TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY-NC-ND

Citation (APA)
Spengler, D., Korovin, S., Prakash, K., Bankhead, P., Debets, R., Balcioglu, H. E., & Smith, C. (2026). CLEAR-IT, a framework for contrastive learning to capture the immune composition of tumor microenvironments. *Communications Biology*, *9*(1), Article 796. https://doi.org/10.1038/s42003-026-09984-2

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-026-09984-2

CLEAR-IT, a framework for contrastive learning to capture the immune composition of tumor microenvironments

Check for updates

Daniel Spengler[1], Serafim Korovin[1], Kirti Prakash[1,2], Peter Bankhead[3], Reno Debets[4], Hayri E. Balcioglu[4,5,6] ✉ & Carlas Smith[1,6] ✉

Accurate phenotyping of cells in the tumor microenvironment is essential for understanding cancer biology but typically requires precise cell segmentation, limiting scalability. Here, we introduce Contrastive Learning Enabled Accurate Registration of Immune and Tumor cells (CLEAR-IT), a self-supervised framework that learns cell-level features from multiplexed images using only cell locations. CLEAR-IT encoders achieve strong linear evaluation performance, improve substantially with hyperparameter optimization, and maintain high accuracy across imaging modalities and with up to 90% fewer labels. When substituted for handcrafted features in a state-of-the-art classifier, CLEAR-IT features yield higher performance, and their combination enables comparable accuracy with less than half of the labeled data otherwise required. The learned representations also support prognostic modeling: using annotations from a single patient, CLEAR-IT-based phenotyping identifies survival-associated tissue features that generalize across two cohorts and modalities. CLEAR-IT provides a segmentation-light, label-efficient approach for scalable cell phenotyping and enhances existing workflows in digital pathology and tumor microenvironment analysis.

Recent advances in multiplexed proteomics have underscored the critical role of tumor microenvironment (TME) spatial composition in cancer prognosis and therapy development[1–4]. As manual characterization of increasing amounts of patient imaging data becomes infeasible, machine learning algorithms for histopathology, and TME characterization in particular, are being developed[5–8]. Classifying individual cells in multiplexed images is essential across many areas of biomedical research and clinical diagnostics. However, current workflows remain labor-intensive and difficult to scale. A bottleneck is the need for accurate cell segmentation. This often involves multiple rounds of manual annotation and refinement to reach the quality required for clinical-grade results[9].

After segmentation, two main strategies are used for cell classification. One is an image-based approach, where deep learning models are trained on annotated and segmented cells to predict phenotypes. An example is CellSighter, which is a fully supervised, end-to-end approach that combines raw image data and segmentation masks to discriminate between different cell types[9]. The second approach, currently outperforming the former, is based on handcrafted features to measure cell expressions. MAPS (Machine learning for Analysis of Proteomics in Spatial biology) exemplifies this strategy by using extracted features such as protein expression and cell size from segmented cells to classify them[10].

While effective, these workflows raise three key challenges: first, whether handcrafted features capture all relevant biological information, second, how such time-intensive pipelines can be made more efficient, and third, the susceptibility to human bias from the manual efforts during cell feature extraction as well as during the fully supervised learning[11]. The overarching goal is to reduce manual effort while maintaining high predictive integrity.

Self-supervised learning (SSL), on the other hand, leverages unlabeled data to learn robust feature representations that maintain high performance with fewer labels. A widely adopted method is the Simple Framework for Contrastive Learning of Visual Representations (SimCLR) algorithm by Chen et al.[12], valued for its straightforward architecture and training procedure. By pre-training on unlabeled images, SimCLR can substantially reduce the need for large annotated datasets. Ciga et al.[13] were the first to apply SimCLR to brightfield histopathology images (primarily hematoxylin

[1]Delft Center for Systems and Control, Delft University of Technology, Delft, the Netherlands. [2]Centre for Genomic & Experimental Medicine, Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, UK. [3]Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, UK. [4]Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus Medical Center, Rotterdam, the Netherlands. [5]Department of Precision and Microsystems Engineering, Faculty of Mechanical Engineering, Delft University of Technology, Delft, the Netherlands. [6]These authors jointly supervised this work: Hayri E. Balcioglu and Carlas Smith. ✉e-mail: h.balcioglu@erasmusmc.nl; c.s.smith@tudelft.nl

and eosin stained) and demonstrated that modality-specific tuning of both the types and strengths of data augmentations was critical to obtain strong representations. In a similar vein, Jiménez-Sánchez et al.'s NaroNet framework was based on an exhaustive hyperparameter sweep while learning features from multiplex fluorescence images, again uncovering optimal parameters that diverged from the original SimCLR defaults[14]. Altogether, these examples demonstrate that SSL, using SimCLR as a prototypical framework, can unlock powerful feature learning from medical images with reduced annotation requirements, provided that critical hyperparameters (especially augmentation settings) are carefully tuned.

In this work, we explore the application of SSL to learn cell features from multiplexed imaging, which involves multiple biomarker channels per tissue section, and second, we avoid iterative segmentation by working directly with image patches centered on approximate cell locations, thus bypassing the need for precise segmentation. We evaluate our method on two distinct modalities: multiplexed immunofluorescence (MxIF) microscopy and multiplexed ion beam imaging (MIBI). Our aim is to develop a flexible, channel-agnostic workflow that minimizes the time required for development and annotation. By processing each channel independently and combining features at the classification stage, we aim to enable accurate, scalable downstream cell classification for imaging modalities with an arbitrary channel count and with minimal manual intervention. The resulting features can also be used together with features from other methods to further enhance classification performance.

Results

Here, we present Contrastive Learning Enabled Accurate Registration of Immune and Tumor cells (CLEAR-IT), a contrastive learning-based approach to extract features of cells from multiplexed tissue images that can be used to classify cell types with minimal annotation effort and maximal integrity (Fig. 1a–c). CLEAR-IT utilizes the SimCLR SSL algorithm for the encoder network[12], thereby decoupling feature extraction from downstream classification and enabling flexible selection of the classifier architecture.

To develop a flexible, channel-agnostic workflow that minimizes the time required for development and annotation, the following steps are taken (see Methods); (i) using only individual channels of a multiplex image and cell centroids as center coordinates, single-channel image patches centered around individual cells are extracted (Supplementary Fig. 1a left); (ii) an encoder is then pre-trained on those patches following the self-supervised (i.e. annotation-free) SimCLR algorithm (Supplementary Fig. 1b); (iii) after pre-training, the encoder outputs of each channel for a given patch are concatenated to obtain a multi-channel representation of an image patch (Supplementary Fig. 1a right, c); (iv) a classifier is then trained to predict cell phenotypes from the multi-channel representations given cell annotations as supervised learning signal (Supplementary Fig. 1c).

Here, steps i and ii produce the CLEAR-IT self-supervised encoder, which learns informative feature representations of cells from unlabeled image patches. Steps iii and iv correspond to a process commonly called linear evaluation: we keep the encoder's parameters fixed (a procedure known as "freezing") and train only a simple SLP (single-layer perceptron, also referred to as a linear classifier) on those fixed features to assign cell phenotype probabilities. This approach lets us directly measure how well the encoder's representations capture biologically relevant information without the results being influenced by the capacity of a more complex classifier[12]. As a metric for the linear evaluation, we compute the precision-recall area under the curve (PR-AUC) of the SLP's predictions (Methods).

In this work, we made use of data from two triple-negative breast cancer (TNBC) cohorts imaged with two different modalities: MxIF microscopy and MIBI. The multiplex immunofluorescence images (TNBC1-MxIF8) are 1010 8-channel images from 62 patients[1] and multiplexed ion beam time-of-flight images (MIBI-TOF) are 41 44-channel images from 41 patients (TNBC2-MIBI44)[4]. For direct comparison with the multiplexed immunofluorescence (TNBC1-MxIF8), we also reduced the MIBI images to 8-channel images (TNBC2-MIBI8) comparable to those of TNBC1-MxIF8 (Methods). For all classification experiments, we split the data by patients into approximately 75% training and 25% test sets (47/15 patients for TNBC1-MxIF8, 30/11 patients for TNBC2-MIBI8 and TNBC2-MIBI44) and further split the training sets into 80% training and 20% validation sets during training (Methods).

For the supervised training and testing of classifiers, we used multi-label (ML) annotations with 6 different class labels that were generated with TME-Analyzer[15] (TME-A_ML6) for both the TNBC1-MxIF8 and TNBC2-MIBI8 datasets (Methods). Second, we made use of previously published multi-class (MC) annotations: multiplexed immunofluorescence with inForm annotations (inForm_MC7 and inForm_ML6)[1] and multiplexed ion beam imaging with DeepCell annotations (DeepCell_MC17 and DeepCell_ML6)[4] (Methods, note the distinction between ML and MC annotations).

To evaluate CLEAR-IT across both modalities, we first carried out separate hyperparameter optimizations, including loss-function parameters and augmentation strategies, for the TNBC1-MxIF8 and TNBC2-MIBI8 datasets, in line with previous findings that SimCLR hyperparameters should be tailored to specific imaging modalities[13,14]. This optimization involved systematically applying combinations of pre-training loss-function parameters, augmentation schemes, and training data composition (Supplementary Fig. 1d), and assessing their impact via linear evaluation by performing cell phenotyping using a linear classifier (Supplementary Fig. 1c) trained and tested on the TME-Analyzer annotations (TME-A_ML6) available for both datasets. Final PR-AUC after optimization was 0.86 for TNBC1-MxIF8 and 0.95 for TNBC2-MIBI8. This corresponds to a 55.6% and 13.7% improvement, respectively, over the base SimCLR hyperparameter configuration. For TNBC1-MxIF8, the largest performance gain (43%) was achieved by optimizing the pre-training loss function, followed by a ~ 7% gain from improved augmentations and a small additional improvement (~ 2%) from increasing the amount of pre-training data. For TNBC2-MIBI8, optimizing the pre-training loss function alone accounted for nearly the entire 13% improvement, while augmentation and increased data only had a small effect (Fig. 1d and Supplementary Figs. 2–11).

To investigate the consistency of the features learned by the CLEAR-IT encoders, we conducted a cross-validation study across different sources of annotations beyond the ones generated by TME-Analyzer. For each of the two datasets, we performed linear evaluation using two independent sets of cell annotations: 1) multiplexed immunofluorescence with TME-Analyzer annotations (TNBC1-MxIF8 + TME-A_ML6), 2) multiplexed immunofluorescence with inForm annotations (TNBC-MxIF8 + inForm_ML6), 3) multiplexed ion beam imaging with TME annotations (TNBC2-MIBI8 + TME-A_ML6) and multiplexed ion beam imaging with DeepCell annotations (TNBC2-MIBI8 + DeepCell_ML6). Performance across datasets was highest when the unseen test annotations originated from the same annotation source (inForm, DeepCell or TME-Analyzer) as the training annotations, highlighting the impact of annotation consistency on downstream phenotyping accuracy (Supplementary Fig. 12a, b, main diagonals, see also Supplementary Fig. 13).

To investigate how many and which images should be annotated, we conducted an experiment where we progressively reduced the number of training images and evaluated the resulting classifier performance. Remarkably, we found that using only 10% of the training set, equivalent to just 0.1% of all available data, was sufficient to maintain high performance across most datasets. An exception was observed for TNBC2-MIBI8 + DeepCell_ML6, where performance dropped significantly when using less than 20% of the training data (1% of all available data) (Supplementary Fig. 14).

In practical settings, annotations are typically made at the patient level rather than on individual images. To simulate this scenario, we evaluated the impact of reducing the number of annotated patients included in the supervised training set. We found that including only 10% of the total number of patients (TNBC1-MxIF8: 6 patients; TNBC2-MIBI8: 4 patients) resulted in <10% drop in performance across all cases (Supplementary Fig. 15).

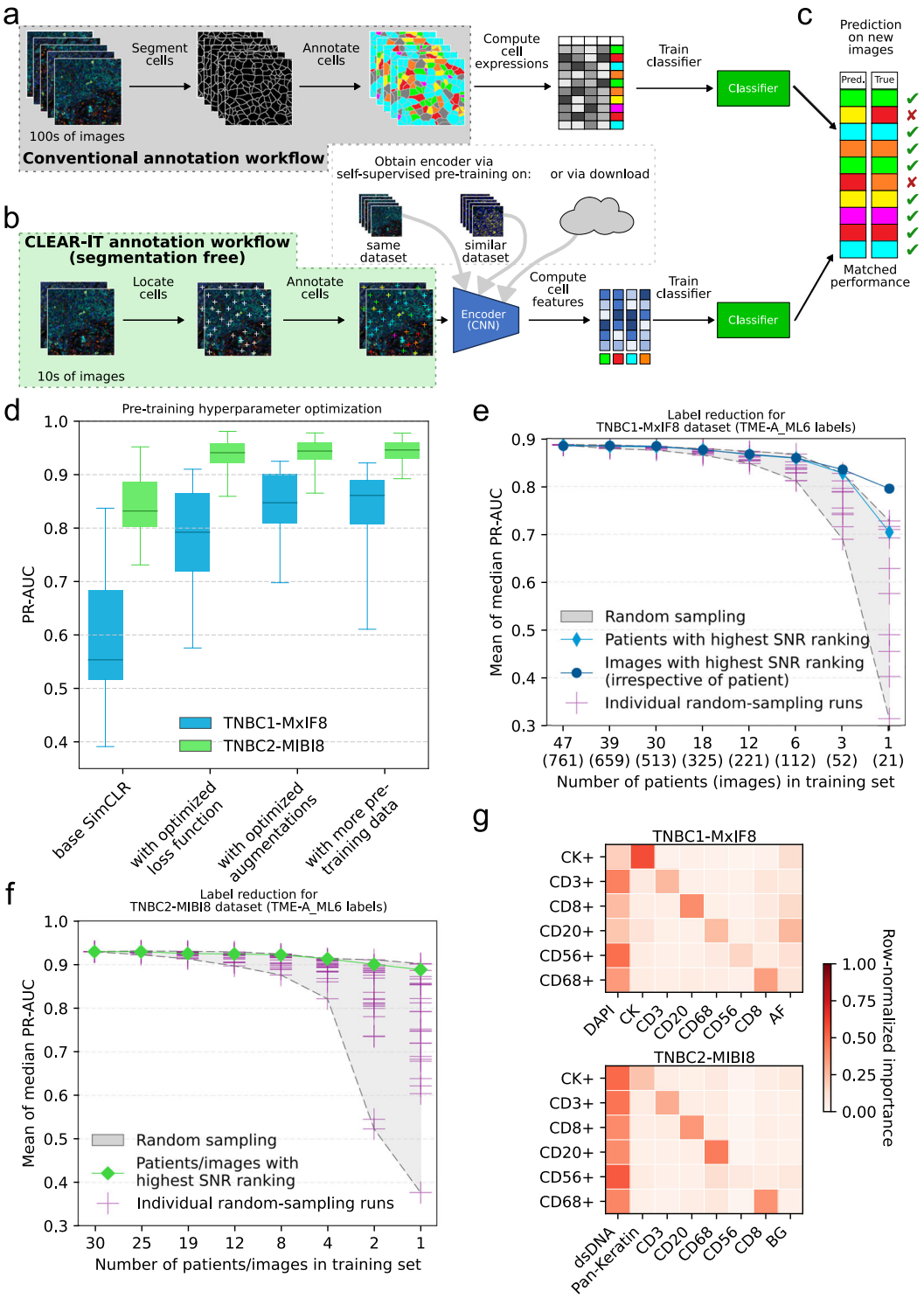


Fig. 1 | CLEAR-IT pipeline and performance. a Conventional training pipelines for cell phenotyping are trained on cell expressions, which require large amounts of segmentation ground truth. **b** Use of feature encodings generated by CLEAR-IT; convolutional neural network (CNN) encoder trained self-supervised without annotations; reduces hands-on-time needed by increasing label efficiency and using only cell locations. **c** Matched performance is achieved between benchmark and CLEAR-IT. **d** PR-AUC box-and-whisker plots showing improvement in linear evaluation performance for the TNBC1-MxIF8 and TNBC2-MIBI8 datasets through optimization of the pre-training hyperparameters for the CLEAR-IT encoder. The boxes represent the interquartile range (25th to 75th percentiles) and the whiskers extend to the 5th and 95th percentiles of the data, excluding outliers. Each box is based on $n = 60$ points (10 evaluation chunks and 6 class labels per chunk). **e**, **f** PR-AUC scores from linear evaluation of the best CLEAR-IT encoders pre-trained on TNBC1-MxIF8 and TNBC2-MIBI8 datasets. Performance is shown as a function of labeled data available during supervised training; reduction of the number of patients or images included in the training set. Shaded gray regions: performance range of $n = 10$ (**e**, TNBC1-MxIF8) and $n = 25$ (**f**, TNBC2-MIBI8) random permutations of patients or images (marked individually by purple + -signs). Diamonds: patients with the highest signal-to-noise ratio ranking. Dark blue circles: images with the highest signal-to-noise ratio irrespective of patient. **g** Heatmaps containing row-normalized (i.e., every row sums to 1) importance scores for the 100 most confident true positive (TP) predictions per class. Each cell of the matrix indicates how much impact a particular channel (column) has on the classifier's prediction of that row's class. AF autofluorescence, BG background.

We observed that data from some annotated patients yielded better classification performance than from others, prompting us to investigate whether this variation was due to differences in data quality rather than label quality. To guide annotation efforts, we analyzed which patients' data contributed most effectively to model performance and found considerable variation between individuals (Fig. 2, Supplementary Table 1, Supplementary Data 2). To assess whether these differences stemmed from data quality, we ranked the data based on signal-to-noise ratio (SNR) (Fig. 3a, Methods), which allowed us to evaluate the correlation between each patient's signal quality and the resulting classification performance (Fig. 3b, c). With this ranking, we observed Spearman correlation of $r = 0.51$ for TNBC1-MxIF8 and $r = 0.92$ for TNBC2-MIBI8, $p < 10^{-3}$ for both. A ranking based on more traditional metrics (mean of image with 99.5th percentile clipped, mean to median, Shannon entropy, see Supplementary Results, Supplementary Fig. 16) resulted in low correlation for TNBC1-MxIF8 ($r \leq 0.13$) and moderate to high correlation for TNBC2-MIBI8 ($r = -0.75$ for Shannon entropy). Notably, the performance with the best ranking patient was 18.6% and 4.6% lower than a full training cohort, for TNBC1-MxIF8 and TNBC2-MIBI8, respectively (Fig. 1e light blue diamonds, f).

For the TNBC1-MxIF8 dataset, we then investigated the performance of training on an equivalent amount of images but irrespective of which patients those images belong to (since there is only one image per patient for TNBC2-MIBI8, this case is identical to the one already investigated). This corresponded to 761 images (equivalent to 47 patients) for the full training set and 21 images (equivalent to 1 patient) for the most reduced training set. Training on the 21 highest-ranked images resulted in a mere 10.1% reduction in performance over training on the full dataset (Fig. 1e dark blue circles).

We found upon visual inspection that cells with a strong signal in the center were classified correctly with higher confidence (Fig. 4, Supplementary Figs. 17 and 18; Supplementary Methods). Motivated by the relationship between this patch-level confidence and image-level SNR, we further investigated how individual pixels across all channels influenced classifier decisions using Shapley Additive Explanations (SHAP)[16]. Here, our analysis was performed by considering the encoder's input. SHAP computes attribution scores for individual pixels, allowing us to visualize, which regions of each patch most strongly influence the encoder's feature outputs and subsequently the classifier's decision (Fig. 5, see also Supplementary Figs. 19–34, Supplementary Results, Methods). For a compact summary across classes, we aggregated absolute SHAP values over space and samples and row-normalized per class to obtain channel/class importance matrices for the 100 most confident true positive (TP) predictions in each dataset (see Methods; Fig. 5a, b for TNBC1-MxIF8 and TNBC2-MIBI8, respectively; corresponding importance heatmaps in Fig. 1g). The matrices show a clear diagonal pattern (class-matching marker channel) together with a consistently elevated nuclear channel importance (DAPI in MxIF, dsDNA in MIBI). For TNBC1-MxIF8 (Fig. 5a, see also Supplementary Figs. 19–26), the mean diagonal (class-matching channel) importance share was 0.19 (≈ 1.52 × a uniform 1/8 baseline), the nuclear share 0.32 (≈ 2.56×), and the combined diagonal + nucleus ("Top-2") captured 0.51 of the attribution mass. For TNBC2-MIBI8 (Fig. 5a, see also Supplementary Figs. 27–34), the mean diagonal importance share was 0.12 (≈ 0.96×), the nuclear share 0.45 (≈ 3.6×), and Top-2 reached 0.57. Together, these results indicate that CLEAR-IT's decisions are anchored in the expected marker and nuclear signals, with class-specific but smaller contributions from other channels.

Finally, we benchmarked the features learned by CLEAR-IT by comparing it as an input to a cell classifier against cell expressions and cell sizes, i.e., the set of "handcrafted features" derived from cell segmentations that serve as input to the state-of-the-art network MAPS (Machine learning for Analysis of Proteomics in Spatial biology)[10].

MAPS consists of a four-layer multi-layer perceptron (MLP), which predicts cell phenotypes. For our benchmark, we copied this architecture and trained MLPs with three different feature sets as input in the cell classification task while varying the amount of training data: (i) learned features from the pre-trained CLEAR-IT encoder, (ii) cell expressions and sizes (standard MAPS input), and (iii) a concatenation of both learned and handcrafted features. We performed the benchmark on four datasets: TNBC1-MxIF8 with inForm_MC7 annotations[8], TNBC2-MIBI44 with DeepCell_MC17 annotations[4] (Fig. 6), CRC-CODEX26 (colorectal cancer imaged via CODEX) with CellSighter_MC14 annotations[9,17], and TONSIL-IMC41 (tonsil tissue imaged via imaging mass cytometry, IMC) with OPTIMAL_MC21 annotations[18] (Supplementary Results, Supplementary Fig. 35, see also Supplementary Figs. 36–43 for detailed confusion matrices). Due to all annotations containing multi-class labels, we choose the micro average F1-score as a metric.

We observed that classification performance varied across datasets depending on the feature set used and the amount of training data (Table 1). For the cancer tissue datasets, the learned features from the pre-trained CLEAR-IT encoders outperformed cell expressions and sizes in the high-label regime, that is, when the highest amount of supervised training data was used for classifier training. Here, the largest difference of 9% for CLEAR-IT features was observed for CRC-CODEX26. For the non-cancerous TONSIL-IMC41, the difference was 11.2% for cell expressions and sizes.

Concatenating the learned CLEAR-IT features with cell expressions and sizes consistently improved performance in the high-label regime across all datasets, as summarized in Table 2, with improvements between 1.9% and 6.8% over the better-performing individual input. Specifically, for the cancer tissue datasets, combining CLEAR-IT features with cell expressions and sizes enabled a reduction to 3/47 patients (6.38%) for TNBC1-MxIF8, 19/30 images (63.33%) for TNBC2-MIBI44, 12/28 images (42.8%) for CRC-CODEX26, and 11/17 images (64.7%) for TONSIL-IMC41 in the supervised training sets, all while maintaining similar performance compared to the best individual feature set using fully supervised training.

To evaluate the clinical relevance of our approach, we investigated whether the cell phenotyping performed based on features from CLEAR-IT encoders in a minimal label scenario was sufficient to identify clinically relevant parameters. Using predictions made by linear classifiers trained on top of CLEAR-IT encoders with minimal labels for TNBC1-MxIF8 (Fig. 1f, 21 images) and TNBC2-MIBI8 (Fig. 1g, 1 patient) as input for our recently published methodology that extracts prognostic parameters[15], we were able to obtain survival classifiers that split the cohorts into two differential overall survival groups (Fig. 7a, b, top). We then cross-validated these classifiers (Fig. 7a, b, bottom) by predicting survival rates for patients from the respective other dataset (TNBC2-MIBI8 or TNBC1-MxIF8), which were considered to be the validation cohort. The predictions made by the TNBC1-MxIF8 minimal label linear classifiers mentioned above (Fig. 1f) were used to compute tumor and stroma area, as well as cellular density and distance TME parameters per image, which were then averaged across all images from each patient to get parameters per patient. This constituted the discovery cohort. Subsequently, using the TNBC1-MxIF8 survival data, we trained a survival classifier following our recently published methodology[15] to distinguish patients based on overall survival (Fig. 7a, top). For validation, we first obtained patient TME parameters in the TNBC2 dataset by using the cell classifier that was trained on the TNBC1-MxIF8 dataset (linear cell classifier trained using CLEAR-IT encoder pre-trained on TNBC1-MxIF8 dataset with annotations from 21 images) to predict cell classes for the TNBC2-MIBI8 features generated by a CLEAR-IT encoder pre-trained on TNBC2-MIBI8. Application of the survival classifier obtained for the TNBC1 cohort to these TNBC2 TME parameters successfully assigned patients to differential survival groups in the TNBC2 dataset (Fig. 7a, bottom), validating the prognostication pipeline. We then tested the alternative scenario, where self-supervised pre-training was performed on unlabeled data, and all supervised training was performed with TNBC2-MIBI8 labels, blind to the TNBC1-MxIF8 labels, and the supervised training of the cell classifier used only labels for cells from a single TNBC2-MIBI8 image (Fig. 1g, 1 patient). This approach again achieved successful assignment of patients into differential survival groups for both cohorts (Fig. 7b), demonstrating that cell phenotype predictions made based on CLEAR-IT

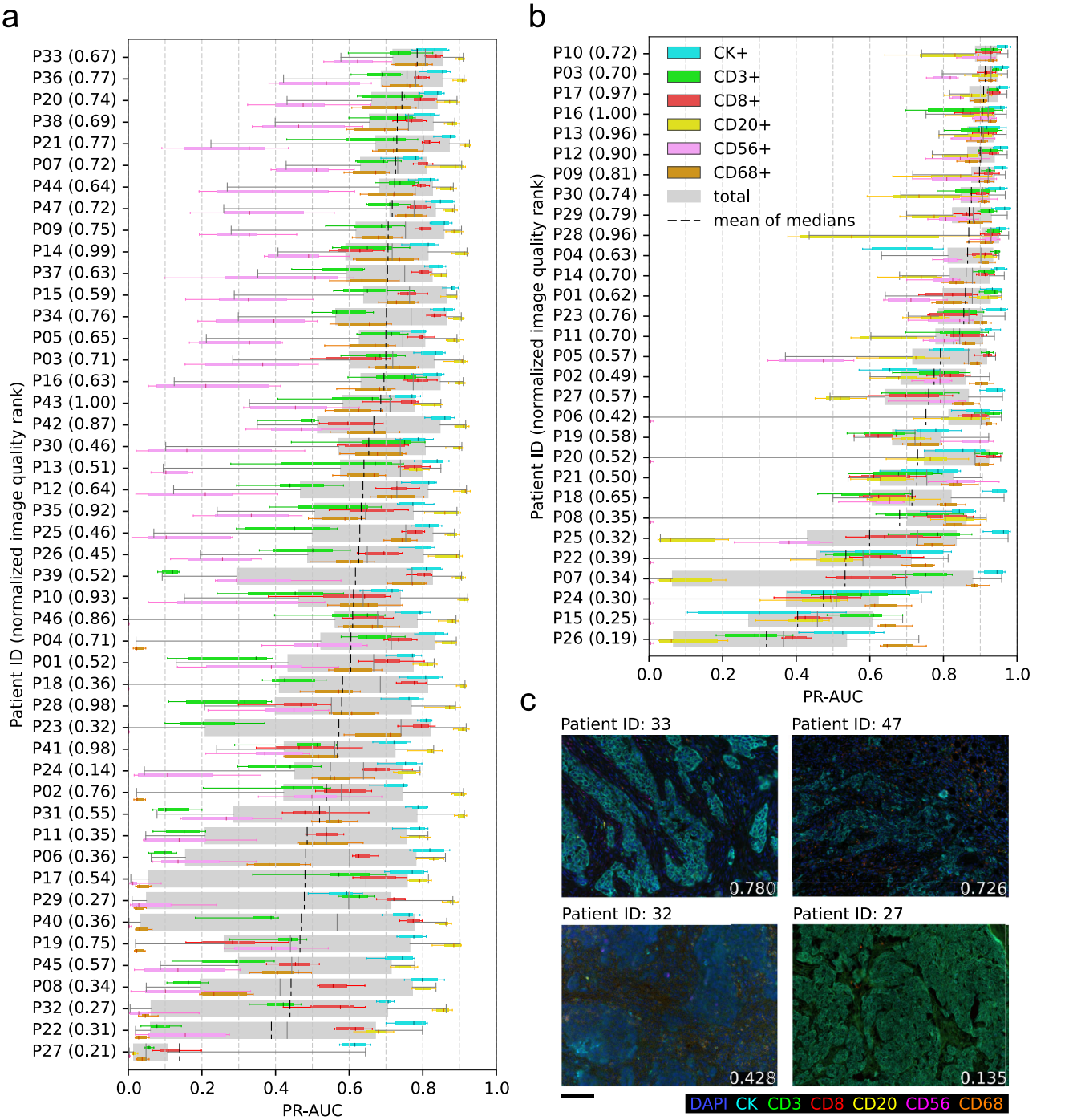


Fig. 2 | Performance of CLEAR-IT classifiers trained on data from single patients. **a**, **b** Overall performance of classifiers trained on data from one patient in the TNBC1-MxIF8 (**a**) and TNBC2-MIBI8 (**b**) datasets. The area under the precision-recall-curve (PR-AUC) of the encoder/linear classifier combination, plotted per cell label and for all cells per train/test set combination. The gray boxes represent the total classification performance irrespective of class labels. The dashed black lines represent the mean of the median PR-AUC scores per class, which are represented by the colored boxes. The y-axis specifies the patient ID and the normalized image-quality rank in parentheses. The boxes represent the interquartile range (25th to 75th percentiles) and the whiskers extend to the 5th and 95th percentiles of the data, excluding outliers. Each colored box is based on $n = 10$ points (10 evaluation chunks for one patient and one class), and each gray box on $n = 60$ points (10 chunks $\times$ 6 classes). Box colors: cyan; CK+, green; CD3+, red; CD8+, yellow; CD20+, magenta; CD56+, orange; CD68+. **c** Example tissue images from patients with best (top row) and worst (bottom row) generalization performance in TNBC1-MxIF8 cohort, when only a single patient's data is used for classifier training. Numbers in bottom right corners indicate the mean of median PR-AUC scores achieved when training classifiers on images of only that respective patient. Scale bar: 100 µm.

features by even very simple classifiers (i.e., a linear classifier), and even at highly reduced label presence, can be used to prognosticate TNBC patients. Furthermore, when investigating the parameters that are involved in survival classifiers, we observed that the parameter ranking generated from predictions made by CLEAR-IT-based classifiers were in good agreement with the published ranking based on predictions made by TME-Analyzer[15] (Supplementary Fig. 44a, b), demonstrating that CLEAR-IT encoders are able to extract features that correspond to clinically relevant TME biology, with minimal user input. More specifically, the parameters that rank highly with CLEAR-IT outputs (Supplementary Tables 2 and 3, Supplementary

a

Channel	CD3	CD8	CD20	CD56	CD68	background
STD of top ...% pixels ranked	0.1% descending	0.1% descending	0.1% descending	0.1% descending	0.1% descending	100% ascending

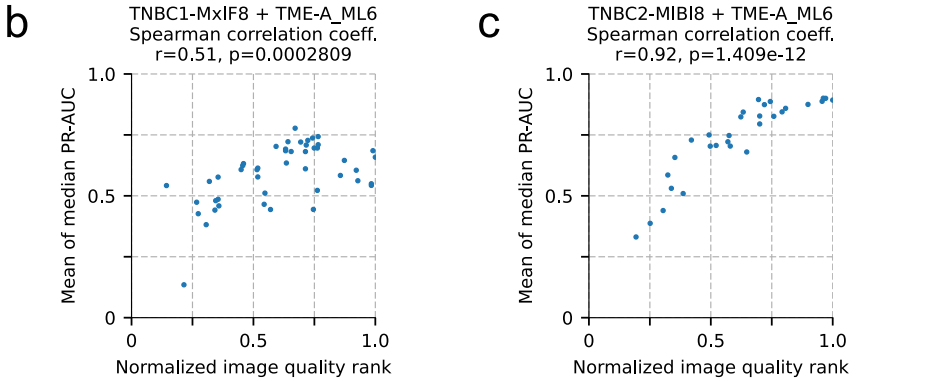


Fig. 3 | Correlation of classification performance with image-quality ranking based on signal-to-noise ratio (SNR). a Channel ranking parameters used for image ranking based on the pixel intensity standard deviation of the highest pixel intensities. **b** Scatter plot showing mean of median area under the precision-recall curve (PR-AUC) scores against normalized image-quality rank for the TNBC1-MxIF8 dataset and TME-A_ML6 annotations. Each dot represents a classifier that is trained on images from only a single patient. The corresponding normalized image-quality rank is computed from images from that respective patient. **c** Same as **b**, but for the TNBC2-MIBI8 dataset and TME-A_ML6 annotations.

Data 2), similar to our previous findings[15], relate to; high T-cell presence and low tumor burden relating to longer survival, and a dual role for macrophages, with macrophages proximity to tumor cells relating to shorter survival and clustering of macrophages among themselves relating to longer survival.

Discussion

In summary, CLEAR-IT demonstrates that contrastive learning-based encoders can be utilized for learning features for the task of cell classification in multiplex images with very sparse labels and cell locations, irrespective of imaging and analysis platforms. The label efficiency and self-supervised pre-training on the data of interest are of particular importance given the recent demonstration of demographic bias in existing computational pathology models and how SSL can address it[11,19].

CLEAR-IT was based on a relatively simple architecture that enabled the exploration of the hyperparameter space to improve its performance, and cross-dataset performance demonstrating transferability across similar datasets. Despite this, classification performance based on features learned by CLEAR-IT was comparable to classification performance based on cell expressions and areas, the state-of-the-art, and the combination of these two sets of features provided further performance gain. The F1-scores for the CLEAR-IT feature concatenated with cell expressions follow the F1-scores of the CLEAR-IT features only, rather than the cell expressions, when there is a drop in performance. We suspect that this is because, per image channel, CLEAR-IT produces 32 features compared to the 1 feature for cell expressions. This can be possibly addressed by introducing weights to the CLEAR-IT features and cell expression for the classifier to treat these inputs equally. Furthermore, when comparing classification performance on CLEAR-IT features and cell expressions, we observed a dependency on cell types, whether one or the other or their concatenation, performs best. While this could be partially attributed to CLEAR-IT features being higher in number, as mentioned earlier, it is also worthwhile to further investigate which type of feature input is beneficial for the cell types of interest and tailor the inputs to a classifier accordingly in order to maximize performance.

To systematically quantify the contribution of each learned feature for specific cell types, one could leverage the PLM-eXplain framework[20] to decompose the high-dimensional CLEAR-IT feature embeddings into a low-dimensional interpretable subspace aligned with known measurements (i.e., cell expression levels and cell sizes) and a residual subspace capturing complementary information. This would both perform implicit dimensionality reduction and enable targeted analysis of the most relevant features for each cell phenotype. We note, however, that some of the observed performance degradation may also stem from inaccuracies in the reference annotations used during supervised training. Future studies incorporating expert-curated labels will be important to disentangle model limitations from annotation errors when trying to investigate the features learned via SSL.

So far, we have limited our investigation to the ResNet-18[21] architecture for the choice of CLEAR-IT encoder in order to follow the SimCLR algorithm more closely, with the choice of the 18-layer variant due to a limit in computational resources. Further exploration of larger and possibly different network architectures, combined with deeper exploration of pre-training hyperparameters (e.g., a more comprehensive grid-search and additional augmentations), could result in higher-quality features being learned. During every supervised classifier training, we kept the pre-trained encoder's parameters fixed (a practice known as freezing), so that the encoder remains unchanged. Allowing the encoder to update its parameters during classifier training (a process known as fine-tuning) can further improve performance, albeit at the cost of computation time[13]. Lastly, investigation of different pre-training algorithms beyond SimCLR could also lead to improved feature learning. In future work, it could also be worthwhile to investigate the application of unsupervised techniques to the learned CLEAR-IT features, for example, clustering (see Supplementary Results, Supplementary Figs. 45–47 for an example of how *k*-means clustering and Leiden clustering could be used to analyze the extracted features). Furthermore, to eliminate the need for cell segmentations or cell locations, it could be worthwhile to investigate the application of the encoder to a full image by separating it into a grid of tiles similar to NaroNet[14].

Irrespectively, CLEAR-IT encoders demonstrate high performance in multiplexed image analysis using sparse labels and without the need for accurate segmentation masks, thereby eliminating a labor-intensive step and reducing hands-on processing time. Moreover, the features learned by CLEAR-IT are comparable to traditional handcrafted features, such as cell expressions and sizes, in predictive performance and yield even higher performance in combination, underscoring their complementary nature. This flexibility enables CLEAR-IT to be integrated into existing analysis pipelines, such as those implemented in QuPath[22], which supports user-in-the-loop training (Fig. 8), where they can be used either standalone or in combination with other feature sets. Importantly, our downstream survival analysis, conducted using a linear classifier trained with minimal supervision (from a single patient) on top of CLEAR-IT features, demonstrated that these features are clinically relevant, as the performance of the survival classifier aligns well with published results. There is also potential for further improvement by employing larger or more complex encoder architectures,

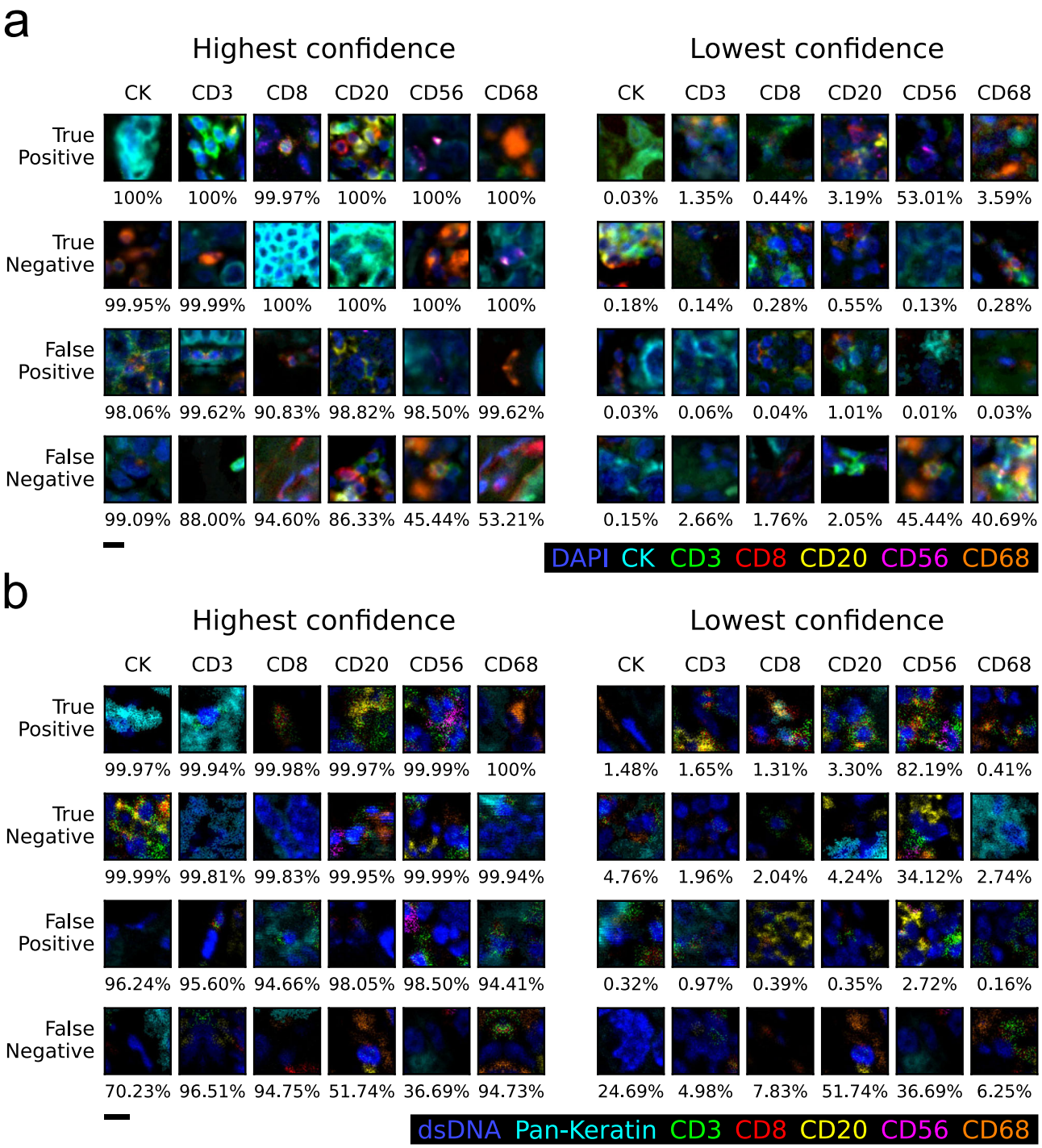


Fig. 4 | Examples of cell phenotype predictions made by CLEAR-IT. a Predictions for the TNBC1-MxIF8 dataset using the best classifier obtained (4th column in Fig. 1d). **b** Predictions for the TNBC2-MIBI8 dataset using the best classifier obtained (4th column in Fig. 1d). For every type of prediction and class label, a random cell from the 100 most confident (left) and 100 least confident (right) predictions is shown. Percentage values represent the mean confidence of the 100 most and least confident predictions, respectively. For better visibility, brightness and contrast have been enhanced for the shown images. Scale bars: 10 µm.

fine-tuning the encoder weights instead of using them in a frozen state, and applying dimensionality reduction techniques to focus on the most informative features for specific cell types, as discussed above.

In conclusion, by applying SSL to multiplex images for the purpose of single-cell phenotyping, CLEAR-IT offers a flexible solution for feature generation without the need for accurate cell segmentation, thereby enabling rapid analysis and supporting research-oriented digital pathology workflows. Our method provides the field of digital pathology and cancer research with a robust algorithm for this purpose.

Methods

Datasets

All datasets used are published and consist of multiplex images of tissue with individual cells to be identified. These are, briefly, TNBC1-MxIF8[1,23], TNBC2-MIBI44[4], CRC-CODEX26[9,17], and TONSIL-IMC41[18].

For the pre-training of a CLEAR-IT encoder, only the images are necessary, as well as a table containing the following information: cell coordinate, image file name, image channel number (this is not required if the images are already separated by channel beforehand). In practice, if cell

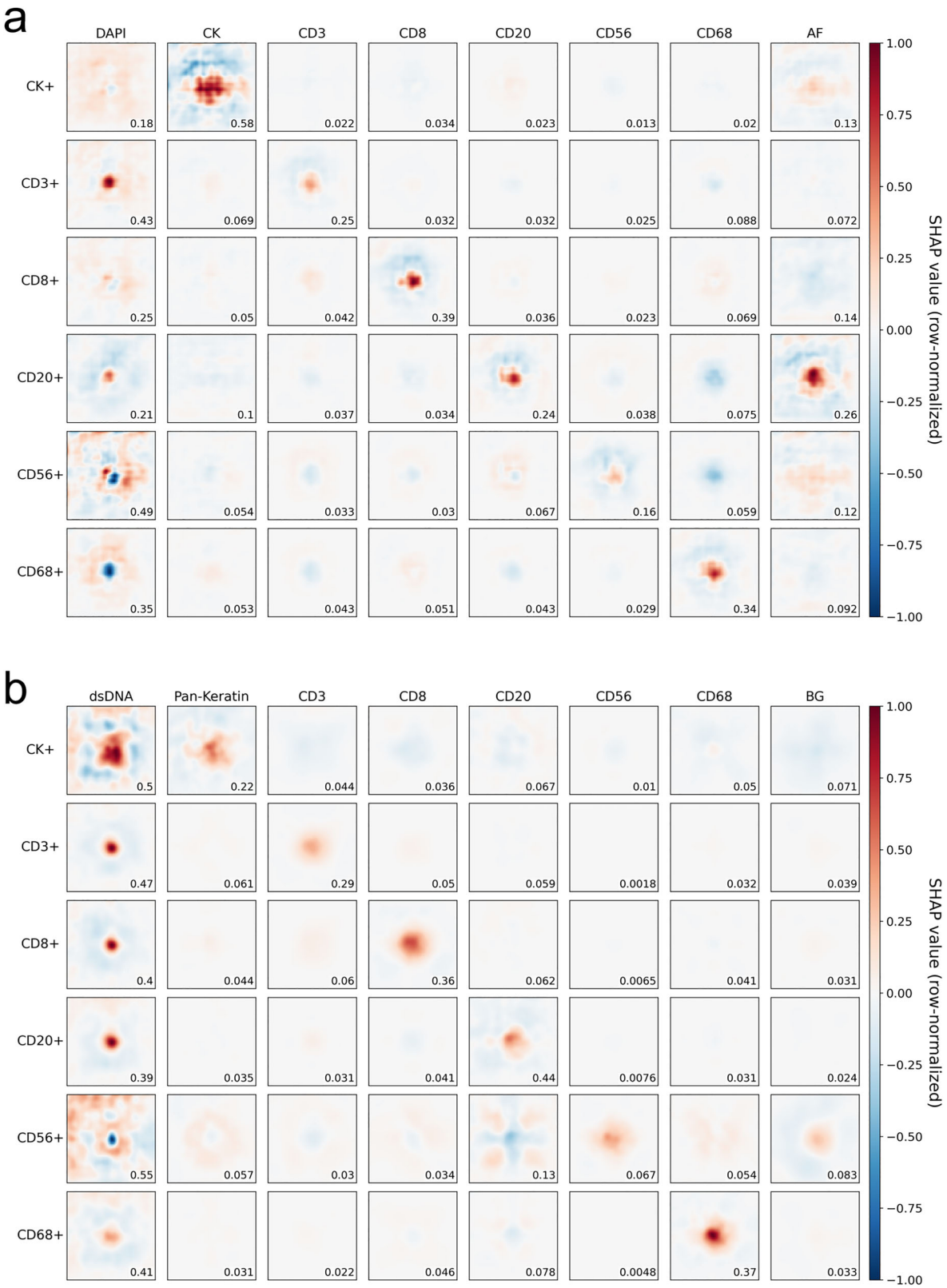


Fig. 5 | Aggregated SHAP values of the 100 most confident TP predictions of a linear classifier on top of CLEAR-IT encoders. For TNBC1-MxIF8 (**a**) and TNBC2-MIBI8 (**b**), for every channel (*x* axis) in the input image patch, the SHAP values indicate how strongly a pixel contributes towards (red) or against (blue) the classifier assigning the class label (*y* axis). SHAP values displayed are averaged over the 100 most confident TP (true positive) predictions per class (after removing duplicates: $n = 539$ (**a**) and $n = 526$ (**b**)), normalized per row, and correspond to the examples shown in Fig. 4a (left, first row) (**a**) and Fig. 4b (left, first row) (**b**). Numbers in lower right corners indicate the row-normalized importance (mean of the absolute value of all SHAP values). AF autofluorescence, BG background.

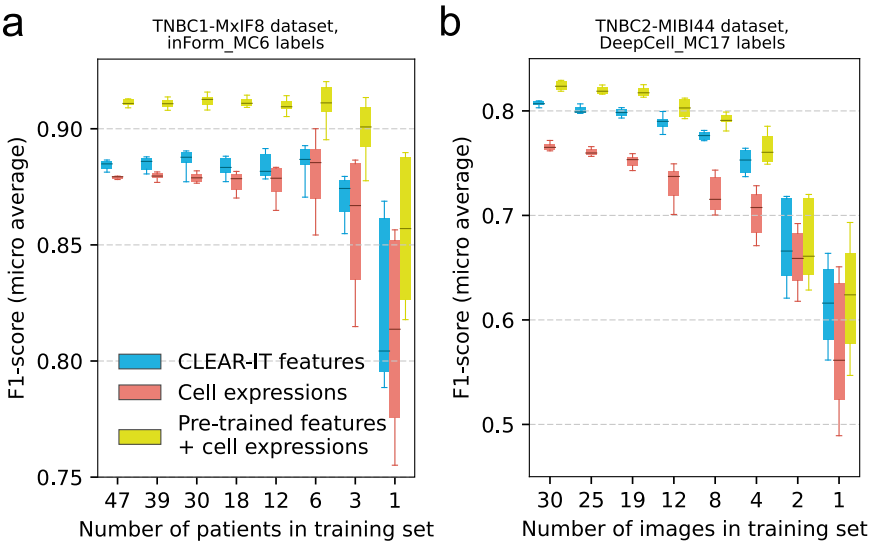


Fig. 6 | CLEAR-IT benchmark testing. a, **b** Box-and-whisker plots showing micro average F1-scores of four-layer MLPs trained on features learned by pre-trained CLEAR-IT encoders (blue), cell expressions and sizes (red), and a concatenation of both (yellow), with reduced amounts of labeled data for the **a** TNBC1-MxIF8 and **b** TNBC2-MIBI44 datasets with 10 bootstrappings. The boxes represent the interquartile range (25th to 75th percentiles) and the whiskers extend to the 5th and 95th percentiles of the data, excluding outliers.

Table 1 | Micro average F1-score comparison between CLEAR-IT and cell expression features across four datasets

Dataset	Training set (N)	CLEAR-IT	Cell Expr.
TNBC1-MxIF8	47 patients (761 images)	**0.885**	0.879
	1 patient (21 images)	0.804	**0.814**
TNBC2-MIBI44	30 images	**0.807**	0.765
	1 image	**0.616**	0.561
CRC-CODEX26	28 images	**0.826**	0.758
	1 image	0.569	**0.648**
TONSIL-IMC41	17 images	0.500	**0.556**
	1 image	0.239	**0.482**

Results are shown for both the full training set and a single image or patient. Best-performing values per row are in bold.

coordinates were unavailable from the annotations, we computed them from available nucleus segmentation using the scikit-image library[24]. In principle, coordinates could also be randomly generated, skipping the need to localize cells for pre-training. Although we have performed experiments where we pre-trained encoders on random locations, we do not discuss their results in the manuscript. Since all available datasets have nucleus segmentation masks provided, we also do not discuss the multitude of methods for obtaining them[25,26] and merely point out the algorithms used by the authors of the respective datasets in the following subsections.

For the supervised training of a classifier on top of a CLEAR-IT encoder, the images are necessary, as well as a table containing the following information: cell coordinate, image file name, and cell class label. Here, the cell coordinates should correspond to actual cell locations, since the encoder uses this information to extract an image patch and compute the cell features. The computed features serve as input to the classifier, and the cell class label serves as a supervised learning signal. Cell class labels can be obtained using any tool or method, as long as they are provided in a form suitable for multi-label or multi-class classification.

TNBC1-MxIF8 dataset

The TNBC1-MxIF8 dataset[1,23] consists of 1010 MxIF microscopy images of cancerous tissue from 62 TNBC patients. The images have a size of 1340 × 1008 pixels with pixel size of 0.5 µm, and have 8 channels: "DAPI", "CK", "CD3", "CD68", "CD8", "CD56", "CD20", "background". Overall, the images include approximately 2.3 million cells whose coordinates and cell phenotypes are specified in two separate sets of annotations, one being multi-label and the other being multi-class. The multi-label annotation set was obtained using TME-Analyzer[15] with segmentations based on StarDist[27] (referred to as "TME-A_ML6 labels") and contains the following 6 classes: "CK+", "CD3+", "CD8+", "CD20+", "CD56+", "CD68+". Supplementary Table 1 (Supplementary Data 2) contains information about cell counts in the training set of TME-A_ML6. The multi-class annotation set was obtained by Hammerl et al.[1] using inForm with segmentations also obtained using inForm (referred to as "inForm_MC7 labels") and contains the following 7 classes: "other", "CK", "CD3", "CD3 CD8", "CD20", "CD56", "CD68". Furthermore, we converted inForm_MC7 to the multi-label label set inForm_ML6: Here, the "CD3 CD8" class in inForm_MC7 was converted to both "CD3+" and "CD8+", being positive in the multi-label equivalent.

TNBC2-MIBI44 dataset

The TNBC2-MIBI44 dataset[4] consists of 41 MIBI-TOF images of cancerous tissue from 41 TNBC patients. The images have a size of 2048 × 2048 pixels with pixel size of 0.4 µm and have 44 channels: "Au", "Background", "Beta catenin", "Ca", "CD11b", "CD11c", "CD138", "CD16", "CD20", "CD209", "CD3", "CD31", "CD4", "CD45", "CD45RO", "CD56", "CD63", "CD68", "CD8", "dsDNA", "EGFR", "Fe", "FoxP3", "H3K27me3", "H3K9ac", "HLADR", "HLA_Class_1", "IDO", "Keratin17", "Keratin6", "Ki67", "Lag3", "MPO", "Na", "P", "p53", "Pan-Keratin", "PD-L1", "PD1", "phospho-S6", "Si", "SMA", "Ta", "Vimentin". The corresponding multi-class annotation set, obtained by Keren et al.[4] (referred to as "DeepCell_MC17 labels"), consists of segmentation masks for approximately 221,000 cells obtained using DeepCell and contains the following 17 classes: "Unidentified", "Tregs", "CD4 T", "CD8 T", "CD3 T", "NK", "B", "Neutrophils", "Macrophages", "DC", "DC/Mono", "Mono/Neu", "Other immune", "Endothelial", "Mesenchymal-like", "Tumor", "Keratin-positive tumor".

TNBC2-MIBI8 dataset

To make the images from the TNBC2-MIBI44 dataset qualitatively similar to those from the TNBC1-MxIF8 dataset, 8 out of the original 44 image channels were extracted: "dsDNA", "Pan-Keratin", "CD3", "CD68", "CD8", "CD56", "CD20", "background". Using these reduced images, a multi-label annotation set was obtained using TME-Analyzer[15] with segmentations based on StarDist[27] (referred to as "TME-A_ML6 labels"), which contains the same 6 classes that were used for the TNBC1-MxIF8 dataset: "CK+", "CD3+", "CD8+", "CD20+", "CD56+", "CD68+".

Table 2 | Micro average F1-score performance of the concatenation of CLEAR-IT features and cell expressions in the high-label regime, along with the relative improvement over the better-performing individual input

Dataset	Concatenation of CLEAR-IT and cell expr.	Improvement over the best individual input
TNBC1-MxIF8	0.911	+2.9% over CLEAR-IT
TNBC2-MIBI44	0.824	+2.1% over CLEAR-IT
CRC-CODEX26	0.842	+1.9% over CLEAR-IT
TONSIL-IMC41	0.594	+6.8% over cell expressions

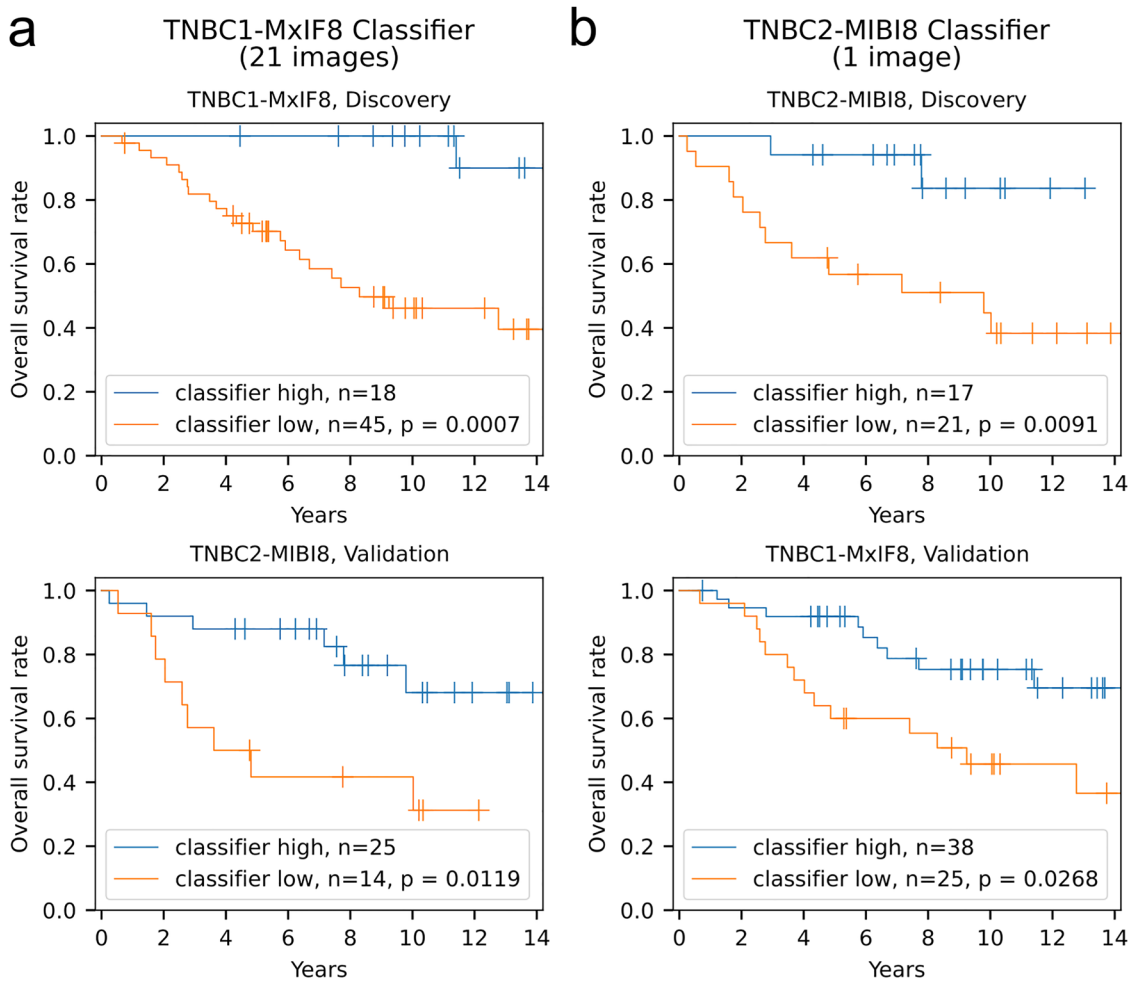


Fig. 7 | CLEAR-IT clinical testing. a, **b** Kaplan–Meier curves for two-group-survivals of patients in the TNBC1-MxIF8 and TNBC2-MIBI datasets based on a CLEAR-IT encoding—minimal label linear cell classifier—survival classifier architecture, with supervised training performed exclusively on discovery cohort; **a** TNBC1-MxIF8 discovery (top) and TNBC2-MIBI8 validation (bottom) cohorts; **b** TNBC2-MIBI8 discovery (top) and TNBC1-MxIF8 validation (bottom) cohorts. *p* values were obtained according to the log-rank test.

To convert the DeepCell_MC17 annotation set to the reduced images, the following label assignments were performed to obtain the DeepCell_ML6 annotation set:

- "CD3+": "Tregs", "CD4 T", "CD3 T", "CD8 T"
- "CD8+": "CD8 T"
- "CD56+": "NK"
- "CD20+": "B"
- "CD68+": "Macrophages"
- "CK+": "Keratin-positive tumor"

CRC-CODEX26 dataset

The CRC-CODEX26 dataset[17] consists of 35 co-detection by indexing (CODEX) images of cancerous tissue from 35 colorectal cancer (CRC) patients. The images have a size of 1920 × 1440 pixels with pixel size of 0.377 μm and have 26 channels: "CD11b", "CD11c", "CD15", "CD163", "CD20", "CD3", "CD31", "CD34", "CD38", "CD4", "CD45", "CD56", "CD57", "CD68", "CD8", "Collagen", "Cytokeratin", "FOXP3", "HLADR", "MUC1", "NAKATPASE", "PDPN", "SYP", "VIM", "SMA", "CD45RA". The corresponding multi-class annotation set, obtained by Amitay et al.[9] with segmentations based on Mesmer[28] (referred to as "CellSighter_MC14 labels"), consists of segmentation masks for 85,179 cells and contains the following 14 classes: "Bcell", "CD3 T", "CD4 T", "CD8 T", "DC", "Endothelial", "Lymphatic", "Macrophage", "Neuron", "Neutrophil", "Plasma", "Stroma", "Treg", "Tumor".

TONSIL-IMC41 dataset

The TONSIL-IMC41 dataset consists of 24 imaging mass cytometry (IMC) images of human tonsil tissue from 7 individuals. The images have sizes ranging from 465 × 464 pixels to 686 × 617 pixels with pixel size of 1 μm and

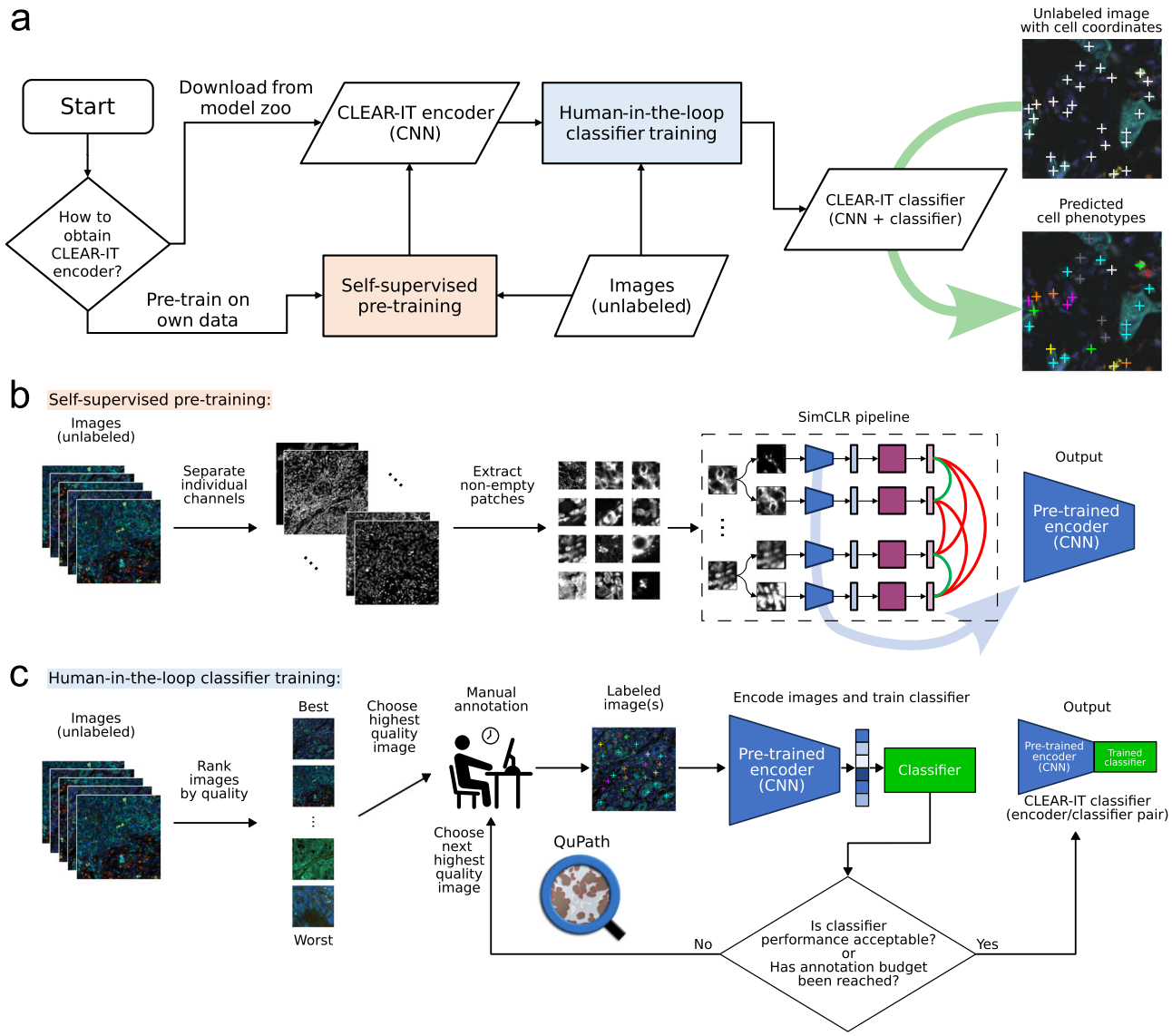


Fig. 8 | CLEAR-IT usage guideline. a Flowchart showing the process from beginning with an unlabeled dataset to obtaining a CLEAR-IT classifier (encoder + classifier pair) that can perform cell phenotyping. Users can obtain a CLEAR-IT encoder either by pre-training it themselves (red box, **b**) or by downloading an already pre-trained one. The classifier on top of the encoder can be trained in a human-in-the-loop process (blue box, **c**). **b** For the self-supervised pre-training, multiplex images are split into their individual channels, and non-empty grayscale patches are extracted. The patches serve as input to the SimCLR pipeline, which produces a pre-trained encoder. **c** For the human-in-the-loop training of the cell classifier, the unlabeled images are ranked by quality. Annotation efforts are then focused on the highest quality images and iterated upon until the classification performance is acceptable or the annotation budget has been reached. The human-in-the-loop classification is facilitated by loading a pre-trained CLEAR-IT encoder into existing software, such as QuPath. Icons in **c** are licensed under CC BY 3.0: "workspace" by Universal Icons, "clock" by Salman Azzumardi.

have 41 channels: "Sars_CoV2_Spike", "CD45RO", "CD45RA", "CD68", "CD8a", "Ki67", "Collagen1", "CD138", "CD163", "IL1R", "CD42b", "MPO", "SARSCov2_Capsid", "B7_Complement", "CD56", "Podoplanin", "CD69", "EPCAM", "CD206", "CD79a", "STING", "TMPRSS2", "AQP5", "CD1c", "IFITM3", "ACE2", "CD57", "p16", "IL6R", "cCaspase3", "CD61", "CD3", "ProSPC", "CD31", "C30-30", "CD4", "HLADR", "CD169", "193Ir", "CD147", "b2M". The images, segmentation masks, and corresponding multi-class annotation set were obtained by Hunter et al. as part of OPTIMAL (an OPTimized Imaging Mass cytometry AnaLysis framework for benchmarking segmentation and data exploration)[18]. The multi-class label set (referred to as "OPTIMAL_MC21 labels") consists of segmentation masks for 109,535 cells and contains the following 21 classes: "Epithelium", "Memory CD8 T cells", "M2 Macrophages", "B cells", "Memory CD4 T cells", "Epithelium (proliferating)", "Folicular B cells (proliferating)", "Mature Macrophages", "Folicular T cells", "Plasma cells", "Endothelium", "Unclassified", "Effector CD4 T cells", "Epithelium and Immune cells", "Naïve CD8 T cells", "anti-Inflammatory Macrophages", "CD4 T cells", "STING+ cells", "Apoptotic cells", "Germinal center Macrophages", "Standard Macrophages".

Image intensity scaling

We leave all images in the same format as they are published (either 8-bit or 16-bit integer). Before feeding images to a neural network, we divide them by a constant to scale the intensity range to [0, 1]: 255 for TNBC1-MxIF8, 265 for TNBC2-MIBI44 and TNBC2-MIBI8, and 65535 for CRC-CODEX26 and TONSIL-IMC41.

Train/test splits

For every experiment, an identical split of data into a train and test set was used. For the considered datasets, the train/test splits are: TNBC1-MxIF8

47/15 patients (761/249 images), TNBC2-MIBI44 and TNBC2-MIBI8 30/11 patients (30/11 images), CRC-CODEX26 28/7 patients (28/7 images), TONSIL-IMC41 5/2 patients (17/7 images). For the supervised training of classifiers, the training set was further split by 80/20 into a training and validation set. To ensure that classifiers are trained and tested on a comparable amount of data for the cross-testing of TNBC1-MxIF8 and TNBC2-MIBI8, the train/test split was set to 30/11 patients for both datasets. Furthermore, for TNBC1-MxIF8, only 3 images per patient (captured at the tumor center) were included. This balanced the amount of training and testing data between the TNBC1-MxIF8 images (1340 × 1008 pixels) and the TNBC2-MIBI8 images (2048 × 2048 pixels) as much as possible to approximately 4 megapixels of image data per patient.

CLEAR-IT training pipeline

The CLEAR-IT training pipeline is summarized in Supplementary Fig. 1. The inference pipeline (Supplementary Fig. 1a) takes a multiplex image patch centered around a cell as input and outputs a vector containing probabilities of the cell belonging to the class labels considered. To achieve this, the multiplex image patch is first split into its individual channel components, which results in one grayscale single-channel image patch for each image channel. The grayscale image patches are then fed to the pre-trained encoder network (specifically, a ResNet encoder[21]), which outputs a feature vector. The individual channel feature vectors are concatenated and fed into a classification network, which predicts the cell's class label.

Unsupervised encoder pre-training

Pre-training of the encoder network follows the SimCLR algorithm[12] and is summarized in Supplementary Fig. 1b. The pre-training dataset consists of grayscale single-channel image patches centered around cells from individual color channels of the multiplex images. A minibatch is constructed by sampling a number of different image patches of random cell locations and random image channels. Every image patch in the minibatch is augmented in two random and distinct ways to create an augmented image pair, which is then encoded by a ResNet-18[21] encoder. The encodings are non-linearly projected by the projection head, which is an MLP with three hidden layers. The loss function acts on the non-linear projections of the encodings to maximize the similarity between the outputs originating from an augmented image pair (positive pair), while minimizing the similarity across projections from augmented images that are not from the same image pair (negative pairs). After pre-training, the MLP is removed, and the ResNet encoder can be used to provide feature embeddings as input for training downstream networks.

The input size to the ResNet-18 encoder is 64 × 64 × 1 pixels (32 × 32 × 1 for the TONSIL-IMC41 dataset), and the output size is either 512 or 32, for a linear classifier or an MLP classifier, respectively, which is used to perform classification. The MLP projection head has an output layer of size 128 and 3 hidden layers, each of size 512, with batch normalization and ReLU (rectified linear unit) activations. Optimization is performed via the LARS (layer-wise adaptive rate scaling)[29] method with a learning rate of $0.3N/256$ (N being the pre-training batch size), and the NT-Xent (normalized-temperature cross-entropy) loss function

$$\mathcal{L}^{\mathrm{NT-Xent}} = -\frac{1}{N}\sum_{i,j\in\mathcal{MB}} \log\left(\frac{\exp\left(\mathrm{sim}\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N}\mathbf{1}_{[k\neq i]}\exp\left(\mathrm{sim}\left(z_i, z_k\right)/\tau\right)}\right)$$

where $i, j \in \mathcal{MB}$ represents all augmented image pairs in the minibatch (i.e. positive pairs), z_i is the non-linear feature projection of the ith image in the minibatch, $\mathrm{sim}(a, b)$ is the cosine similarity between vectors a and b,$\mathbf{1}_{[k\neq i]}$ is an indicator function that is equal to 1 iff $k \neq i$ and 0 otherwise, and τ is the temperature parameter that is used to scale the relative importance of positive and negative pairs.

During the optimization of pre-training augmentations, we apply combinations of augmentations in the following order:

1. **Gaussian blur**: Apply Gaussian blur with intensity x, where x determines the kernel size as $2x + 1$ and the standard deviation randomly drawn from $\sigma \in [0, x]$.
2. **Brightness/Contrast**: Adjust brightness and/or contrast of the image by x, which determines the intensity that is randomly drawn from $\beta \in [-x, x]$.
3. **Translation**: Move the center of the crop in a random radial direction. The translation distance is drawn from $\rho \in [0, x]$.
4. **Zoom**: Zoom in or out of the image. The zoom factor is randomly drawn from $Z \in [x_{\mathrm{zmin}}, x_{\mathrm{zmax}}]$ where $x = 1$ means that no zoom is applied.
5. **Rotation**: Randomly apply rotation in multiples of 90°.
6. **Flipping**: Randomly apply horizontal and/or vertical flipping.

The optimization of the hyperparameters is illustrated in Supplementary Fig. 1d.

Supervised classifier training

The supervised classifier training is the second step in the training pipeline and is summarized in Supplementary Fig. 1c. The pre-trained encoder is used as many times in parallel as there are color channels in the input image and the computed feature encodings are concatenated and input to a classification network. For every experiment, we keep the weights of the pre-trained ResNet encoder frozen and only train the classifier to make predictions based on the generated feature embeddings.

In order to evaluate the quality of the features learned by the pre-trained encoder, we use a single-layer perceptron (SLP), in other words, a linear classifier, to make predictions about cell classes. This is referred to as linear evaluation[12]. The linear classifier consists of a single layer with input size $512 \times C$ and output size K, where C is the number of color channels in the input image, and K is the number of classes to predict. We train every linear classifier with a batch size of 64, a 30% dropout layer, using the Adam optimizer with both learning rate and weight decay of 10^{-4} and minimizing the binary cross-entropy loss. For every training run, we hold out a random 20% of the training set for validation, train for 50 epochs and choose the weights from the epoch where the classifier achieved the lowest loss on the validation set.

Since the class distribution in the used datasets is highly imbalanced, we undersampled the majority classes to ensure balanced training sets (see Supplementary Methods for an explanation of the algorithm).

3-round encoder optimization

The encoder hyperparameter search for optimizing downstream classification performance is performed in 3 separate rounds. Every change in hyperparameters results in a separate ResNet encoder-linear classifier pair being trained as described previously. The optimization goal is the maximization of the classifier's performance, which we measure with the area under the PR-AUC and report the mean of medians for PR-AUC of each class. See Supplementary Results for further details.

Cross-testing experiments

For the experiments that compare classifier performance across TNBC1-MxIF8 and TNBC2-MIBI8 and their two respective annotation sets, we use two encoders that are pre-trained on the same dataset that is used for the supervised test set, i.e., the encoders are pre-trained on TNBC1-MxIF8 dataset for the upper half of the table in Supplementary Fig. 12a and on the TNBC2-MIBI8 dataset for the lower half of the table in Supplementary Fig. 12b. For both encoders, the remaining hyperparameters are as follows: Pre-training batch size $N = 256$, loss function temperature $\tau = 0.05$, minimum zoom factor $x_{\mathrm{zmin}} = \frac{2}{3}$, maximum zoom factor $x_{\mathrm{zmax}} = \frac{4}{3}$, and pre-training dataset size $D = 819{,}200$ image patches sampled from 31 patients from TNBC1-MxIF8 or TNBC2-MIBI8.

Label-reduction by patient count

For the experiments that investigate the effects of reducing the number of patients in the supervised training set, we use the same encoders as for the cross-testing experiments, again depending on the supervised test set. For both datasets, we begin by using the maximum amount of patient data available and reduce the patient counts, resulting in classifiers trained with data from $P_1 \in \{47, 39, 30, 18, 12, 6, 3, 1\}$ patients for the TNBC1-MxIF8 dataset and $P_2 \in \{30, 25, 19, 12, 8, 4, 2, 1\}$ patients for the TNBC2-MIBI8 dataset. For every case with a reduced number of patients, 10 separate classifiers are trained on image data from different combinations of randomly and independently sampled patients.

Supervised classifier training on individual patient data

For training the supervised classifiers on image data from individual patients, we use the same encoder pre-trained on the TNBC1-MxIF8 dataset (Fig. 2a) and TNBC2-MIBI8 dataset (Fig. 2b), respectively. Since, for some patients, there are none or only very few cells of a certain phenotype, balancing the training sets based on the least frequent cells is infeasible. Instead, we set a fixed target count $T = 500$ cells per label, for the amount that every class should be represented in the training sets. If the available amount is less than T, all available cells are included in the training set.

Patient image-quality ranking for correlation analysis

For determining the quality of images belonging to a patient, we compute a metric that serves as a surrogate for the image SNR. We consider the brightest pixels in a signal channel as representing the true signal of interest, while the overall pixel distribution in the background channel is indicative of noise. In practice, a high standard deviation among the top 0.1% brightest pixels in signal channels suggests a strong, well-defined signal that stands out against the noise, whereas a low standard deviation in the background channel indicates lower noise levels[30]. Following these principles, we rank images as follows:

1. Calculate the standard deviations of pixels across all images belonging to an individual patient for the following channels:
   - CD3, CD8, CD20, CD56, CD68 channels: Standard deviation of top 0.1% brightest pixels.
   - Background channel: Standard deviation of all pixels.
2. For each patient, assign an ordinal rank from 1 to N, where N is the number of patients, based on the following criteria:
   - CD3, CD8, CD20, CD56, CD68 channels: High rank for high standard deviation.
   - Background channel: High rank for low standard deviation.
3. For each patient, sum the ranks assigned for each channel.
4. Divide all rank sums by the highest obtained.

The resulting rank scores range between $1/N$ for the lowest image quality and 1 for the highest image quality per patient.

Relating CLEAR-IT output to clinical data

The ranking of parameters and survival classifiers was obtained as described previously[15]. In brief, 50 center parameters per patient were used, namely: densities of immune cell populations in tumor and stroma compartments ($n = 12$); areas of these compartments ($n = 2$); and distance z-scores between phenotypes in whole tissue ($n = 36$), and were ranked with a reiterative nested Monte Carlo approach. Here, the patient dataset was randomly split into train and test sets, and interim classifiers were constructed in the train set. First, the representative values of parameters for long and short survival were obtained through repeating the following steps 1000 times:

1. Random selection of 12 patients;
2. Splitting these 12 patients for each of 50 parameters into 2 groups of 6 patients according to the median value of parameters;
3. Testing the statistical significance of survival differences between the two patient groups per parameter according to the log-rank tests, and finally.
4. Upon significance, recording the parameter mean values for the shorter and longer survival groups.

After 1000 repetitions, the means of statistically significant parameters were calculated for shorter and longer survival groups, and the parameters were ranked based on the degree of separation (p value) between the survival groups obtained in the train set. Second, the patients in the train set were split into two groups, based on most parameters for each patient being closer to the recorded means of shorter or longer survivors, where the lowest ranking parameter was reiteratively excluded until statistically significant performance was reached. Third, the surviving interim classifier (i.e., the resulting set of parameters and their recorded means) was tested in the test set. In the case when statistical significance was also observed for the test set, then the set of parameters that defined the classifier was considered to have a first hit. These three steps were repeated 5000 times, and the parameters were ranked based on the total number of hits in the interim classifiers. Subsequently, the top 10 ranking parameters were used to obtain the final survival classifier using the steps 1–4 above on the whole discovery cohort, but with a repetition of 5000 instead of 1000, and then applied to the discovery and validation cohort.

CLEAR-IT feature benchmark against MAPS algorithm

In order to benchmark the quality of feature encodings produced by a pre-trained CLEAR-IT encoder, we utilize the MAPS[10] architecture for classification. Briefly, MAPS consists of an MLP with four hidden layers that performs cell classification based on cell expressions and cell size, resulting in an input size of $C + 1$, where C is the number of channels in the image, and an output size of K, where K is the number of class labels. In order to perform classification of CLEAR-IT features, we pre-train encoders to output feature vectors of size 32 and adapt the MAPS input size to be $32 \times C$. Lastly, we investigate the performance of concatenating both inputs, resulting in an input size of $C \times 33 + 1$. When training MAPS on cell expressions, we use the following training parameters: batch size 128, 25% dropout, Adam optimizer with learning rate of 10^{-4} and minimizing the binary cross-entropy loss. When training MAPS on CLEAR-IT features and their concatenation with cell expressions, we use the following training parameters: batch size 1024, 25% dropout, Adam optimizer with learning rate of 10^{-4} and minimizing the binary cross-entropy loss.

Calculation of cell size and cell expressions

In order to provide input to the MAPS architecture, we calculate the cell sizes and cell expressions from the available images. Cell sizes are determined by counting the pixels belonging to a particular cell in the segmentation masks. Cell expressions per channel are calculated by summing the pixel intensities inside segmented regions per channel and dividing this by the cell size.

SHAP analysis of selected high- and low-confidence predictions

We performed pixel-level attribution via SHAP[16] on the cells whose predictions are illustrated in Fig. 4 for both TNBC1-MxIF8 and TNBC2-MIBI8. For each of the six marker classes (CK, CD3, CD8, CD20, CD56, CD68) and each prediction outcome (TP, TN, FP, FN), we pooled the 100 most and 100 least confident examples (see Supplementary Notes for calculation of prediction confidence), removed duplicate cell IDs across classes, and extracted 64×64 pixel patches centered on those cells from the eight-channel images.

With the encoder and classifier in evaluation mode (frozen encoder), we wrapped the model in a PyTorch DeepExplainer using an all-zero baseline of the same shape. SHAP values were computed in batches over all selected patches, producing an attribution tensor of shape $(N, C, 3, H, W, K)$, where N is the number of patches, $C = 8$ channels, $H = W = 64$, and $K = 6$ classes. We then averaged these values over the batch, yielding per-channel, per-class heatmaps of size (C, H, W, K) (Supplementary Figs. 19–34). To improve visual coherence, each spatial map was Gaussian-blurred ($\sigma = 2$ pixels).

Channel-class importance and summary quantities

For each scenario (e.g., top-confidence TP), let $S_{n,c,i,j,k}$ denote the SHAP value for sample n, channel $c \in \{1, \ldots, C\}$, pixel (i, j) with $H = W = 64$, and class $k \in \{1, \ldots, K\}$. After Gaussian smoothing of spatial maps ($\sigma = 2$), we aggregate absolute attributions over space and samples to obtain a channel-class importance matrix

$$I_{c,k} = \frac{1}{N\,H\,W} \sum_{n=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} |\, S_{n,c,i,j,k} \,|.$$

To compare channels within each class, we row-normalize by class

$$R_{c,k} = \frac{I_{c,k}}{\sum_{c'} I_{c',k}} \text{ so that } \sum_{c} R_{c,k} = 1 \text{ for each } k,$$

and visualize R in Fig. 1g (channels as columns, classes as rows). For a compact per-class summary we report: (i) the diagonal share $d_k = R_{m(k),k}$, where $m(k)$ pairs each class with its marker channel by name matching (including the special case Pan-Keratin ↔ CK+); (ii) the nuclear share $u_k = R_{n,k}$ with n chosen as dsDNA for MIBI and DAPI for MxIF; and (iii) the Top-2 share $t_k = \min\left(1,\, d_k + u_k\right)$. Dataset-level summaries are the mean values across classes of these quantities; for interpretability, we also report fold-enrichment relative to a uniform $1/C$ baseline, defined as (observed share)/$(1/C)$. All matrices and summaries were computed with the encoder and classifier in evaluation mode on the 100 most confident TP predictions per class (deduplicated by cell ID).

Hardware and software implementation

Code has been developed and tested on a machine with an Intel Core i9-10900X CPU, 128 GB RAM and an NVIDIA GeForce RTX 3090 GPU (24GB VRAM) within the PyTorch Release 24.07 Docker container provided by NVIDIA Optimized Frameworks. Image augmentations are implemented using Kornia[31] and certain image processing steps using scikit-image[24].

Training and inference times depend on the dataset. Encoder pre-training during the hyperparameter optimizations took ~2 hours per network, and supervised training of a linear classifier took around 10 minutes. Excluding loading times, inference can be performed at a speed of ~4500 cell image patches (of size 64 × 64 × 8 pixels) per second.

Statistics and reproducibility

All statistical analyses were performed in Python using standard scientific computing packages. No statistical methods were used to predetermine sample sizes; sample sizes were dictated by the availability of publicly released multiplex imaging datasets. All experiments using human tissue images were conducted once per dataset, and reproducibility was assessed by repeated training-validation splits, bootstrapping, or cross-evaluation procedures as described below.

For linear evaluation experiments, each reported performance value represents the median PR-AUC across classes, computed on held-out test sets defined at the patient level. When assessing label-reduction or image-quality effects, classifiers were retrained independently for each condition; each "replicate" corresponds to an independently trained encoder-classifier pair using a distinct subset of patients or images. For benchmarking with 4-layer MLP classifiers, results reflect 10 bootstrapped training runs per condition.

Correlation analyses between image-quality metrics and classification performance were evaluated using Spearman's rank correlation coefficient. Survival analyses were conducted using Kaplan–Meier estimators with log-rank tests. All clustering quality metrics (adjusted mutual information, adjusted Rand index, homogeneity) were computed on test-set embeddings only.

All code, deterministic preprocessing steps, and pre-trained CLEAR-IT encoders are publicly available to ensure reproducibility of feature extraction, model training, and evaluation[32,33].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

• TNBC1-MxIF8 dataset[1,23]: Images available to download upon request at https://doi.org/10.4121/126d8103-6de5-4493-a48e-5d529fef471e – Segmentation masks available to download at https://doi.org/10.4121/ebc792ad-4767-4aef-b8ff-ae653e901e3f • TNBC2-MIBI44 dataset[4]: – Images available to download at https://www.weizmann.ac.il/mcb/Keren/resources– Segmentation masks available to download at https://www.angelolab.com/mibi-data• CRC-CODEX26 dataset[17]: – Images available to download at https://doi.org/10.7937/TCIA.2020.FQN0-0326 – Segmentation masks provided by authors of ref. 9• TONSIL-IMC41 dataset[18]: – Images and segmentation masks available to download at https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BSST1047. For survival analyses, overall survival outcome data for the TNBC1-MxIF8 and TNBC2-MIBI cohorts were taken from the corresponding published studies[1,4]; no new clinical outcome data were collected in this work. Source data underlying the graphs in the main paper are provided in Supplementary Data 1. Data underlying the tables shown in Supplementary Information are provided in Supplementary Data 2. All data tables, pre-trained encoder models, and supervised classifier models are available to download at https://doi.org/10.4121/ebc792ad-4767-4aef-b8ff-ae653e901e3f.

Code availability

The source code used in this study is openly available at https://github.com/qnano/CLEAR-IT and archived in a persistent repository[33].

Received: 12 November 2024; Accepted: 24 March 2026;
Published online: 16 April 2026

References

1. Hammerl, D. et al. Spatial immunophenotypes predict response to anti-PD1 treatment and capture distinct paths of T cell evasion in triple negative breast cancer. *Nat. Commun.* **12**, 5668 (2021).
2. Cabrita, R. et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **577**, 561–565 (2020).
3. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
4. Keren, L. et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**, 1373–1387.e19 (2018).
5. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
6. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
7. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
8. Feng, Y. et al. Spatial analysis with SPIAT and spaSim to characterize and simulate tissue microenvironments. *Nat. Commun.* **14**, 2697 (2023).
9. Amitay, Y. et al. CellSighter: a neural network to classify cells in highly multiplexed images. *Nat. Commun.* **14**, 4302 (2023).
10. Shaban, M. et al. MAPS: pathologist-level cell type annotation from tissue images through machine learning. *Nat. Commun.* **15**, 28 (2024).
11. Movva, R., Koh, P. W. & Pierson, E. Using unlabeled data to enhance fairness of medical AI. *Nat. Med.* **30**, 944–945 (2024).

12. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *Proc. Mach. Learn. Res.* **119**, 1597–1607 (2020).
13. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* **7**, 100198 (2022).
14. Jiménez-Sánchez, D. et al. NaroNet: discovery of tumor microenvironment elements from highly multiplexed images. *Med. Image Anal.* **78**, 102384 (2022).
15. Balcioglu, H. E. et al. TME-analyzer: a new interactive and dynamic image analysis tool that identified immune cell distances as predictors for survival of triple negative breast cancer patients. *npj Imaging* **2**, 21 (2024).
16. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).
17. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359 (2020).
18. Hunter, B. et al. OPTIMAL: an OPTimized Imaging Mass cytometry AnaLysis framework for benchmarking segmentation and data exploration. *Cytom. A* **105**, 36–53 (2024).
19. Vaidya, A. et al. Demographic bias in misdiagnosis by computational pathology models. *Nat. Med.* **30**, 1174–1190 (2024).
20. van Eck, J., Gogishvili, D., Silva, W. & Abeln, S. PLM-explain: divide and conquer the protein embedding space. *Bioinformatics* **42**, btaf631 (2026).
21. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
22. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep*. **7**, 16878 (2017).
23. Spengler, D. & Balcioglu, H. E. TNBC1-MxIF8: Multiplex immunofluorescence images of triple-negative breast cancer tissue. *4TU.ResearchData.* https://doi.org/10.4121/126d8103-6de5-4493-a48e-5d529fef471e (2026).
24. van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
25. Wen, T. et al. Review of research on the instance segmentation of cell images. *Comput. Methods Prog. Biomed.* **227**, 107211 (2022).
26. Haase, R. et al. A hitchhiker's guide through the bio-image analysis software universe. *FEBS Lett.* **596**, 2472–2485 (2022).
27. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 265–273 (Springer, 2018).
28. Greenwald, N. F. et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
29. You, Y., Gitman, I. & Ginsburg, B. Large batch training of convolutional networks. Preprint at https://arxiv.org/abs/1708.03888 (2017).
30. Gonzalez, R. C. & Woods, R. E. *Digital Image Processing*, 3rd edn. (Pearson Prentice Hall, Upper Saddle River, NJ, USA, 2008).
31. Riba, E., Mishkin, D., Ponsa, D., Rublee, E. & Bradski, G. Kornia: an open source differentiable computer vision library for PyTorch. In *Proc. IEEE Winter Conference on Applications of Computer Vision* 3663–3672 (2020).
32. Spengler, D. & Balcioglu, H. E. CLEAR-IT supplementary data: models, embeddings, and analysis outputs for tumor microenvironment cell phenotyping. *4TU.ResearchData.* https://doi.org/10.4121/ebc792ad-4767-4aef-b8ff-ae653e901e3f (2026).
33. Spengler, D. & Balcioglu, H. E. CLEAR-IT, a framework for contrastive learning to capture the immune composition of tumor microenvironments. *4TU.ResearchData.* https://doi.org/10.4121/365ab556-b03f-49d9-b8b8-58f48aae85ec (2026).

Acknowledgements

Icons in Fig. 8c are licensed under CC BY 3.0: "workspace" by Universal Icons, "clock" by Salman Azzumardi. We thank Dora Hammerl, John Martens (both from Erasmus MC, the Netherlands), Yael Amitay, Leeat Keren, Yuval Bussi (all from Weizmann Institute of Science, Israel), Michael Angelo (Stanford University, USA), Christian Schürch (University of Tübingen, Germany), Garry Nolan (Stanford University, USA), Bethany Hunter, Andew Filby, George Merces (all from Newcastle University, UK), Paul Rees (Swansea University, UK), Muhammad Shaban (Brigham and Women's Hospital, USA), Yunhao Bai (Broad Institute, USA), Huaying Qiu, Sizun Jiang (both from Beth Israel Deaconess Medical Center, USA), and Faisal Mahmood (Brigham and Women's Hospital, USA) for making their datasets and analyses available. We also thank members of the laboratory of tumor immunology (Erasmus MC, the Netherlands), Rolf Harkes (Netherlands Cancer Institute, the Netherlands), Thibaut Goldsborough (University of Edinburgh, UK), and Dimitri Kromm (Delft University of Technology, the Netherlands) for valuable discussions. This work was supported in part by a contribution from the National Growth Fund program NXTGEN Hightech, Erasmus MC's Academic Center of Excellence for Tumor Immunology and Immune Therapy (ACE TI-IT), and the Wellcome Trust 223750/Z/21/Z. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

D.S. led the work and developed the CLEAR-IT architecture; D.S., C.S., and H.B. conceptualized the work; D.S. and H.B. curated the data, performed analyses, and prepared figures; S.K. provided technical feedback for the CLEAR-IT architecture; P.B. wrote the code for integrating CLEAR-IT into QuPath; D.S., K.P., C.S., and H.B. wrote the manuscript; R.D., C.S., and H.B. supervised the work; all authors read the manuscript and provided critical feedback.

Competing interests

The authors declare the following competing interests: Reno Debets has received research support from MSD and Bayer, personal fees from Bluebird Bio, Genticel, and other support from Pan Cancer T outside the submitted work (all paid to the Erasmus MC Cancer Institute), as well as European patent application nos 21152822.9, 24192717.7, and 25163944.9 (pending to Erasmus MC). The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-026-09984-2.

Correspondence and requests for materials should be addressed to Hayri E. Balcioglu or Carlas Smith.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Johannes Stortz and Aylin Bircan. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026