Self-supervised deep learning for highly efficient spatial immunophenotyping

Hanyun Zhang[1,2], Khalid AbdulJabbar[1,2], Tami Grunewald[3], Ayse U Akarca[4], Yeman Hagos[1,2], Faranak Sobhani[1,2], Catherine SY Lecat[5], Dominic Patel[5], Lydia Lee[5], Manuel Rodriguez-Justo[5], Kwee Yong[5], Jonathan Ledermann[3,6], John Le Quesne[7,8,9], Shelley Hwang[10], Teresa Marafioti[4], Yinyin Yuan[1,2*]

[1]Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK
[2]Division of Molecular Pathology, The Institute of Cancer Research, London, UK
[3]University College London Hospital, London, UK
[4]Department of Cellular Pathology, University College London Hospital, London, UK
[5]Research Department of Hematology, Cancer Institute, University College London, UK
[6]UCL Cancer Institute and UCL Hospitals, Department of Oncology, University College London, UK
[6]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK
[7]Cancer Research UK Beatson Institute, Glasgow, UK
[8]Institute of Cancer Sciences, University of Glasgow, Glasgow, UK
[9]NHS Greater Glasgow and Clyde Pathology Department, Queen Eliabeth University Hospital, UK
[10]Department of Surgery, Duke University School of Medicine, Durham, USA
*Corresponding author.

Corresponding author's email:
Yinyin Yuan, yyuan6@mdanderson.org

Summary

Background

Efficient biomarker discovery and clinical translation depends on fast and accurate analytical output from crucial technologies such as multiplex imaging. However, reliable cell classification often requires extensive annotations. Label efficient strategies are urgently needed to reveal diverse cell distribution and spatial interactions in large-scale multiplex dataset.

Methods

This study proposed Self-supervised Learning for Antigen Detection (SANDI) for accurate cell phenotyping while mitigating the annotation burden. The model first learns intrinsic pair-wise similarities in unlabelled cell images, followed by a classification step to map learnt features to cell labels using a small set of annotated references. We acquired four multiplexed immunohistochemistry dataset and one imaging mass cytometry dataset, comprising 2825 to 15258 single cell images to train and test the model. The efficacy of SANDI was tested among various annotation burdens. We further assessed the potential of SANDI to identify biological meaningful cell-cell interactions.

Findings

With 1% annotations (18 – 114 cells), SANDI achieved weighted F1-scores ranging from 0.82 to 0.98 across the five datasets, which outperformed the other self-supervised methods and was comparable to the fully supervised classifier trained on 1828 - 11459 annotated cells (-0.002 - -0.053 of weighted F1-score). In ovarian cancer, analysis of this single cell data reveals spatial expulsion between PD1 expressing T helper cells and T regulatory cells, suggesting an interplay between PD1 expression and T regulatory cell-mediated immunosuppression.

Interpretation

By striking a fine balance between minimal expert guidance and the power of deep learning to learn similarity within abundant data, SANDI presents new opportunities for efficient, large-scale learning for histology multiplex imaging data.

Funding

Y.Y. acknowledges funding from Cancer Research UK Career Establishment Award (CRUK C45982/A21808), CRUK Early Detection Program Award (C9203/A28770), CRUK Sarcoma Accelerator (C56167/A29363), CRUK Brain Tumor Award (C25858/A28592), Breast Cancer Now (2015NovPR638), Rosetrees Trust (A2714), Children's Cancer and Leukaemia Group (CCLGA201906), NIH U54 CA217376, NIH R01 CA185138, CDMRP Breast Cancer Research Program Award BC132057, European Commission ITN (H2020-MSCA-ITN-2019), and The Royal Marsden/ICR National Institute of Health Research Biomedical Research Centre.

Keywords

Deep learning; Self-supervised learning; Cell classification; Multiplex imaging; Multiplex Immunohistochemistry; Imaging mass cytometry; Biomarkers

Research in context

Evidence before this study

We searched PubMed (https://pubmed.ncbi.nlm.nih.gov/) for studies using self-supervised learning or weakly-supervised learning to identify cell phenotypes in multiplex images. We found two relevant studies (PMID: 35758799; PMID: 35217454). In 2022, Murphy et al. trained a self-supervised model to learn features relevant to the targeted genes from unlabelled immunohistochemistry images of kidney, and to predict the cell type specificity of the image using single-cell transcriptomic data as references. The approach estimated the presence of cell types in a tissue region stained with a single marker but was unable to locate and classify single cells defined by a combination of antibodies. In the same year, Daniel Jiménez-Sánchez et al. proposed a deep learning framework to associate clinical characteristics of the patient to tumour microenvironment elements inferred from the multiplex-stained cancer tissues. The study regarded cell phenotyping as a side product of the pipeline, which was designed to reveal the cell types contributed to the clinical parameters rather than unbiasedly classify all cells targeted by the markers. To date, dedicated approaches for cell classification on multiplex images, especially on multiplex immunohistochemistry images where the intensities and combinations of staining are inferred from RGB images, has not yet been proposed.

Added value of this study

The current study designed a self-supervised based pipeline for label-efficient cell classification on multiplex immunohistochemistry and mass cytometry images. The method was evaluated on five datasets containing slides from ovarian cancer, lung squamous cell carcinoma, ductal carcinoma in situ, myeloma and pancreas. The method dramatically reduced the annotation to 1%, equalling to 18 – 114 cells across five datasets, while achieving a performance comparable to the model trained on 1828 - 11459 cells. Therefore, in the context of current research, this new study presents an efficient and accurate method with new functionalities to 1) classify single cells on multiplex stained tissue sections with a small set of user-specified examples. 2) be adopted to multiplex immunohistochemistry images without colour deconvolution or marker channel separation. 3) automatically recommended prone-to-misclassified cells for manual correction to efficiently improve model performance. 4) facilitate hypothesis-driven anlaysis of cellular spatial distributions on a large scale.

Implications of all the available evidence

By mitigating the annotation burden for accurate cell classification, the proposed pipeline demonstrated great potential to accelerate the multiplex imaging analysis, which would promote the biomarker discovery and clinical applications.

Introduction

The abundance and spatial distribution of cell subsets are crucial to our understanding of disease progression and response to therapies[1]. Rapid development of multiplex imaging techniques such as multiplex immunohistochemistry (mIHC) and imaging mass cytometry (IMC) has enabled the accurate quantitative localization of cellular markers in situ[2]. However, co-expression of antigens and the coexistence of abundant and rare cell types impose unique challenges for automated cell phenotyping in these images[3].

The field of multiplex image analysis is currently dominated by supervised learning[4]. Model training is often required for each panel with specific marker colours and cellular

locations, resulting in dramatically increased annotation burden as the number of panels increases. Typically, fully supervised methods require >1000 annotations per cell type per panel[3,5], summed to >10 hours of work from a pathologist to annotate for a panel with 3 markers. Also, existing methods could be sensitive to the class imbalance issue often observed in multiplex images[3]. Unsupervised methods which mainly rely on colour decomposition, are often limited to 4-6 colour channels[6], and can be prone to background staining noise[7].

To leverage the latest advantages of deep learning and to minimize the annotation burden, we proposed to apply self-supervised deep learning-based approach that utilizes intrinsic features from unlabelled data to facilitate cell phenotyping. Unlike supervised models trained using manual labels, self-supervised learning models can learn the inherent similarities of unlabelled data without pre-existing knowledge (Fig. 1a). Self-supervised learning has shown great promise in the classification of natural scene images[8,9], haematoxylin and eosin histology images[10,11], and microscope cell image data[12,13]. Additionally, previous applications of self-supervised learning on immuno-stained tissue sections either aimed at estimating cell type compositions in a region[14], or revealing cell types associated with patient-level clinical characteristics[15]. So far, dedicated approaches have not yet been developed for classification of single cells on multiplex images with their unique experimental set up, often consisting of multiple panels and therefore resulting in particularly heavy annotation burden.
Mitigating the annotation bottleneck of cell classification using self-supervised learning can produce fast and precise mapping of cell phenotypes, thereby accelerating the biomarker discovery and the clinical translation of multiplex imaging.

Here we propose SANDI with a self-supervised learning framework leading to a significant reduction of pathologist time. By leveraging the intrinsic similarities in unlabelled cell images, SANDI was able to perform cell classification with a small reference set containing as few as 10 annotations per type, while achieving a comparable performance with that of the supervised model trained on thousands of cell annotations.

We validated the efficacy of SANDI by comparing its performance with the fully supervised model, and two state-of-the-art self-supervised frameworks, SimCLR[8] and

MoCo[9] across a range of annotation burdens. We also examined the performance of SANDI with automatically selected reference set, as an approach to further reduce the necessary annotations for desirable classification accuracy. We conducted the experiments on four mIHC datasets and one IMC dataset, consisting of slides from ovarian cancer[16], lung squamous cell carcinoma (LUSC), ductal carcinoma in situ (DCIS)[17], myeloma[18] and pancreas[19] (Table 1). We focused on the classification of immune cell types, whose distribution and abundance are known to have impact on the disease progression and prognosis of different cancer types, and are therefore being targeted by a majority of multiplex imaging studies.

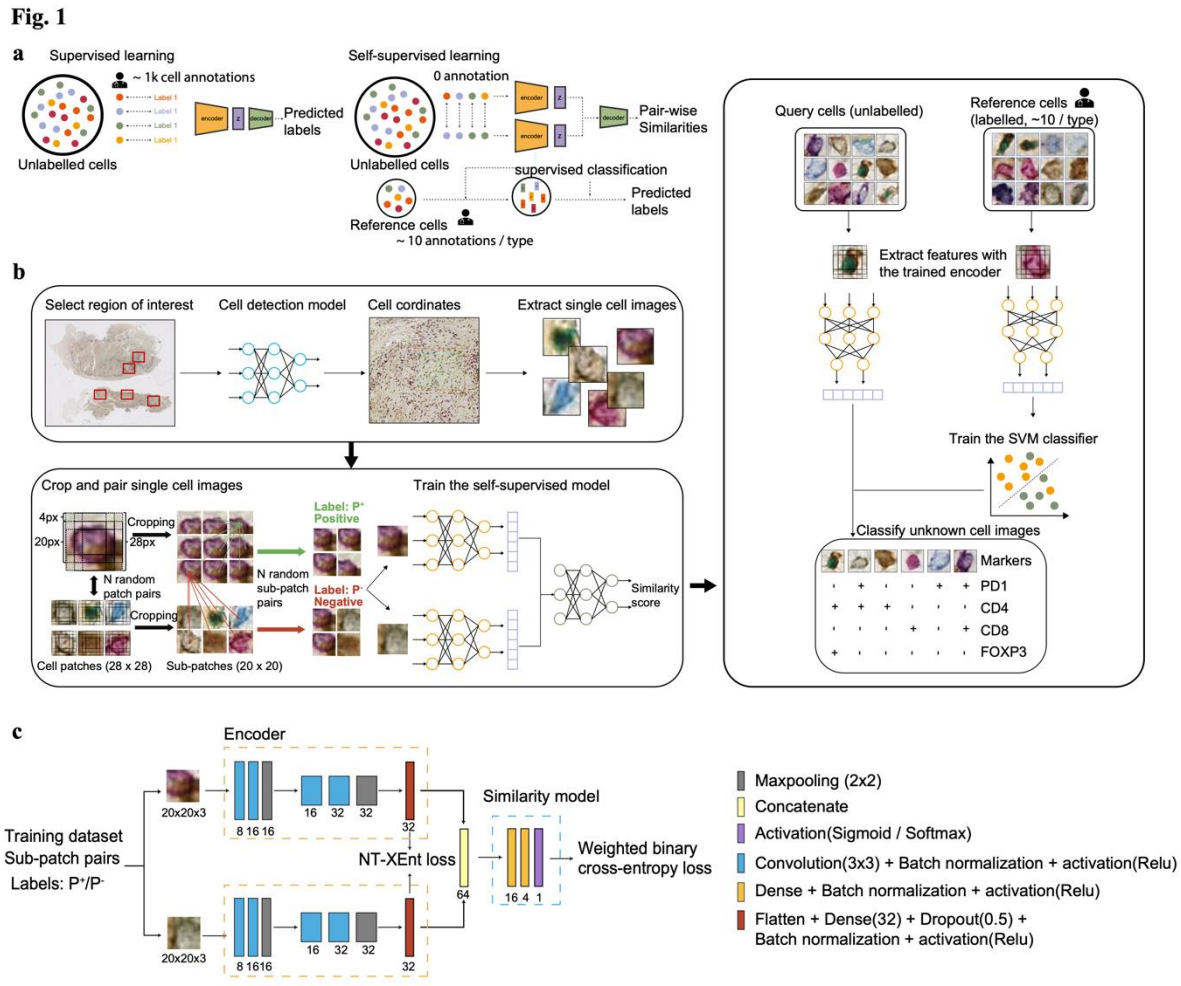

Figure. 1 Overview of the SANDI pipeline. a, Cartoons illustrating label-based and pairwise comparison-based training strategies of supervised and self-supervised learning. Supervised training is based upon a large number of manual labels, whereas self-supervised learning first infers distinct features of cell types by learning from the pairwise similarities, then classifies unlabeled cells using a small reference set. **b,** Schematic representation of the SANDI pipeline. In the data preparation process, we selected multiple regions on the WSI that contain a variety of cell types. Then a pre-trained cell detection model was applied to the selected regions to map the coordinates of cells. Single-cell patches of 28x28 pixels were retrieved to constitute the training dataset. The patches were then randomly paired and cropped into 20x20 pixel sub-patches. Subpatch pairs that originated from the same patch were labeled as positive (P^+), otherwise negative (P^-). Pairs of input sub-patches

were processed by two identical encoders to generate a feature vector of 32 dimensions. The encoded features were concatenated as inputs for the similarity model, which learnt to discriminate between P^+ and P^-. The output score represents the pairwise similarities between a pair of sub-patches. A small set of cells were labeled by the pathologists as references. Both the reference and the unknown cell image patches were cropped into 9 overlapping sub-patches of 20x20 pixels, which were then processed by the trained encoder to yield a feature vector of 9x32 dimensions. A support vector machine (SVM) classifier was trained on features extracted from the references and to classify features extracted from unknown cells. **c,** Architecture of the self-supervised model.

Material and methods

Datasets

For experiments conducted in the study, the model was trained and validated on four mIHC datasets and one IMC dataset, including 9 ovarian cancer slides stained with CD8/CD4/FOXP3/PD1, 4 LUSC slides with CD8/CD4/FOXP3/haematoxylin, 12 DCIS slides with FOXP3, 6 Myeloma slides with CD8/CD4/FOXP3, and 100 IMC slides with CD4/CD8 channels extracted. Details of the five datasets are summarized in Table 1. Slides were scanned at 40x magnification and were down-sampled to 20x before processing.

Table 1. Composition of the 5 datasets used in the study.

Dataset	**Contributors**	**cell phenotypes**	**No. of annotations**		**Total no. of annotations**	
			Training	Testing	Training	Testing
Ovarian T cells	Tami Grunewald et al.[16]	CD4+FOXP3+	292	197	1828 (4 slides)	997 (5 slides)
		CD4+FOXP3-	596	168		
		PD1+CD8+	726	347		
		PD1-CD8+	139	203		
		PD1+CD4+	39	60		
		PD1+CD8-CD4-	36	22		
LUSC T cells	Teresa Marafioti and John Le Quesne	CD4+FOXP3+	746	228	2407 (2 slides)	1383 (2 slides)
		CD4+FOXP3-	1225	696		
		CD8+	204	200		
		Haematoxylin-stained	232	259		
DCIS FOXP3	Hwang et al.[17]	FOXP3+	1030	576	11459 (7 slides)	3799 (5 slides)
		FOXP3-	10429	3223		
Myeloma	Yong et al.[18]	CD8+	866	979	3269 (4 slides)	1588 (2 slides)
		CD4+FOXP3-	2244	493		
		CD4+FOXP3+	159	116		
IMC CD4-CD8	Damond et al.[19]	CD4+	987	828	3954 (80 slides)	1085 (20 slides)
		CD8+	2967	257		

Overview of the SANDI pipeline

The SANDI pipeline incorporated key strategies tailored for digital pathology to: (1) rapidly generate abundant examples of each cell type in regions of interest selected by pathologists, which can be achieved in minutes; (2) perform a series of operations to assign cell pairs as similar or dissimilar, generate shift-invariant representation of the cells, and extract distinctive features from unlabelled cell images; (3) convert learnt features into cell phenotyping based on a small set of references using a Support Vector Machine (SVM) classifier (Fig.1b).

The self-supervised model of SANDI was built on a convolution neural network with two identical encoders[20] (Fig. 1c). The model was trained to discriminate between pairs of subpatches that originated from the same cell image (P^+), and different cell images (P^-) (Methods). Each subpatch was encoded into a vector of 32 features (Fig. 1c). The objective of the training step was to minimize the loss function as a combination of normalized temperature-scaled cross-entropy loss (NT-XEnt)[8] and the weighted cross-entropy loss (Methods). The loss function is designed to keep features derived from the same cell in close proximity, and features derived from different cells to be far away in the feature space.

The trained self-supervised model of SANDI was able to extract discriminating features for different cell phenotypes by learning to predict the pair-wise similarities (Fig. 1c). To convert the encoded features into cell identities, we collected a small set of representative cell images as references. The encoder of the trained self-supervised model was used to extract features from both subpatches of reference and unknown cells. A linear SVM trained on features of the references was used to classify unknown cells.

Single-cell patches sampling

All slides were analysed for single-cell detection using a pre-trained deep learning model[21] prior to the proposed pipeline. To build the dataset for self-learning purposes, the first step was typically to sample single-cell patches from the whole slide image (WSI)[22]. In an ideal situation where the percentage of each cell type present in the dataset is balanced, we can randomly sample from the pool of all detected cells and

expect an equal chance of capturing each cell type of interest. However, in pathological data, cell type imbalance is common, which might cause some rare cell types to be missed out by random sampling.

To tackle this problem and to investigate the impact of data imbalance on the model performance, we introduced a data sampling step to capture a variety of cell phenotypes and ensure the inclusion of rare cell types. First, small regions on the WSI enriched with diverse cell types were manually identified. Then, a pathologist will label the class of each cell within these regions by annotating the cell centre using different colours to denote different cell types. The selection of regions ensures that a considerable number of each cell type are included in the training dataset. Manual labels revealed the composition of cell types within the regions and provided ground truth for model evaluation. A 28x28 pixel patch around each dot annotation was retrieved to form the dataset. All patches from slides used for model training were pooled together and randomly allocated to training or validation set with a 4:1 ratio.

Patch cropping and pairing

Given a dataset containing *n* 28x28 pixel (12.32x12.32 μm^2) single-cell image patches $D_n = \{x_i, \ldots, x_n\}$, we first generated all possible combinations $C_2 = \{(x_i, x_j) \in D | i \neq j\}$. For each batch, *N* pairs(x_i, x_j) were randomly sampled from C_2 without replacement. For each pair of single-cell image patches, the acquired patches x_i, x_j were each randomly cropped into 20x20 pixel (8.8x8.8 μm^2) sub-patches x_{d_i, s_i}. Sub-patches retrieved from the same patch and the paired patch were labelled as positive (P^+) and negative (P^-) respectively, indicating that they were from the same cell or different cells. These are described as follows:

$$P^+ = \left\{\left(x_{d_i,s_i}, x_{d_j,s_j}\right) \in C_2 |\ d_i = d_j, s_i \neq s_j\right\} \quad (1)$$

$$P^- = \left\{\left(x_{d_i,s_i}, x_{d_j,s_j}\right) \in C_2 |\ d_i \neq d_j, s_i \neq s_j\right\} \quad (2)$$

The total number of $P^+$and P^- in a batch is *2N* with *N* set to 256 in the experiment. RGB-valued images were normalized to the range [0,1] before being fed into the network. The rationale behind comparing sub-patches randomly cropped from

single-cell images is to mimic the inspection by pathologists where a slight shift in the field of view does not affect the judgment of cell identities.

Network architecture and training

As shown in Fig.1c, the self-supervised network consists of two identical encoders conjoined at their last layers, followed by a single branch responsible for computing the pairwise similarity between the outputs of the two encoders. Each encoder contains a series of convolution, activation, batch normalization, max-pooling, and dropout layers, which encode the input image into a vector of 32 features. The single branch concatenates the outputs from two encoders and feeds them through a dense layer, followed by linear activation, batch normalization, Relu activation, and Sigmoid activation. The last layer generates a value between 0 and 1, which corresponds to the predicted similarity score between the image pairs. A higher score indicates more similarity between the two images.

For cell phenotyping purposes, the network was expected to generate a high score for cells from the same class and a low score for cells from distinct classes. However, since the network was trained to identify similar or dissimilar pairs randomly sampled from the unlabelled dataset, two images from the same class might have been labelled as negative during the data preparation, which biased the network towards features that discriminate against images from the same class. To reduce the impact of uncertainty in negative labels, we modified the binary-entropy loss function by applying lower weights to the P^- than to P^+.

$$L_{wbce} = -\frac{1}{N}\sum_{i=1}^{N}\left(w^+ log\left(f_s(P_i^+)\right) + w^- log\left(f_s(P_i^-)\right)\right) \quad (3)$$

Where f_s denotes the similarity branch, *N* is the total number of P^+ or P^- within a batch. w^+, w^-denote the pre-defined weights applied to the entropy loss of positive pairs P_i^+ and negative pairs P_i^-. In the experiment, $w^+$and w^-were set as 0.7 and 0.3 respectively.

To further constrain the latent representations to maximize the agreement between P^+, we combined L_{wbce} with the normalized temperature-scaled cross entropy loss (NT-XEnt)[8], which is expressed as

$$L_{NT-XEnt} = -log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} l_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \quad (4)$$

where z_i denotes the l_2 normalized embedding of sub-patch x_{d_i,s_i}, sim denotes cosine similarity, $l_{[k \neq i]}$ equals to 1 if $k \neq i$, otherwise 0. τ denotes the temperature parameter, which is set as 0.1 in the experiment. For a given sub-patch x_{d_i,s_i}, the NT-XEnt loss treats the sub-patch x_{d_i,s_j} originated from the same patch as positive samples, and all the other $(2N-2)$ sub-patches within the batch as negative samples.

The combined loss is the combination of L_{wbce} and $L_{NT-XEnt}$, as given by

$$L_{combined} = L_{wbce} + L_{NT-XEnt} \quad (5)$$

For rigorous assessment of models, all training was performed on an Intel i7-9750H CPU for 100 epochs with a batch size of 256 and optimized using Adam with a learning rate of 10^{-3}. The model with the least validation loss was selected for evaluation.

Reference-based cell classification

Identification of cells from multiplex images is dependent on stain concentrations and the morphology of the cell, which can be affected by experimental artifacts and out-of-focus regions. The noise in the data/label is a well-known issue affecting model performance in digital histology image analysis[23,24]. Motivated by the need to reduce the annotation burden, we selected a set of reference images $R_n = \{x_i, \ldots, x_n\}$ from the training dataset *D* as representations of each cell type. Each cell in a hold-out testing set is treated as a query image x_{q_i}. Both the reference image x_{r_i} and query image x_{q_i} were cropped into 9 20x20 pixel sub-patches and processed by the trained encoder to yield the latent embeddings $f(x_{ri,si})$ and $f(x_{qi,si})$ of size 32x9. Assembling features of sub-patches allows the local regions neighbouring to the cell to be incorporated for downstream classification, which has shown to generate more accurate predictions[21].

An SVM classifier with a linear kernel implemented in the libsvm library[25] was trained on feature embeddings of references $f(x_{ri,si})$ and predicted cell phenotypes for embeddings of unlabelled samples $f(x_{qi,si})$.

Automatic expansion of the reference set

Although SANDI can obtain high accuracy using a limited number of labels, being trained on a small set of representatives may lead to an underestimation of the intra-cell-type variations in stain intensities, colour combinations, and morphologies[3,26,27]. By contrast, a larger training set can expose the model to higher variability in the data but can also deteriorate model performance if poor-quality data is included[26,28]. An ideal approach to capture a good level of variation while ensuring adequate data quality is to leverage information learnt by self-supervised training to inform the pathologist of cells that are prone to misclassification and thereby, create ground truth feedback to improve model performance. For this purpose, we proposed the automatic expansion method for iteratively adding human interpretation of the least confident instances as training events.

The flowchart illustrating the pipeline is shown in Fig. 2a. Firstly, we nominated 1 image for each cell phenotype as a representative, and then the minimal Euclidean distance $dist$ between embeddings of unlabelled images $f(x_{qi,si})$ and each reference image $f(x_{ri,si})$ was used to determine the cell type. This distance-based classification method is described by:

$$p\left(y|dist\left(f(x_{ri,si}), f(x_{qi,si})\right)\right) \quad (6)$$

Second, as an automated reference set expansion, for each group of cells as class *K*, the cell with the maximum Euclidean distance to any of the reference cells from the same class *K* was selected and manually labelled. These newly selected cells were then added to the previous reference set, while ignoring repeated instances. The two steps were repeated for 10 rounds and the weighted F1-score computed on the testing set was examined using the reference set from each round.

Assessment of model performance

To evaluate the model performance under various annotation budgets, we trained linear SVM classifiers on feature embeddings of randomly sampled training subsets containing 1%, 3%, 5%, 10%, 20%, and 30% of annotated samples of each cell type. The training of SVM was repeated five times on different randomly sampled training sets, and the mean weighted F1-scores were reported on hold-out testing sets containing cells from slides excluded from training. Results were compared against the performance of SVM trained features generated by two state-of-the-art self-supervised methods SimCLR[8] and MoCo[9], and a supervised classifier trained on 10%, 20%, 30%, and 100% of the annotations.

For fair comparisons, the supervised classifier, SimCLR, and MoCo were constructed with the same encoder as SANDI, and only random flipping was applied for data augmentation. All methods were trained on the same training/validation set split, and were tested on the same hold-out testing set as SANDI.

Performance of the model was evaluated using the weighted F1-score, which is the average of F1-score for each class weighted by the number of their instances:

$$weightedF1 = \frac{1}{n}\sum_{i=1}^{k} n_i * \frac{2TP}{2TP+FP+FN} \quad (7)$$

Where n is the total number of instances, k is the number of classes, and n_i is the number of instances for class i.

Results

Evaluation of SANDI for cell classification across various annotation burdens

The effectiveness of SANDI in discriminating diverse cell types was first evaluated by visualizing the embeddings of testing images in the latent space, which was performed using the t-distributed stochastic neighbour embedding (t-SNE). To capture the variability in cell appearance, each testing image was represented by the embeddings of nine sub-patches in its neighbourhood. The t-SNE plot revealed compact and

distinguishable clusters corresponding to each cell type, suggesting that the model has captured discriminative features for different cell classes (Fig. 2a).

To investigate the size of reference set required for SANDI to achieve reasonable performance, we first trained linear SVM on feature embeddings of randomly sampled reference sets containing 1%, 3%, 5%, 10%, 20%, and 30% of annotated samples of each cell type. When the budget was limited to 1%, the number of annotations ranged from 1 for PD1+CD8-CD4- cells in the ovarian T dataset, to 104 for the FOXP3- cells in the DCIS FOXP3 dataset (Table 1).

Across 5 datasets, SANDI achieved an impressive performance using only 1% of annotations (18 - 114 cells, Fig. 2b), comparable to a supervised classifier constructed using the same encoder trained on 1,828 - 11,459 annotations (-0.002 - -0.053 of weighted F1-score, Table 2). With a budget of below 30% of annotation data (11-3129 cells per type), SANDI outperformed the supervised classifier, and the other two state of the art self-supervised frameworks SimCLR[8] and MoCo[9] in all the five datasets (Fig. 2b, Table 2). Even when the size of the reference set was limited to 1%, 3%, and 5% of annotations, SANDI still achieved a higher or comparable (+- 0.05) weighted F1-score than the supervised classifier trained on 10% of annotations (Fig. 2b, Table 2). Thus, SANDI can obtain an adequate classification accuracy using 100 times fewer annotations than the conventional supervised training methods. Importantly, the superiority of SANDI in the ovarian T cells, lung squamous cell carcinoma (LUSC) T cells and myeloma datasets with substantial data imbalance suggests a key advantage of SANDI in multiplex image analysis. Furthermore, we observed that the weighted cross-entropy loss improved over the non-weighted version; and when combined with the contrastive loss NT-XEnt to learn co-occurring modalities, resulted in best overall performance regardless of image types (Table S1). Thus, SANDI is capable of boosting the performance of unbiased cell identification regardless of cell abundance, possibly because of its loss function design and independence of prior-defined labels[29].

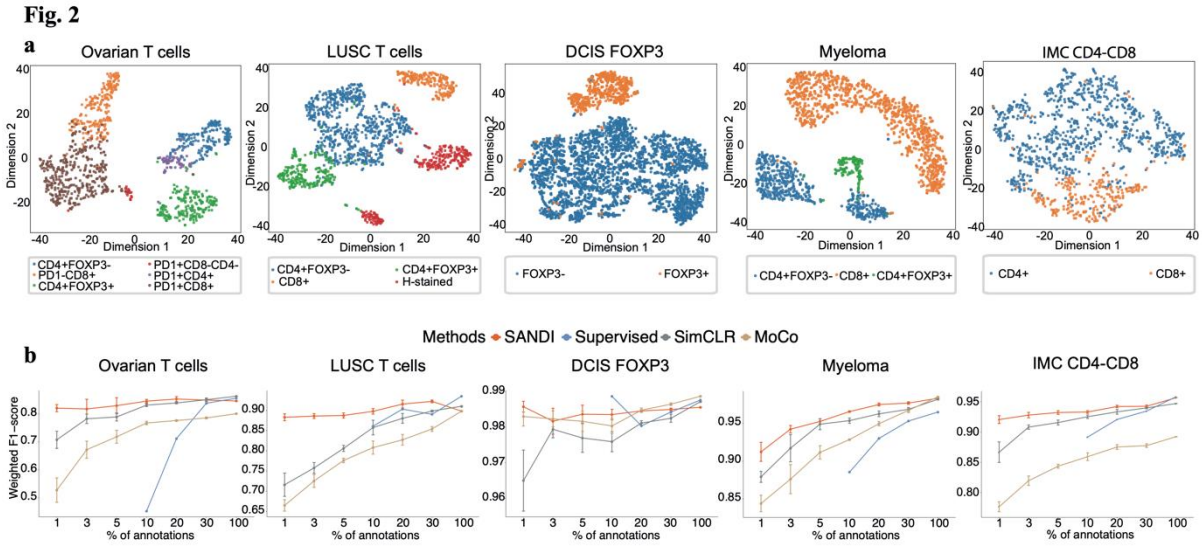

Fig. 2 Performance of SANDI on five datasets. a, The t-SNE representation of test image embeddings. Cell labels are represented as color codes. **b,** Comparison of the performance based on weighted F1-score of three self-supervised methods (SANDI, SimCLR and MoCo) and the supervised classifier over increasing amounts of training data. For SANDI, SimCLR and MoCo, the mean and standard error from five random sampling were shown.

Table 2. The weighted F1-score of the SVM classifier trained on features generated by different methods and with various percentages of annotations. All results are the average over 5 trials with different random seeds.

Annotations	1%	3%	5%	10%	20%	30%	100%
Ovarian T cells							
No. of cells	18	54	91	182	365	548	1828
Supervised classifier	-	-	-	0.452	0.711	0.838	0.856
SimCLR	0.707	0.782	0.789	0.831	0.839	**0.850**	**0.863**
MoCo	0.527	0.671	0.718	0.767	0.776	0.785	0.800
SANDI	**0.820**	**0.817**	**0.829**	**0.845**	**0.853**	**0.849**	0.846
LUSC T cells							
No. of cells	24	72	120	240	481	722	2407
Supervised classifier	-	-	-	0.861	0.903	0.890	0.935
SimCLR	0.716	0.757	0.806	0.857	0.880	0.898	0.910
MoCo	0.664	0.725	0.776	0.808	0.827	0.854	0.898
SANDI	**0.883**	**0.886**	**0.887**	**0.898**	**0.916**	**0.922**	**0.934**
DCIS FOXP3							
No. of cells	114	343	572	1145	2291	3437	11459
Supervised classifier	-	-	-	0.989	0.980	0.984	0.988
SimCLR	0.965	0.979	0.977	0.976	0.981	0.982	0.987
MoCo	0.983	0.982	0.982	0.980	0.985	**0.986**	**0.989**

SANDI	**0.986**	**0.982**	**0.984**	**0.984**	**0.985**	**0.985**	0.986
Myeloma							
No. of cells	32	98	163	326	653	980	3269
Supervised classifier	-	-	-	0.885	0.930	0.953	0.965
SimCLR	0.879	0.917	0.949	0.953	0.962	0.968	0.982
MoCo	0.844	0.876	0.912	0.928	0.949	0.967	**0.985**
SANDI	**0.912**	**0.942**	**0.952**	**0.965**	**0.975**	**0.977**	0.983
IMC CD4-CD8							
No. of cells	39	118	197	395	790	1186	3954
Supervised classifier	-	-	-	0.892	0.921	0.935	**0.958**
SimCLR	0.867	0.908	0.916	0.925	0.933	0.940	0.947
MoCo	0.777	0.820	0.844	0.859	0.875	0.878	0.892
SANDI	**0.921**	**0.930**	**0.932**	**0.933**	**0.942**	**0.942**	**0.957**

Performance with the automatic expansion of the reference set

To effectively select reference images that contribute the most to model performance improvement, we designed an automatic expansion of the reference set. This is achieved by iteratively estimating the confidence of cell phenotyping performed by the trained model, and recommending the least confident instances for manual labelling (Fig. 3a, Methods). The reference set was initialized with one arbitrarily selected image for each cell type (Fig. 3b). With 10 iterations, we gathered a reference set containing 10 most diverse representations of the same cell type. Initial references and example cells classified at the 10th iteration were shown in Fig. S1. As expected, performance fluctuated at 1 random reference per cell type, but quickly gained stability (Fig. 3c), achieving higher weighted F1-scores than randomly sampled reference sets with about the same number of annotations (3% for Ovarian T cells and LUSC T cells, 1% for Myeloma and IMC CD4-CD8 datasets, Table 2, 3). These results suggest that the confidence-based reference selection scheme can effectively boost classification accuracy using as few as 10 annotations per cell type.

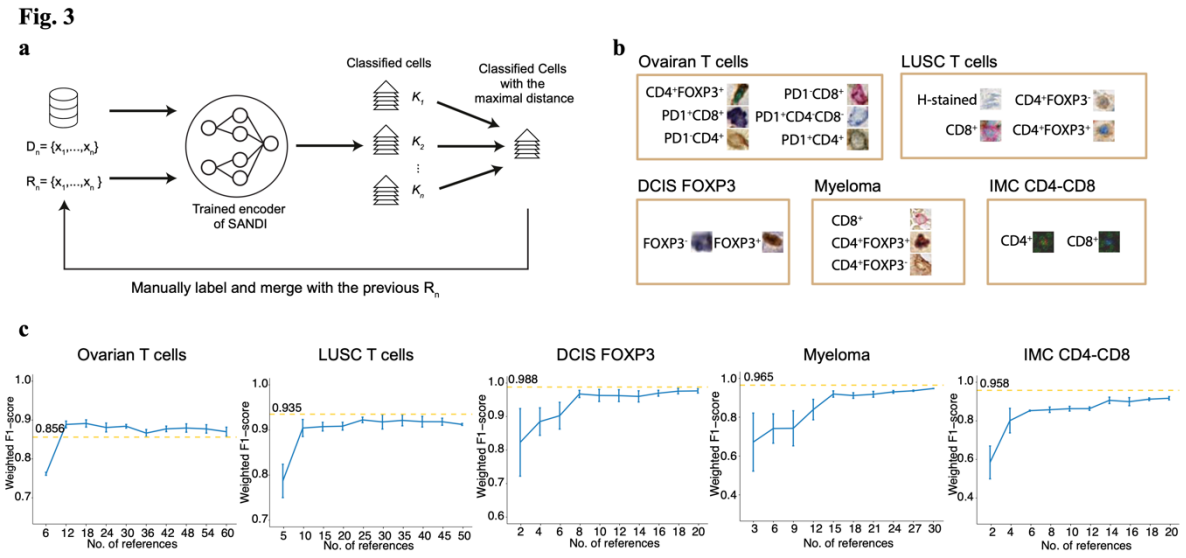

Fig. 3 Automatic expansion of reference sets. a, The automatic expansion scheme of reference sets to effectively select reference images that contribute most to the improvement of model performance. The unlabeled cell images from the training set *D* and an initial reference set R_n containing 1 reference image for each cell type *K* were provided in the first round. Images in *D* and R_n were cropped to 9 20x20 pixel sub-patches and processed by the trained feature encoder. The unlabeled cells were assigned with cell type *K* based on the Euclidean distance between embeddings of references and unlabeled cells. The instance with the maximal Euclidean distance was selected for manual labeling and merged with R_n from the previous round to form the new reference set. In the experiment, the process was repeated 10 times. **b,** Examples of initial reference sets for each of the five datasets. **c,** Weighted F1-score on testing sets for the linear SVM classifier trained on the reference set generated from each round of automatic expansion. The process was repeated three times with different initial reference images. The error bar indicates the standard error. The yellow horizontal line denotes the weighted F1 score achieved by the supervised classifier trained on 100% annotations.

Table 3. Weighted F1-score on testing set for the linear SVM classifier trained on reference set generated from each round of automatic expansion of reference set. Bold values are within 0.005 below the best.

Datasets	Rounds	1	2	3	4	5	6	7	8	9	10
Ovarian T cells	Ref. Size	6	12	18	24	30	36	42	48	54	60
	Weighted F1-score	0.758	**0.903**	0.878	0.866	0.878	0.873	0.889	0.894	0.891	0.892
LUSC T cells	Ref. Size	4	8	12	16	20	24	28	32	36	40
	Weighted F1-score	0.852	0.934	0.919	0.946	**0.952**	**0.948**	**0.953**	0.944	0.939	0.924
DCIS FOXP3	Ref. Size	2	4	6	8	10	12	14	16	18	20
	Weighted F1-score	0.771	**0.986**	**0.982**	0.976	0.973	0.956	0.960	0.955	0.957	0.970
Myeloma	Ref. Size	3	6	9	12	15	18	21	24	27	30
	Weighted F1-score	0.840	0.789	0.831	0.830	0.906	0.900	0.924	0.934	0.942	**0.948**
IMC CD4-CD8	Ref. Size	2	4	6	8	10	12	14	16	18	20
	Weighted F1-score	0.441	0.652	0.850	0.830	0.878	0.871	0.927	0.922	**0.924**	**0.929**

SANDI reveals association between PD1 expression and T regulatory cell proximity in the Ovarian T cells dataset

To examine the capability of SANDI in identifying biological meaningful cellular distributions, we performed it on cells auto-detected by a pre-trained neural network[5] on 9 slides from the Ovarian T cells dataset. It is worth noting that the auto-detected dataset contains tissue backgrounds that were over-detected as cells (Fig. S2). Despite such noise within the data, SANDI trained on 4431 auto-detected cells from 19 regions achieved a weighted F1-score of 0.855 with the linear SVM classifier trained on 20% of randomly selected training samples and evaluated on the same testing set as previously described, suggesting its robustness against incorrect detection of cells. Additionally, SANDI is capable of correcting over-detected cells using patches of tissue background as references (Fig. S2).

We applied SNADI to classify the six immune cell subsets using the 10th iteration of the automatic expansion scheme. The classified cells exhibit a diverse composition across the 9 samples (Fig. 4a), with PD1-CD4+FOXP3-, PD1+CD8+ and PD1-CD8+ being the top three abundant cell types (Fig. 4b). We observed negative associations between percentages of PD1-CD4+FOXP3- T helper cells (Th) and PD1-CD8+ cells (Rho = -0.922, p = 0.0004), PD1-CD4+FOXP3+ T regulatory cells (Treg) and PD1+CD8-CD4- cells (Rho = -0.720, p = 0.029), and between PD1+CD8+ cells and PD1+CD8-CD4- cells (Rho = -0.759, p = 0.018, Fig. 4c). PD1 expression has been associated with activation and exhaustion of CD8+ and CD4+ T cells[30]. To quantify the impact of PD1 expression on the T regulatory cell (Treg) mediated immunosuppression, we measured the distance of PD1+ and PD1- T cells to the nearest PD1-CD4+FOXP3+ Treg cell. We constrained the analysis to distance within 250um, which was shown to be the maximal distance of effective cell-cell interactions[31]. This approach showed that PD1-CD4+FOXP3- T helper cells (Th) were nearer to Treg cells than PD1+CD4+FOXP3- Th cells (Fig. 4d), whereas PD1+CD8+ cells were closer to Treg cells compared to PD1-CD8+ cells (Fig. 4e). It has been documented that CD4+ cells with low PD1 expression displayed reduced cytokine production and was associated with poor overall survival in follicular lymphoma[32]. By contrast, high expression of PD1 is known to charaterise the dysfunctional CD8+ T cells[30], and the irreversible exhaustion is partly attributed to the Treg interaction[33].

These findings raised the possibility that high PD1 expression on CD8+ T cytotoxic cells may be linked to increased interaction with Treg cells and co-orchestrate immunosuppression, while having an opposite effect on Th cells. Overall, these results demonstrated the potential of SANDI not only to classify cellular components but also to facilitate hypothesis-driven analysis of cell-cell interactions in complex tissues.

Figure 4.

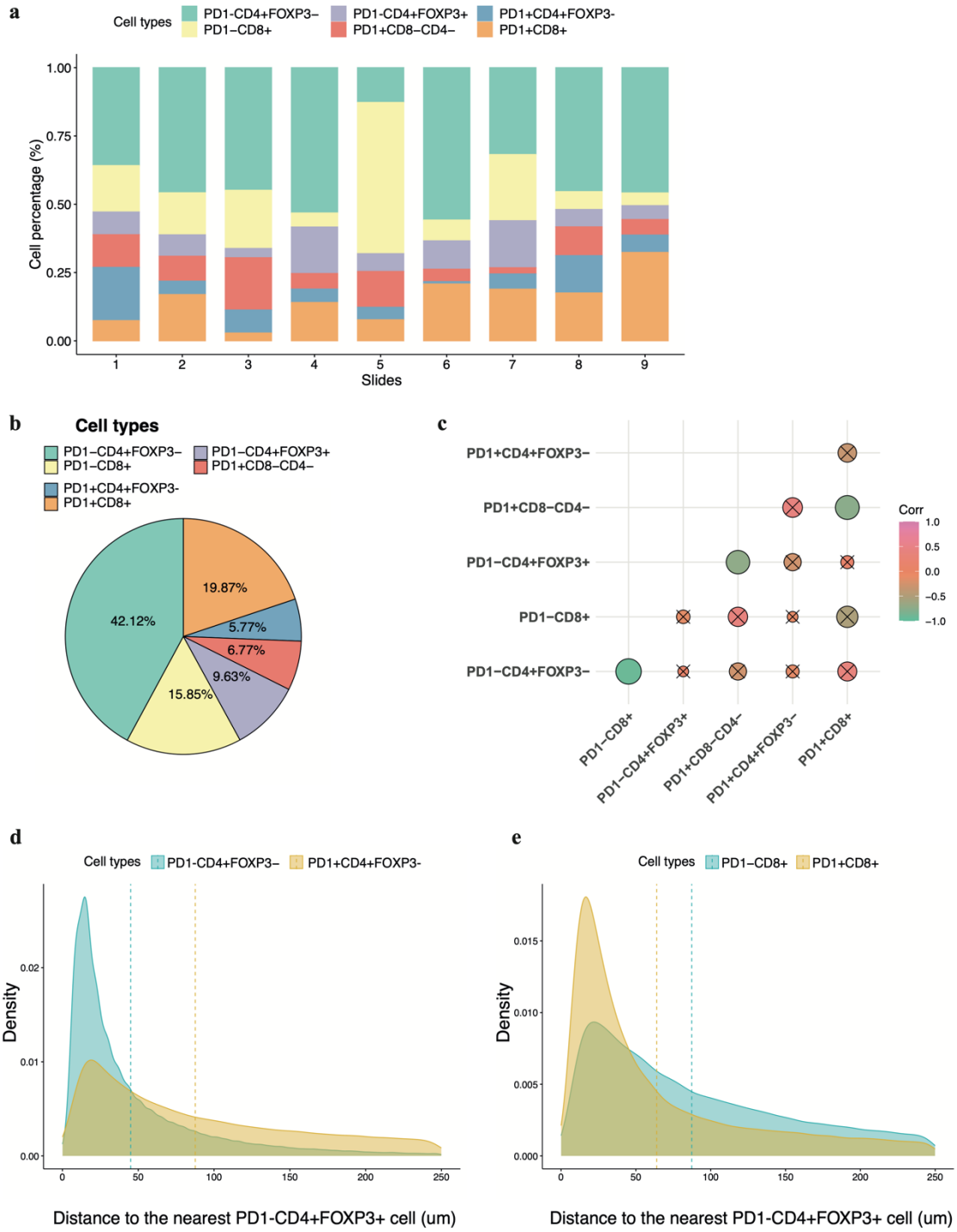

Fig. 4 Cellular composition and cell-cell distance in the Ovarian T cells dataset revealed by SANDI. a. The percentage of immune cell subsets in each of the 9 ovarian slides. **b.** Overall compositions of six immune cell subsets. **c,** Correlation heatmap to illustrate the association between percentages of six immune cell subsets. **d.** Density plots showing the distribution of PD1-CD4+FOXP3-, PD1+CD4+FOXP3-, PD1-CD8+, and PD1+CD8+ T cells within 250um to the nearest PD1-CD4+FOXP3+ Treg cells. The mean distance is shown as the horizontal line.

Discussion

In this work, we developed and demonstrated the performance of a self-supervised framework SANDI for cell classification on multiplex images to minimize the workload of pathologists. The results obtained in datasets acquired from 5 different sites show that, with an average of 10 labels per cell type, the performance of SANDI was comparable to that of the fully supervised classifier trained on more than 1800 single cell annotations. Specifically, SANDI achieved a weighted F1-score -0.002 below that of the fully supervised classifier in the Ovarian T dataset. We also showed that SANDI outperformed other self-supervised frameworks when the annotation budget was below 10%, indicating that our proposed framework is highly effective at reducing the number of annotations required for accurate classification. We achieved these results by using a self-supervised model that learns the distinct features of cell classes using pairwise similarities between subpatches of the same cell and different cells as labels. Additionally, we showed that the trained encoders can help identify cells that are prone to misclassification, thus guiding the annotation efforts towards cells that can effectively improve classification accuracy. With SANDI applied to the Ovarian T cell slides, we revealed a distinct association of PD1 expression on CD8+ and CD4+ cells with the Treg mediated immunosuppression. These results demonstrate the capability of SANDI in deconstructing cellular spatial organisation at scal and suggest its potential application in biomarker discovery and clinical translations.

This work has several limitations. First, the pipeline still requires manual selection of regions that contain a variety of cell phenotypes to ensure that a considerable number of cells of interest are included in the training. Future work to evaluate automated methods to guide region selection will help address this issue. Second, the training images of the self-supervised model is currently limited to cell-containing images, which involves a pre-trained detecting model applied prior to the pipeline to locate image patches of single cells. Classification on automatically detected cells showed that SANDI was capable of distinguishing cell-containing images from the tissue background when representative images of background were provided as references. It would be of interest to identify background patches using the self-supervised model

trained on randomly cropped image patches to reduce false positives in the cell detection. Additionally, the increase in classification performance as the automatic expansion of the reference set was inconsistent. Therefore, a labelled validation set is required to determine the optimal iteration and corresponding reference set that boosts the classification accuracy. Lastly, future work should tailor this method to other multiplex imaging techniques, such as phenocycler[34] and multiplexed ion beam imaging[35] to facilitate the cell phenotyping in the context of a large number of antibodies.

In conclusion, SANDI enables cost-effective cell phenotyping in multiplex images with minimal manual inputs, which facilitates the analysis of large-scale datasets and paves the way for translating multiplex image analysis into clinical practice. By employing the prediction of pairwise similarity as the pretext task, self-supervised learning leverages intrinsic information from the rich amount of unlabelled data independent of prior knowledge of cell phenotypes. This strategy greatly reduces the expert annotations required for desired classification performance, and establishes self-supervised learning as a promising new technology in medical artificial intelligence.

Data availability

The images and annotations of the DCIS FOXP3, Myeloma, and the IMC CD4-CD8 datasets can be obtained from the corresponding publication. Raw data of the Ovarian T cells and the LUSC T cells datasets are available upon reasonable request.

Code availability

The scripts for implementing and validating the pipeline are available at https://github.com/yuerua/SANDI.

Supplementary

Table S1. Weighted F1-score of cell classification using SANDI trained with different loss functions.

Figure S1. Initial references and example cell images classified using the automatic expansion scheme.

Figure S2. Example region from the ovarian T cell dataset containing auto-detected cells.

Acknowledgement

Y.Y acknowledges funding from Cancer Research UK Career Establishment Award (CRUK C45982/A21808), CRUK Early Detection Program Award (C9203/A28770), CRUK Sarcoma Accelerator (C56167/A29363), CRUK Brain Tumor Award (C25858/A28592), Breast Cancer Now (2015NovPR638), Rosetrees Trust (A2714), Children's Cancer and Leukaemia Group (CCLGA201906), NIH U54 CA217376, NIH R01 CA185138, CDMRP Breast Cancer Research Program Award BC132057, European Commission ITN (H2020-MSCA-ITN-2019), and The Royal Marsden/ICR National Institute of Health Research Biomedical Research Centre. K.W, L.L and MR-J are partly funded by the "UCL/UCLH NIHR Biomedical Research Centre.

Contributions

H.Z and Y.Y conceived and designed the study. Y.Y supervised the work. H.Z conducted the validation experiments and analysed the results. K.A provided insights for the pipeline. T.G and J.L provided the Ovarian T cells dataset. J.L.Q, T.M, and A.U.A provided the LUSC T cell dataset. F.S and S.H provided the DCIS FOXP3 dataset. Y.H, C.S.Y.L, D.P, L.L, MR-J and K.Y provided the Myeloma dataset. H.Z, K.A, Y.Y wrote the manuscript with input from all the authors.

Competing interests

The authors declare the following competing financial interests: a patent has been filed for the methodology reported in the paper (applicant, the Institute of Cancer Research; inventors, Hanyun Zhang and Yinyin Yuan; application number, UK patent GB 2106397.9 and PCT/EP2022/061941). Y.Y. has received speakers bureau honoraria

from Roche and consulted for Merck and Co Inc. The funders had no role in the design of the study; the collection, analysis, or interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication. The authors declare no competing non-financial interests that may have influenced the publication process.

References

1. Binnewies M, Roberts EW, Kersten K, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature Medicine*. 2018;24(5):541-550. doi:10.1038/s41591-018-0014-x

2. Tan WCC, Nerurkar SN, Cai HY, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications*. 2020;40(4):135-153. doi:10.1002/cac2.12023

3. Fassler DJ, Abousamra S, Gupta R, et al. Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images. 2020;15(1). Accessed October 15, 2020. https://diagnosticpathology.biomedcentral.com/articles/10.1186/s13000-020-01003-0

4. Serag A, Ion-Margineanu A, Qureshi H, et al. Translational AI and Deep Learning in Diagnostic Pathology. *Frontiers in Medicine*. 2019;6. Accessed September 23, 2022. https://www.frontiersin.org/articles/10.3389/fmed.2019.00185

5. AbdulJabbar K, Raza SEA, Rosenthal R, et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nature Medicine*. Published online May 27, 2020:1-9. doi:10.1038/s41591-020-0900-x

6. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*. 2017;7(1). doi:10.1038/s41598-017-17204-5

7. Geread RS, Morreale P, Dony RD, et al. IHC Color Histograms for Unsupervised Ki67 Proliferation Index Calculation. *Front Bioeng Biotechnol*. 2019;7:226. doi:10.3389/fbioe.2019.00226

8. Chen T, Kornblith S, Norouzi M, Hinton G. *A Simple Framework for Contrastive Learning of Visual Representations*.; 2020. Accessed November 17, 2020. https://github.com/google-research/simclr.

9. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum Contrast for Unsupervised Visual Representation Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Published online 2020:9726-9735. doi:10.1109/CVPR42600.2020.00975

10. Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, Rajpoot N. Self-Path: Self-Supervision for Classification of Pathology Images with Limited Annotations. *IEEE*

Transactions on Medical Imaging. 2021;40(10):2845-2856. doi:10.1109/TMI.2021.3056023

11. Ciga O, Xu T, Martel AL. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*. 2022;7:100198. doi:10.1016/J.MLWA.2021.100198

12. Kobayashi H, Cheveralls KC, Leonetti MD, Royer LA. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat Methods*. 2022;19(8):995-1003. doi:10.1038/s41592-022-01541-z

13. Wong KS, Zhong X, Low CSL, Kanchanawong P. Self-supervised classification of subcellular morphometric phenotypes reveals extracellular matrix-specific morphological responses. *Sci Rep*. 2022;12(1):15329. doi:10.1038/s41598-022-19472-2

14. Murphy M, Jegelka S, Fraenkel E. Self-supervised learning of cell type specificity from immunohistochemical images. *Bioinformatics*. 2022;38(Supplement_1):i395-i403. doi:10.1093/bioinformatics/btac263

15. Jiménez-Sánchez D, Ariz M, Chang H, Matias-Guiu X, de Andrea CE, Ortiz-de-Solórzano C. NaroNet: Discovery of tumor microenvironment elements from highly multiplexed images. *Medical Image Analysis*. 2022;78:102384. doi:10.1016/j.media.2022.102384

16. Zhang H, Grunewald T, Akarca AU, et al. Symmetric Dense Inception Network for Simultaneous Cell Detection and Classification in Multiplex Immunohistochemistry Images. *MICCAI Computational Pathlogy (COMPAY) Workshop*. 2021;156:246-257.

17. Sobhani F, Muralidhar S, Hamidinekoo A, et al. Spatial interplay of tissue hypoxia and T-cell regulation in ductal carcinoma in situ. *npj Breast Cancer*. 2022;8(1):1-11. doi:10.1038/s41523-022-00419-9

18. Hagos YB, Lecat CS, Patel D, et al. Cell abundance aware deep learning for cell detection on highly imbalanced pathological data. *Proceedings - International Symposium on Biomedical Imaging*. 2021;2021-April:1438-1442.

19. Damond N, Engler S, Zanotelli VRT, et al. A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metabolism*. 2019;29(3):755-768.e5. doi:10.1016/j.cmet.2018.11.014

20. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context Encoders: Feature Learning by Inpainting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2016;2016-Decem:2536-2544. doi:10.1109/CVPR.2016.278

21. Sirinukunwattana K, Raza SEA, Tsang YW, Snead DRJ, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Transactions on Medical Imaging*. 2016;35(5):1196-1206. doi:10.1109/TMI.2016.2525803

22. Falk T, Mai D, Bensch R, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods*. 2019;16(1):67-70. doi:10.1038/s41592-018-0261-2

23. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clinical Cancer Informatics*. Published online 2019. doi:10.1200/cci.18.00157

24. JM T, G A, M A, et al. The Society for Immunotherapy of Cancer statement on best practices for multiplex immunohistochemistry (IHC) and immunofluorescence (IF) staining and validation. *Journal for immunotherapy of cancer*. 2020;8(1). doi:10.1136/JITC-2019-000155

25. Chang CC, Lin CJ. LIBSVM: A Library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. 2011;2(3). doi:10.1145/1961189.1961199

26. Nalepa J, Kawulok M. Selecting training sets for support vector machines: a review. *Artificial Intelligence Review*. 2019;52(2):857-900. doi:10.1007/s10462-017-9611-1

27. Frénay B, Verleysen M. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*. 2014;25(5):845-869. doi:10.1109/TNNLS.2013.2292894

28. Tsyurmasto P, Zabarankin M, Uryasev S. Value-at-risk support vector machine: Stability to outliers. *Journal of Combinatorial Optimization*. 2014;28(1):218-232. doi:10.1007/s10878-013-9678-9

29. Liu H, HaoChen JZ, Gaidon A, Ma T. Self-supervised Learning is More Robust to Dataset Imbalance. In: ; 2021. Accessed September 29, 2022. https://openreview.net/forum?id=vUz4JPRLpGx

30. Hashimoto M, Kamphorst AO, Im SJ, et al. CD8 T Cell Exhaustion in Chronic Infection and Cancer: Opportunities for Interventions. *Annu Rev Med*. 2018;69:301-318. doi:10.1146/annurev-med-012017-043208

31. Francis K, Palsson BO. Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94(23):12258-12262. doi:10.1073/pnas.94.23.12258

32. Yang ZZ, Grote DM, Ziesmer SC, Xiu B, Novak AJ, Ansell SM. PD-1 expression defines two distinct T-cell sub-populations in follicular lymphoma that differentially impact patient survival. *Blood Cancer Journal*. 2015;5(2):e281-e281. doi:10.1038/bcj.2015.1

33. Ngiow SF, Young A, Jacquelot N, et al. A Threshold Level of Intratumor CD8+ T-cell PD1 Expression Dictates Therapeutic Response to Anti-PD1. *Cancer Research*. 2015;75(18):3800-3811. doi:10.1158/0008-5472.CAN-15-1082

34. Goltsev Y, Samusik N, Kennedy-Darling J, et al. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*. 2018;174(4):968-981.e15. doi:10.1016/J.CELL.2018.07.010

35. Keren L, Bosse M, Marquez D, et al. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*. 2018;174(6):1373-1387.e19.

doi:10.1016/J.CELL.2018.08.039/ATTACHMENT/DB711C24-528F-47F0-B379-F3DB037CA6BD/MMC3.XLSX