E.

In contrast, the performance of privileged Siamese models is very poor, with this baseline achieving the worst performance on every test dataset. This is because by mapping into a single shared latent space, the primary and privileged models can only learn features which are shared between inputs, and nothing can be learned from vacuous inputs.

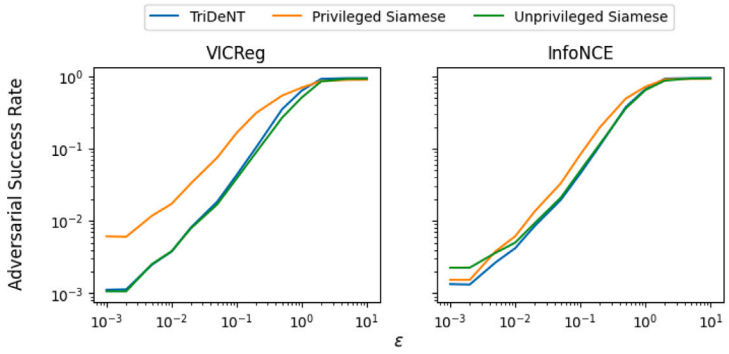
Despite TriDeNT employing no explicit feature selection mechanisms, we see that this method can dynamically select features which optimise its objective. When privileged information is useful, features are selected which are correlated with the privileged information. When the privileged information is irrelevant or harmful, it is ignored in favour of the primary features that would be learned by unprivileged methods. This is critical for the real-world usefulness and general applicability of TriDeNT Ψ, which can be used to effectively distil any source of privileged information without the potential for damaging performance, as routinely happens with privileged models.

### 3.7. *Robustness of TriDeNT Ψ to domain shift and adversarial attacks*

A growing concern with computational pathology models is the robustness of models to discrepancies between their training data and evaluation data. This can be either in the form of unintentional differences caused by different image acquisition methods, causing a domain shift, or in the form of intentional adversarial attacks (Ghaffari Laleh et al., 2022). These are malicious attacks that are designed to augment the input data in a way that causes the model to make an incorrect classification while being imperceptible to humans. The robustness of models to these attacks is of interest for translation of models to clinical

**Table 4**
Results for models trained on the PanNuke dataset containing H&E patches paired with redundant or harmful privileged information – blank patches or randomly shuffled nuclear segmentation masks – to assess whether TriDeNT Ψ can mitigate the impact of detrimental privileged information. We denote the best performance in a category in bold, and the second best with an underline.

| Paired Data | Loss | Method | Privileged | NCT | PanNuke | Singapore | MHIST |
|---|---|---|---|---|---|---|---|
| Blank Patches | VICReg | TriDeNT Ψ | ✓ | 0.6943 | 0.8125 | 0.7592 | 0.7257 |
| | | Siamese | ✓ | 0.5378 | 0.7227 | 0.6142 | 0.5670 |
| | | | ✗ | **0.7301** | **0.8682** | **0.7754** | **0.7421** |
| | InfoNCE | TriDeNT Ψ | ✓ | **0.7680** | 0.8092 | 0.7875 | 0.7345 |
| | | Siamese | ✓ | 0.6466 | 0.7063 | 0.6200 | 0.6034 |
| | | | ✗ | 0.7199 | **0.8668** | **0.8015** | **0.7451** |
| Shuffled Patches | VICReg | TriDeNT Ψ | ✓ | **0.7429** | 0.8235 | **0.7932** | 0.7277 |
| | | Siamese | ✓ | 0.5075 | 0.7254 | 0.6313 | 0.5865 |
| | | | ✗ | 0.7301 | **0.8682** | 0.7754 | **0.7421** |
| | InfoNCE | TriDeNT Ψ | ✓ | **0.7598** | 0.7751 | 0.7803 | 0.7316 |
| | | Siamese | ✓ | 0.6267 | 0.7084 | 0.6269 | 0.6689 |
| | | | ✗ | 0.7199 | **0.8668** | **0.8015** | **0.7451** |




(a)

| Loss | Method | Privileged | SSR ↓ |
|---|---|---|---|
| VICReg | TriDeNT Ψ | ✓ | 0.6850 |
| | Siamese | ✓ | 1.6903 |
| | | ✗ | **0.6248** |
| InfoNCE | TriDeNT Ψ | ✓ | **0.8679** |
| | Siamese | ✓ | 1.1392 |
| | | ✗ | 0.9929 |

(b)

**Fig. 6.** (a) Average adversarial robustness to PGD attacks for models trained on Segpath stains. For full results see Tables S13 and S14. (b) Standardised Success Rate (SSR) for these models. Lower values are better, with bold indicating best performance and underline indicating second best.

use due to the criticality of their predictions to human health. These models must not be vulnerable to small perturbations, either as a result of malicious actors or unintentional variation in measurement.

A concern with TriDeNT Ψ and other models using privileged information could be that the successful distillation of information leads models to be less robust. Privileged models are learning at least some weaker features than unprivileged models, and consequently could in theory be easier to exploit than unprivileged models which learn only strong features in the primary information (see Section 2.3 and Fig. 2(a) for definitions and discussion of weak/strong features).

In Fig. 3 we have presented results on the Camelyon test set, which assesses models' domain transfer performance when trained on images from three hospitals and evaluated on images from another, with large differences between hospitals. This shows that TriDeNT Ψ can achieve very strong performance on different domains, implying that the features learned with the TriDeNT Ψ training regime are more robust. This is in contrast to unprivileged models, which achieve little better than random guessing.

We present results for the adversarial robustness of all SegPath models in Fig. 6. We used a white-box adversarial attack – Projected Gradient Descent (PGD) (Madry et al., 2017) – to assess the robustness of each model, and find that TriDeNT Ψ has a similar robustness to unprivileged models, while privileged Siamese models are considerably less robust. This is because TriDeNT Ψ retains the strong primary features as well as the weaker, privileged features, and consequently is less affected than the privileged Siamese models which are reliant solely on the weaker privileged features. This is an important finding for the clinical relevance of TriDeNT Ψ compared to unprivileged models, as it implies that not only do these models learn features which are more robust to the overt feature shift of domain transfer, they also learn features which are similarly robust to potential adversarial attacks. The results in Fig. 6(b) show a standardised success rate metric, which is defined as the mean standardised value for each model

$$SSR_m := \frac{1}{|E|} \sum_{\epsilon \in E} \frac{SR_{m\epsilon}}{\|SR_\epsilon\|}, \tag{11}$$

where $m$ is a model in the set of models (TriDeNT Ψ, privileged Siamese, and unprivileged Siamese), $E$ is the set of perturbation strengths $\epsilon$, and $SR_{m\epsilon}$ is the success rate of the attack with perturbation strength $\epsilon$ on model $m$. This determines the magnitude of the difference between models across all values of $\epsilon$, while accounting for the different scales of these values, with a smaller value indicating better adversarial robustness.

## 4. Discussion and conclusions

In this work, we have proposed TriDeNT Ψ, a modelling approach which has been demonstrated to effectively integrate privileged sources of data into single-source models during training to improve performance. The model works by providing two supervisory signals to the primary encoder, which dynamically respond to the features which the primary encoder can extract. Experiments have shown that this approach can greatly outperform standard Siamese privileged and unprivileged methods, and even supervised learning, without significantly increasing the computational overheads. There are a vast number of biomedical datasets which contain paired data, such as paired -omics datasets (Weinstein et al., 2013; Schorn et al., 2021), different imaging methods (e.g. PESO Bulten et al., 2019), and even multiple images of the same source, such as the 7-pt skin lesion dataset (Kawahara et al., 2018), CheXpert (Irvin et al., 2019) and CheXphoto (Phillips et al., 2020).

### 4.1. Integrating privileged data is invaluable for research and discovery

The utility of TriDeNT Ψ for research applications can be found not only in increasing the efficacy of primary data models for prediction accuracy, but also in training models to extract coarse-grained features which are relevant to the privileged input. Our results demonstrate that models trained with TriDeNT Ψ will perform better on tasks where the privileged information is more task relevant. This is demonstrated, for example, where H&E prediction of HER2+ status is greatly improved by pairing with HER2 IHC stains (Table 2), where immune-related privileged SegPath stains lead to better performance on the TIL task (Fig. 3(a)), and where performance on metastasis- and malignancy-related tasks are most improved by privileged pan-CK stains (Fig. 3(a)). This is also shown qualitatively with the GradCAM activation heatmaps in Fig. 5(b). A typical use case is that a scientist with a paired dataset could train a model to then evaluate an unpaired dataset, without needing to acquire more paired data. We have shown that the features which are found by privileged methods are significantly different from those found by unprivileged methods. This means that TriDeNT Ψ could enable the identification of novel morphological clusters that are functionally important, such as those in our analyses in Figs. 5(a) and S5 which might not emerge from other methods of training or training on the new dataset alone.

### 4.2. TriDeNT Ψ does not need large datasets

Self-supervised methods typically require very large datasets (Reed et al., 2022), however our results, especially those for PanNuke (Section 3.3) and ALS-ST (Section 3.4), demonstrate that TriDeNT Ψ offers improvements over comparable baselines for comparatively small pretraining datasets. We also studied the effect of evaluation dataset size in Section 3.1.1, showing that TriDeNT Ψ continues to achieve strong performance even when the classifier head is trained on a tiny dataset. This performance can only be expected to improve further if pretrained models are used, either from a general source such as ImageNet (Russakovsky et al., 2015) or from more specific pretraining tasks. We expect that this would be particularly useful in cases where the privileged paired model is pretrained, as teacher–student distillation would likely lead to greater performance in the student (primary) model.

### 4.3. TriDeNT Ψ can incorporate image annotations into representation learning

Our experiments have demonstrated that models trained using TriDeNT Ψ learn significantly different features to those trained in standard self-supervised settings, and that this can be leveraged to manually encode information by the user. For example, we demonstrated in Section 3.3 that the model can be made to learn features related to nuclear segmentation masks, without requiring human prior knowledge of what those precise features might be. This offers new opportunities to make better use of the manual annotations which are provided with many datasets but typically only used as target labels for supervised learning. We have shown that these annotations can be used to create more generalist, robust and effective models when transferred to other tasks, either related or unrelated to the annotations.

Of course, not all annotations are manual, and machine-generated annotations, such as those from HoVer-Net (Graham et al., 2019), could be incorporated into training procedures. Currently there exist a huge number of models which have been trained for one specific task, such as nuclear segmentation (e.g. Graham et al. 2019), tissue type annotation (e.g. Kather et al. 2016), virtual restaining (e.g. Xu et al. 2019), feature detection (e.g. Aubreville et al. 2023), etc., and all of these could be incorporated into new generalist models using TriDeNT Ψ.

### 4.4. Future research

While TriDeNT Ψ offers a new capability for multi-modal distillation in medical imaging, further improvements can be made. Model weights were always frozen for downstream tasks, so the tasks detailed in this work are all zero-shot, meaning fine-tuning these models could lead to improved performance. Design choices were primarily made for simplicity and parity with previous work, and hyperparameters were chosen based on previous work on Siamese networks (e.g. Bardes et al. 2021, Chen et al. 2020), so it is highly likely that the results are skewed in favour of these Siamese networks. Training was also only carried out for 100 epochs (200 for the spatial transcriptomics examples) on a batch size of 128, so models could potentially improve further with longer training times and larger batch sizes. Despite this, we have shown that TriDeNT Ψ outperforms these methods, often by a considerable margin. Improvements could be made by adjusting the loss function, or by implementing more elaborate interactions between branches. The scope of this study was limited to histopathology, however TriDeNT Ψ could be broadly applicable to other domains in both imaging and other modalities. We expect TriDeNT Ψ to have extensive applications for multiplexed imaging, as the best way of integrating these multiple sources of information has not been established.

We also anticipate that utilising different network architectures on different branches could yield interesting results, such as pairing convolutional neural networks (CNNs) with graph neural networks, transformers or simply a larger CNN. We showed this is a possibility in Section 3.4, where a primary CNN is paired with an multilayer perceptron for the privileged spatial transcriptomics data. This would enable different features and patterns to be identified and could lead to models which utilise the efficiency of CNNs with the power of these additional methods.

## Code and data availability

The TriDeNT Ψ codebase is available at github.com/lucasfarndale/TriDeNT. All datasets used are publicly available from the following links:

- ALS-ST – als-st.nygenome.org;
- BCI – bupt-ai-cz.github.io/BCI/;
- Camelyon – wilds.stanford.edu/datasets/;
- IMP 1K/4K – rdm.inesctec.pt/km/dataset/nis-2023-008;
- IMP Cervix – rdm.inesctec.pt/km/dataset/nis-2024-003;
- MHIST – bmirds.github.io/MHIST;
- NCT Colorectal Cancer – 10.5281/zenodo.1214455;
- PANDA – kaggle.com/c/prostate-cancer-grade-assessment/data;
- PanNuke – warwick.ac.uk/fac/cross_fac/tia/data/pannuke;
- SegPath – dakomura.github.io/SegPath;
- Singapore Prostate Cancer – 10.5281/zenodo.7152243.
- TIL – 10.5281/zenodo.6604094;

The exact patchings and dataset splits are available from the authors upon reasonable request where this is permitted by the dataset's license.

## CRediT authorship contribution statement

**Lucas Farndale:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Robert Insall:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Ke Yuan:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Lucas Farndale is supported by the MRC, United Kingdom grant MR/W006804/1, Robert Insall is supported by EPSRC, United Kingdom grant EP/S0300875/1 and Wellcome grant, United Kingdom 221786/Z/20/Z. Ke Yuan acknowledges support from EPSRC, United Kingdom EP/R018634/1, Cancer Research UK (EDDPGM-Nov21/100001 and DRCMDP-Nov23/100010), BBSRC BB/V016067/1 and Prostate Cancer UK MA-TIA22-001.

The authors would like to extend our gratitude to Adalberto Claudio-Quiros and Kai Rakovic for the helpful feedback and discussion.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2025.103479.

## Data availability

Links are provided in the manuscript text.

## References

Abousamra, S., Gupta, R., Hou, L., Batiste, R., Zhao, T., Shankar, A., Rao, A., Chen, C., Samaras, D., Kurc, T., et al., 2022. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. Front. Oncol. 11, 806603.

Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A., 2017. Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992.

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N., 2022. Masked siamese networks for label-efficient learning. In: European Conference on Computer Vision. Springer, pp. 456–473.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N., 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629.

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopfleisch, R., Ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., et al., 2023. Mitosis domain generalization in histopathology images—the MIDOG challenge. Med. Image Anal. 84, 102699.

Balestriero, R., LeCun, Y., 2024. Learning by reconstruction produces uninformative features for perception. arXiv preprint arXiv:2402.11337.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al., 2018. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans. Med. Imaging.

Bardes, A., Ponce, J., LeCun, Y., 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G., 2022. Deep neural networks and tabular data: A survey. IEEE Trans. Neural Networks Learn. Syst..

Bulten, W., Bándi, P., Hoven, J., Loo, R.v.d., Lotz, J., Weiss, N., Laak, J.v.d., Ginneken, B.v., Hulsbergen-van de Kaa, C., Litjens, G., 2019. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. Sci. Rep. 9 (1), 864.

Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al., 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature Med. 28 (1), 154–163.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. 33, 9912–9924.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S., 2017. Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612.

Farndale, L., Insall, R., Yuan, K., 2023. More from less: Self-supervised knowledge distillation for routine histopathology data. In: Cao, X., Xu, X., Rekik, I., Cui, Z., Ouyang, X. (Eds.), Machine Learning in Medical Imaging. Springer Nature Switzerland, Cham, pp. 454–463.

Farndale, L., Walsh, C., Insall, R., Yuan, K., 2024. Synthetic privileged information enhances medical image representation learning. arXiv preprint arXiv:2403.05220.

Feyereisl, J., Aickelin, U., 2012. Privileged information for data clustering. Inform. Sci. 194, 4–23.

Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator: L 2 theory. Z. Wahrscheinlichkeitstheorie Und Verwandte Geb. 57 (4), 453–476.

Gamper, J., Koohbanani, N.A., Benes, K., Khuram, A., Rajpoot, N., 2019. Pan-Nuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European Congress on Digital Pathology. Springer, pp. 11–19.

Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., et al., 2022. Adversarial attacks and adversarial robustness in computational pathology. Nat. Commun. 13 (1), 5711.

Ghahremani, P., Marino, J., Hernandez-Prera, J., de la Iglesia, J.V., Slebos, R.J., Chung, C.H., Nadeem, S., 2023. An AI-ready multiplex staining dataset for reproducible and accurate characterization of tumor immune microenvironment. arXiv preprint arXiv:2305.16465.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. 58, 101563.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284.

He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J., 2020. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng. 4 (8), 827–834.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023. A visual–language foundation model for pathology image analysis using medical Twitter. Nature Med. 1–10.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, (01), pp. 590–597.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. PMLR, pp. 4904–4916.

Jing, L., Vincent, P., LeCun, Y., Tian, Y., 2021. Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348.

Kaczmarzyk, J., Abousamra, S., Kurc, T., Gupta, R., Saltz, J., 2022. Dataset for tumor infiltrating lymphocyte classification (304,097 image patches from TCGA). URL https://doi.org/105281.

Karaletsos, T., Belongie, S., Rätsch, G., 2015. Bayesian representation learning with oracle constraints. arXiv preprint arXiv:1506.05011.

Kather, J.N., Halama, N., Marx, A., 2018. 100,000 histological images of human colorectal cancer and healthy tissue. http://dx.doi.org/10.5281/zenodo.1214456.

Kather, J.N., Weis, C.-A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Sci. Rep. 6 (1), 1–11.

Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2018. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Heal. Inform. 23 (2), 538–546.

Kiela, D., Bottou, L., 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 36–45.

Komura, D., Onoyama, T., Shinbo, K., Odaka, H., Hayakawa, M., Ochi, M., Herdiantoputri, R.R., Endo, H., Katoh, H., Ikeda, T., et al., 2023. Restaining-based annotation for cancer histology segmentation to overcome annotation-related limitations among pathologists. Patterns 4 (2).

Lambert, J., Sener, O., Savarese, S., 2018. Deep learning under privileged information using heteroscedastic dropout. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8886–8895.

LeCun, Y., 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27.

Lee, D., Aune, E., Langet, N., Eidsvik, J., 2022. Vnibcreg: Vicreg with neighboring-invariance and better-covariance evaluated on non-stationary seismic signal time series. arXiv preprint arXiv:2204.02697.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J., 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208.

Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., Jin, M., 2022. BCI: Breast cancer immunohistochemical image generation through pyramid Pix2pix. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1815–1824.

Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V., 2015. Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Maniatis, S., Äijö, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., Fagegaltier, D., Andrusivová, Ž., Saarenpää, S., Saiz-Castro, G., et al., 2019. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science 364 (6435), 89–93.

Neto, P.C., Montezuma, D., Oliveira, S.P., Oliveira, D., Fraga, J., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Reinhard, S., et al., 2024. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. Npj Precis. Oncol. 8 (1), 56. http://dx.doi.org/10.1038/s41698-024-00539-4.

Neto, P.C., Oliveira, S.P., Montezuma, D., Fraga, J., Monteiro, A., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S., 2022. iMIL4PATH: A semi-supervised interpretable approach for colorectal whole-slide images. Cancers 14 (10), 2489.

Oliveira, S.P., Montezuma, D., Moreira, A., Oliveira, D., Neto, P.C., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S., 2023. A CAD system for automatic dysplasia grading on H&E cervical whole-slide images. Sci. Rep. 13, 3970. http://dx.doi.org/10.1038/s41598-023-30497-z.

Oliveira, S.P., Neto, P.C., Fraga, J., Montezuma, D., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., Cardoso, J.S., 2021. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. Sci. Rep. 11 (1), 1–15. http://dx.doi.org/10.1038/s41598-021-93746-z.

Oner, M.U., Ng, M.Y., Giron, D.M., Xi, C.E.C., Xiang, L.A.Y., Singh, M., Yu, W., Sung, W.-K., Wong, C.F., Lee, H.K., 2022. An AI-assisted tool for efficient prostate cancer diagnosis in low-grade and low-volume cases. Patterns 3 (12).

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Phillips, N.A., Rajpurkar, P., Sabini, M., Krishnan, R., Zhou, S., Pareek, A., Phu, N.M., Wang, C., Jain, M., Du, N.D., et al., 2020. CheXphoto: 10,000+ photos and transformations of chest X-rays for benchmarking deep learning robustness. In: Machine Learning for Health. PMLR, pp. 318–327.

Qiao, C., Li, D., Guo, Y., Liu, C., Jiang, T., Dai, Q., Li, D., 2021. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. Nature Methods 18 (2), 194–202.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillory, D., Metzger, S., et al., 2022. Self-supervised pretraining improves self-supervised pretraining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2584–2594.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al., 2020. The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. Cell 181 (2), 236–249.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R., 2016. Progressive neural networks. arXiv preprint arXiv:1606.04671.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al., 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 23 (1), 181–193.

Schorn, M.A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D.D., Aksenov, A.A., Aleti, G., Moghaddam, J.A., Aron, A.T., Aziz, S., et al., 2021. A community resource for paired genomic and metabolomic data mining. Nat. Chem. Biol. 17 (4), 363–368.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: Deep learning is not all you need. Inf. Fusion 81, 84–90.

Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., Shen, P., 2021. An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. Front. Digit. Heal. 3, 637386.

Sturges, H.A., 1926. The choice of a class interval. J. Amer. Statist. Assoc. 21 (153), 65–66.

Vapnik, V., Izmailov, R., 2017. Knowledge transfer in SVM and neural networks. Ann. Math. Artif. Intell. 81 (1–2), 3–19.

Vapnik, V., Vashist, A., 2009. A new learning paradigm: Learning using privileged information. Neural Netw. 22 (5–6), 544–557.

Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., et al., 2021. A petri dish for histopathology image analysis. In: Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings. Springer, pp. 11–24.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The cancer genome atlas pan-cancer analysis project. Nature Genet. 45 (10), 1113–1120.

Xu, Z., Huang, X., Moro, C.F., Bozóky, B., Zhang, Q., 2019. GAN-based virtual re-staining: a promising solution for whole slide image analysis. arXiv preprint arXiv:1901.04059.

Yang, H., Tianyi Zhou, J., Cai, J., Soon Ong, Y., 2017. MIML-FCN+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1577–1585.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S., 2021. Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. PMLR, pp. 12310–12320.

**Lucas Farndale** is a 2nd year Ph.D. student at the CRUK Scotland Institute and the University of Glasgow. His research focuses on developing new computational and mathematical methods for biomedical imaging, with a focus on multimodal data integration. He received an MSci Mathematics degree from the University of Glasgow in 2022.

**Robert Insall** is Professor of Computational Cell Biology at University College London and Professor of Mathematical and Computational Cell Biology at the University of Glasgow. A true cross-disciplinary researcher, trained in biochemistry and cell biology but with publications in mathematics, physics and computer science. Approaches include biochemistry and quantitative microscopy, mathematical and computational modelling, cell biology and machine learning. Recently funded by Wellcome Trust, MRC, EPSRC, InnovateUK, Datalabs Edinburgh, and CRUK.

**Ke Yuan** is a Senior Lecturer in Machine Learning and Computational Biology jointly appointed at the Schools of Computing Science, the School of Cancer Sciences at the University of Glasgow and the Cancer Research UK Scotland Institute. Before Glasgow, he was a postdoctoral research fellow at the Cancer Research UK Cambridge Institute from 2012 to 2016. He received an M.Sc. and Ph.D. from the University of Southampton in 2008 and 2013, respectively.