

## Deep learning and transfer learning identify breast cancer survival subtypes from single-cell imaging data

Shashank Yadav<sup>1,2</sup>, Shu Zhou<sup>1,2</sup>, Bing He<sup>1</sup>, Yuheng Du<sup>1</sup> & Lana X. Garmire<sup>1</sup>  <sup>1</sup>

### Abstract

**Background** Single-cell multiplex imaging data have provided new insights into disease subtypes and prognoses recently. However, quantitative models that explicitly capture single-cell resolution cell-cell interaction features to predict patient survival at a population scale are currently missing.

**Methods** We quantified hundreds of single-cell resolution cell-cell interaction features through neighborhood calculation, in addition to cellular phenotypes. We applied these features to a neural-network-based Cox-nnet survival model to identify survival-associated features. We used non-negative matrix factorization (NMF) to identify patient survival subtypes. We identified atypical subpopulations of triple-negative breast cancer (TNBC) patients with moderate prognosis and Luminal A patients with poor prognosis and validated these subpopulations by label transferring using the UNION-COM method.

**Results** The neural-network-based Cox-nnet survival model using all cellular phenotype and cell-cell interaction features is highly predictive of patient survival in the test data (Concordance Index > 0.8). We identify seven survival subtypes using the top survival features, presenting distinct profiles of epithelial, immune, and fibroblast cells and their interactions. We reveal atypical subpopulations of TNBC patients with moderate prognosis (marked by GATA3 over-expression) and Luminal A patients with poor prognosis (marked by KRT6 and ACTA2 over-expression and CDH1 under-expression). These atypical subpopulations are validated in TCGA-BRCA and METABRIC datasets.

**Conclusions** This work provides an approach to bridge single-cell level information toward population-level survival prediction.

### Plain language summary

It may be possible to separate patients with cancer into different groups or subtypes based on the features of their tumor, such as the interactions between different types of cells in the tumor. In this study, we develop a computer-based model to calculate the interactions between cells in breast cancer images. We use these interactions to identify seven subtypes of patients with breast cancer with differences in their survival. We identify some subpopulations of patients with atypical survival outcomes. This work may ultimately help clinicians and researchers to identify patients with breast cancer at increased risk of poorer outcomes and to tailor their treatments accordingly.

<sup>1</sup>Department of Computational Medicine and Bioinformatics, University of Michigan, Michigan, MI 48105, USA. <sup>2</sup>These authors contributed equally: Shashank Yadav, Shu Zhou.  email: [lgarmire@med.umich.edu](mailto:lgarmire@med.umich.edu)

**B**reast cancer became the most commonly diagnosed cancer in 2020, with an estimated 2.3 million new cases globally<sup>1</sup>. Predicting survival in breast cancer can aid clinicians in making prompt prognostic decisions and deciding the direction of treatment. The relevance of prognosis in oncology is projected to grow in the future as new prognostic indicators allow for more precise treatment therapies<sup>2</sup>. However, insufficient knowledge about the intercellular interaction between the tumor and tumor microenvironment is a big roadblock in applying personalized therapies. The breast cancer tumor ecosystems in breast cancer consist of neoplastic epithelial cells forming the tumor core, as the 'tumor microenvironment' is composed of several types of immune cells, fibroblasts, adipocytes, and mesenchymal cells<sup>3,4</sup>. These diverse cell types alter molecular and cellular programs and present dynamic spatial heterogeneity as the disease progresses. Such temporal and spatial changes are responsible for differential responses to anti-cancer therapies and subsequent clinical outcomes. Hence, it is essential to develop a comprehensive understanding of breast cancer heterogeneity by elucidating the contribution of tumor, tumor microenvironment, and their intricate interactions<sup>5,6</sup>.

Investigations at single-cell resolution<sup>7,8</sup> have lately enabled detailed elucidation of tumor-microenvironment interactions. Various components of tumor-immune-stromal relationships have been identified, and tumor heterogeneity was also analyzed to distinguish breast cancer subtypes based on epithelial and immune cell populations<sup>9–11</sup>. Potential biomarkers have also been discovered for personalized cancer immunotherapy through Single-cell RNA sequencing (scRNA-seq)<sup>12</sup>. However, one major limitation of scRNA-seq is the loss of spatial information crucial in understanding tumor heterogeneity *in situ*<sup>13</sup>. To address these issues, spatially resolved assays, such as single-cell transcriptomic and proteomic techniques, have been developed to study the distribution of cells in cancers<sup>14</sup>. Recently, Imaging Mass Cytometry (IMC)-based approaches have quantified tumor heterogeneity with spatial context and identified new breast cancer molecular subtypes in large population cohorts<sup>15,16</sup>. However, most of these studies that detect molecular subtypes are unsupervised, without explicitly fitting phenotypes such as survival in the learning process. Rather, survival is used as a post hoc metric to evaluate the subtypes<sup>17</sup>. Moreover, these unsupervised approaches cannot be directly used to predict uncovered patients, limiting the practical utility of the subtype findings.

In this study, we asked whether quantifying single-cell level cell-cell interactions could provide meaningful insights for prognosis prediction for breast cancer patients. We computed single-cell level imaging features that capture cell-cell interactions in breast cancer tissues from a previous single-cell imaging mass cytometry study on 259 breast cancer patients with survival data<sup>16</sup>. We applied a neural network-based method Cox-nnet previously developed in our group<sup>18</sup> on these features to predict explicitly the patient survival outcome. Using the top survival features, we uncovered seven single-cell interaction-based patient subtypes that show better associations with survival compared to the current mainstream molecular subtypes in breast cancer. We characterized these survival subtypes with distinct profiles of cellular phenotypes as well as cell-cell interactions. Using this survival subtyping classification, we identified two atypical patient subpopulations: a subgroup of triple-negative breast cancer patients with moderate prognosis and a subgroup of Luminal A patients with poor prognosis. We further utilized the transfer learning approach and validated the presence of such unconventional subgroups and their biomarkers in the TCGA and METABRIC breast cancer datasets.

## Methods

**Dataset and extracted features.** The dataset analyzed in the study is obtained from a previously published study containing 259 patients<sup>16</sup>. The dataset contains the IMC data, phenograph neighborhood information, clinical features (Tumor Grade, ER Status, PR Status and HER2 Status), and patient prognosis outcome (overall survival time, disease-free survival time, and alive/dead status). The previous work defined 27 cellular phenotypes that described the histopathological landscape of breast cancer, and direct neighbor cells are defined within 4 pixels (4  $\mu$ m) distance to the target cell. In this study, we used the clinical features, cellular phenotype density (count of cells of each phenotype per unit area), and cellular neighborhood information for analysis, as detailed below:

**Clinical features.** The clinical features set comprises the Tumor Grade, ER Status, PR Status, and HER2 Status for each patient. Tumors in which cells appear highly dissimilar to normal cells tend to proliferate, and the tumor grade is assigned based on the extent of proliferation. In the traditional clinicopathological classification of breast cancer, patients are classified into Luminal A, Luminal B, Triple-Negative, and HER2-Enriched classes. This classification is based on patients' clinical features such as ER (positive/negative), PR (positive/negative), and HER2 (positive/negative) status<sup>19,20</sup>.

**Cellular phenotypic features.** For each patient, the cell phenotype density is quantified as the counts of each cellular phenotype per unit area in the IMC image of the tumor tissue. Out of the original 27 cellular phenotypes, there are six immune, seven stromal, and fourteen epithelial cellular phenotypes. The six immune phenotypes are B Cell, T and B Cell, T-Cell<sub>1</sub>, Macrophage<sub>1</sub>, T-Cell<sub>2</sub>, Macrophage<sub>2</sub>. The seven stromal phenotypes are Endothelial, Vimentin<sup>hi</sup> Fibroblasts, Small circular Fibroblasts, Small elongated Fibroblasts, Fibronectin<sup>hi</sup> Fibroblasts, Large Elongated Fibroblasts, SMA<sup>hi</sup>-Vimentin<sup>hi</sup> Fibroblasts. The 14 epithelial cellular phenotypes are Hypoxic Epithelial Cells, Apoptotic Epithelial Cells, Proliferative Epithelial Cells, p53<sup>+</sup> EGFR<sup>+</sup> Epithelial Cells, Basal CK Epithelial Cells, CK7<sup>+</sup>CK<sup>hi</sup>Cadherin<sup>hi</sup> Epithelial Cells, CK7<sup>+</sup>CK<sup>+</sup> Epithelial Cells, Epithelial<sup>low</sup> Epithelial Cells, CK<sup>low</sup>HR<sup>low</sup> Epithelial Cells, CK<sup>+</sup>HR<sup>hi</sup> Epithelial Cells, CK<sup>+</sup>HR<sup>+</sup> Epithelial Cells, CK<sup>+</sup>HR<sup>low</sup> Epithelial Cells, CK<sup>low</sup>HR<sup>hi</sup>p53<sup>+</sup> Epithelial Cells, Myoepithelial Cells.

**Cell-cell interaction features.** The phenotypic features in the original report are limited, as they do not assess the interactions between cellular phenotypes and between tumor-tumor microenvironments, which are important parts of tissue heterogeneity. We utilized the available phenograph<sup>21</sup> neighborhood-information data from each patient and quantified the binary interactions between cellular phenotypes. The phenograph neighborhood of the IMC image is described as a numerous cellular network spread out in the mass cytometry image, where individual cells are represented as nodes of the cellular network. Starting with 27 cellular phenotypes, we calculated 378 pairwise phenotype-phenotype features.

Cell-cell interactions are restricted to those within the same cellular community. We iterate through each cell of a cellular community in the image, enumerate the binary interactions the cell makes with its neighbors, and multiply the sum of the interactions by the clustering coefficient  $C$  of the cellular community as the weight. Then, we repeat this process for all the cellular communities across the mass cytometry image, sum the 378 interactions, and divide them by the area occupied by all

the cells. We define the result of this procedure as the ‘Cell-Cell Interaction Score’ (CCIS).

$$CCIS_{\forall Pair(x,y), x \in P, y \in P} = \frac{\sum_{\forall N} (C * B_{xy})}{A} \quad (1)$$

where  $C$  is the clustering coefficient for each cluster,  $B_{xy}$  is the binary interaction counts between a particular pair of phenotypes  $x$  and  $y$ ,  $A$  is the total area occupied by all the cells in a mass cytometry image,  $P$  is the set of all phenotypes and  $N$  is the number of cellular communities in an image.

At a more detailed level, these 378 can be further categorized by immune/stromal/epithelial cell types in the interacting pairs and classified as immune-immune (21 features), immune-stromal (42 features), immune-epithelial (84 features), stromal-stromal (28 features), stromal-epithelial (98 features), epithelial-epithelial (105 features) interactions. Among the feature pairs, 1 immune-stromal feature (macrophage interacting with Large Elongated Fibroblasts), 2 immune-epithelial features (macrophage interacting with Hypoxic Epithelial Cells and macrophage interacting with  $CK7^+CK^{hi}$ Cadherin $^{hi}$  Epithelial Cells), 1 stromal-stromal feature (Large Elongated Fibroblasts self-interaction) and 1 stromal-epithelial (Large Elongated Fibroblasts interacting with Myoepithelial cells) are dropped since these pairwise phenotype-phenotype interactions calculations returned 0 (lack of neighbors within the immediate vicinity), hence the final number of features would be 373 features. We plot the feature heatmaps using the ‘heatmap’ package in R.

**Survival modeling.** We use a variety of neural-network-based Cox-nnet models as the one-stage Cox-nnet-v2 models for the phenotypic feature set, tumor-microenvironment feature set, tumor-core feature set, and pairwise combinations. Further, we develop a two-stage Cox-nnet model by combining the hidden layer output of previously trained one-stage models. For each one-stage Cox-nnet-v2 model, the hidden layer nodes are equal to the square root (rounded up) of input nodes. The Cellular Phenotypic (CP), tumor-microenvironment interaction (TMI), and tumor-core interaction (TCI) feature-based models have 27, 268, and 105 input features, respectively, and hence we obtain 6, 17, and 11 nodes in the respective hidden layer. The two-stage Cox-nnet model combines the hidden layer output features. As a result, 34 input features are used in the second stage. For comparison, we use the Cox-PH model on clinical features as the baseline model.

**Survival feature ranking.** We calculate the feature importance of the 34 input features using the ‘Variable Importance’ function of the Cox-nnet-v2 model. Out of the 34 input features in the two-stage Cox-nnet model, we calculate the importance scores, separate them into the original three sets of 6, 17, and 11 features, respectively, and associate them with the one-stage Cox-nnet-v2 models. For each model, we took a dot product of the hidden layer feature-importance vector with the model weights to get the importance scores vector of the original features.

$$\hat{O} = W^T \cdot \hat{H}, \quad (2)$$

where  $\hat{O}$  is the importance vector of the original features (shape1).  $W$  is the model weight matrix (shape  $h \times p$ ).  $H$  is the importance vector for hidden layer (shape  $h \times 1$ ).

**Unsupervised analysis for patient subtype detection.** We normalize the feature importance values between 0 and 1 and select the top features from each of the three feature sets. For the phenotypic set, we set the threshold as 0.5 and selected all the features with a higher importance score. For the

tumor-microenvironment and tumor-core feature sets, we select features with an importance score greater than 0.75. In total, out of the 40 features we choose 50 features and perform NMF-based consensus clustering using the NMF R package v0.23.0. This technique has been used in molecular subtype detection in cancer<sup>22,23</sup>. We carry out a hyperparameter search for the NMF rank and varied it from 3 to 15, finding that the maximum values of ‘cophenetic scores’ and ‘silhouette coefficients’ are reached at an NMF rank of seven, and hence we choose the optimum NMF rank as seven for subsequent analysis.

**Feature correlation analysis and comparison between NMF-defined and clinically defined subtypes.** We calculate the correlations between the features associated with each NMF-defined subtype using Spearman’s Correlation Coefficient in R and plot the Circos plot of the correlation using the circlize package in R<sup>24</sup>. How the NMF-defined classes intersect with the clinicopathological classification is determined by a Sankey Plot. We plot the Sankey plot using the plotly package in Python.

**Class label transfer and comparison with TCGA-BRCA and METABRIC datasets.** We verify our results for ‘TNBC-Good Survival’ and ‘Luminal A-Poor Survival’ patient subpopulations in two external datasets, TCGA-BRCA and METABRIC<sup>25,26</sup>. Here, we utilize the mass cytometry counts for 30 protein-based biomarkers and calculate the average expression of each biomarker over all the cells present in the IMC image. We treat it as pseudo-bulk protein expression and a proxy for bulk mRNA expression to facilitate the comparison with bulk mRNA expression-based external datasets. Then, we repurpose the UNION-COM algorithm<sup>27</sup> to perform patient matching between our dataset, the TCGA-BRCA dataset, and the METABRIC dataset. The UNION-COM method was originally developed to perform topological alignment and label transfer on single-cell multi-omics datasets. It takes two expression matrices as input, calculates the joint embedding between the two datasets, and maps samples from one dataset to another. In the TCGA-BRCA dataset, we find 28 genes corresponding to 28 protein-based biomarkers out of 30 (except *TWIST1* and *mTOR*). In the METABRIC dataset, we find all 30 genes corresponding to the 30 protein-based biomarkers. For each TNBC and Luminal A subpopulation, the pseudo-bulk protein expression matrix from the single-cell dataset is used as input, and the mRNA expression matrix is applied as the external dataset in the UNION-COM method. We run the UNION-COM method using default parameters, and it returns the pairwise distance between patients between our dataset and the external dataset. We set a 99.5% similarity cutoff and select patients matching our ‘TNBC-Good Survival’ and ‘Luminal A-Poor Survival’ patient subpopulations, respectively. Survival plots are made using the lifelines Python package, and differential gene expressions are done using the limma package in R<sup>28</sup>.

**Cross-check the TNBC survival subtypes with the molecular subtypes of TNBC.** We cross-compared the seven subtypes in the previous study<sup>29</sup> with the two survival subtypes we identified. We used the six subtype markers (*MKI67*, *MYC*, *EGFR*, *CDH1*, *SNAI2* and *TWIST1*) in that study that overlap with the 30 protein-based biomarkers in the single-cell imaging dataset here. The six markers are *MKI67* and *MYC* as the proliferation markers, *EGFR* as the growth factor/Myoepithelial markers, and *SNAI2* and *TWIST1* as the epithelial-mesenchymal transition-associated (EMT-associated) markers. The raw data from the previous study are not available; we relied on the pixels in this heatmap as the normalized gene expression values for the

comparative study. We first transformed the heatmap of Supplementary Fig. 6<sup>29</sup> of the six markers from RGB (red, green and blue) to HSV(Hue, Saturation and Value) scale and filtered out all the red and green-related colors by setting the HSV thresholds. Then we transformed the HSV scale to the gray scale and mapped the intensity to log-transformed gene values according to the figure legend. Specifically, we map green-related colors from  $-2$  to 0 and red-related colors from 0 to 2. We plotted the level of the six markers from the imaging data alongside the heatmap from the previous study. We performed hierarchical clustering with the average expression level of markers from the two survival subtypes defined by us and the TNBC molecular subtypes obtained from the previous study.

## Results

**Feature engineering from the breast cancer single-cell images.** The breast cancer cohort contains 259 patients with survival data, as reported earlier<sup>16</sup>. The summary of the cohort's patients, including clinical variables such as tumor grades, clinical features (ER, PR, HER2), and clinicopathological classes, is shown in Supplementary Table 1. We first extracted 27 pre-defined cellular-phenotype (CP) features based on image mass cytometry data (Fig. 1a, Supplementary Data 1). These features describe epithelial, immune, and stromal cell phenotypes at the single-cell resolution in the original study ("Methods"). We next used networks of cells to represent the cells and cellular communities with the spatial arrangement as shown in the imaging data. We then calculated phenograph neighborhood information based on the cellular phenotypes in the tissue ("Methods"). This process results in an additional 378 cell-cell interaction features that can be broadly divided into two subsets. The first subset contains 273 features related to the tumor-microenvironment interaction (TMI) features (Fig. 1a). The TMI features are computed from pairwise interactions among the three types of cells: immune, stromal, and epithelial cells. Specifically, they represent immune-immune (21 features), immune-stromal (42 features), immune-epithelial (84 features), stromal-stromal (28 features), and stromal-epithelial (98 features) interactions. The second subset contains 105 tumor-core interaction (TCI) features, which are exclusively epithelial-epithelial interactions between epithelial cells of various cellular phenotypes (Fig. 1a). The distributions of these 378 features in patients together with clinically defined breast cancer subtypes are illustrated in Fig. 1b. Anchoring on the TCI features, hierarchical clustering demonstrates broad but distinct cellular heterogeneity among patients, far more complex than that defined by the clinical subtypes. Compared to TCI features, TMI features describing immune-epithelial and stromal-epithelial interactions have similar but less distinct heterogeneity. On the other hand, all CP features show far less global correlations with those TCI patterns presented in epithelial-epithelial interactions.

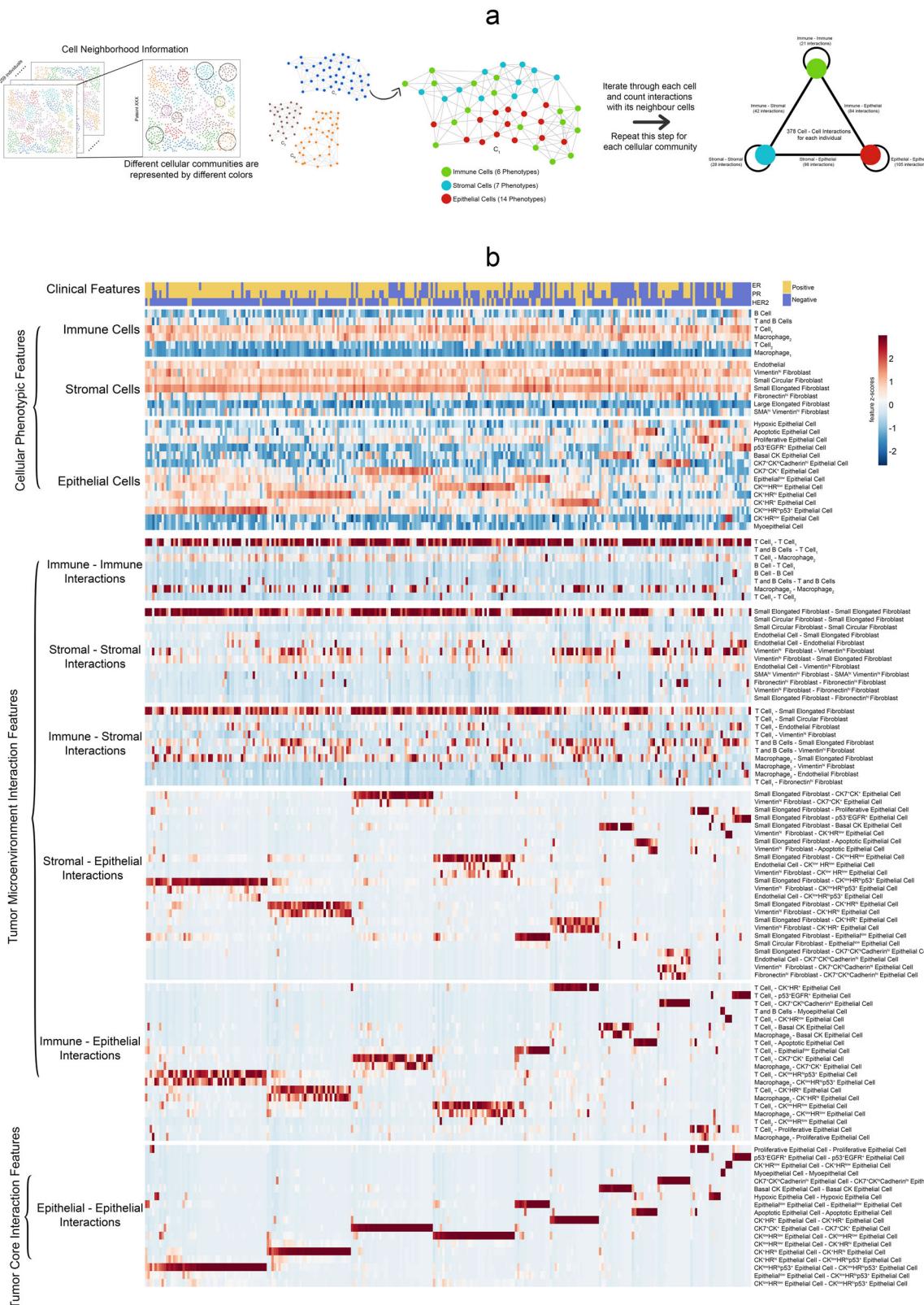
**Survival prediction using single-cell phenotype features.** A major goal of our study is to identify single-cell level features associated with patient survival and further evaluate the relative contributions of these features toward patient survival. To this end, we used the recently developed Cox-nnet neural-network-based survival prediction models from our group (Ching et al.<sup>18</sup>), which had shown advantages in single or multiple data modalities<sup>30,31</sup>, compared to the conventional Cox-PH method<sup>32</sup>. Here as a comparison to Cox-nnet models, we built a Cox-PH model using the clinical information including ER status, PR status, HER2 status, and tumor grade and TNM (tumor-node-metastasis) stagings (Set I) as the baseline model. We constructed a series of two-stage Cox-nnet models (Fig. 2a) on the

combinations (Set II, III, IV, and V) of CP features, TMI features and TCI features (Set II, III, IV, and V). Two-stage Cox-nnet models are complex models that use simpler Cox-nnet models as individual building blocks<sup>30</sup> to integrate different feature sets or data modalities. In the first stage of training, each Cox-nnet model is built to fit a specific set of data (CP, TMI or TCI features) to predict survival. The hidden nodes in the first-stage Cox-nnet model are then combined as the input features to train a second-stage Cox-nnet model. For each Cox-nnet model, we used L2 regularization to reduce overfitting.

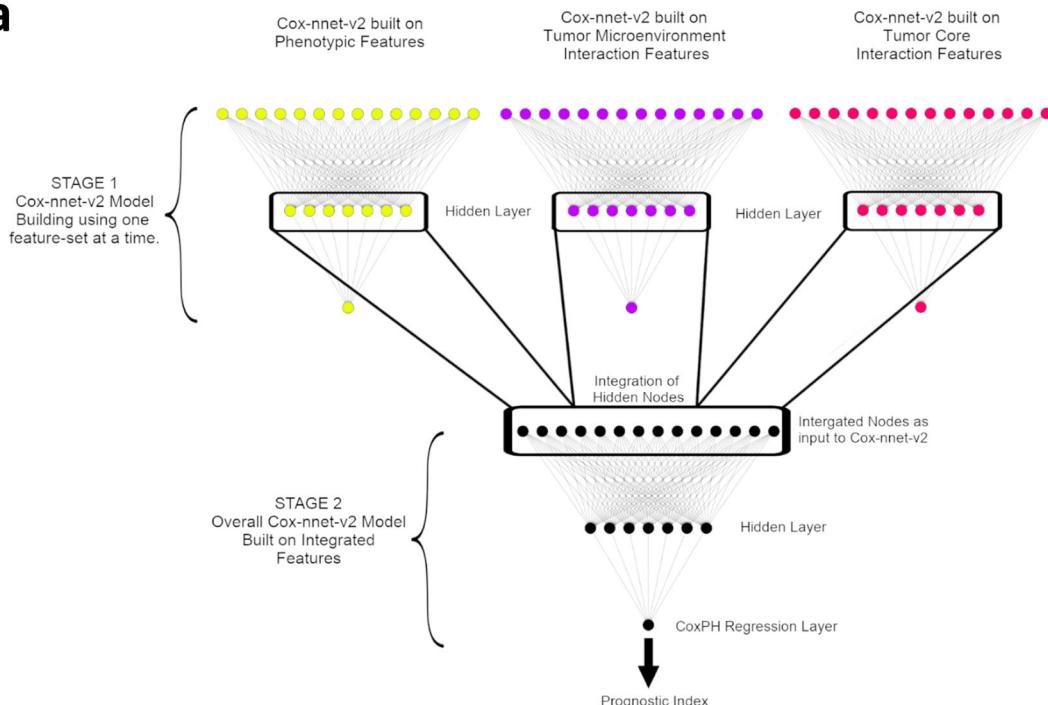
To rank the relative contributions of different features, we estimated the relative importance of the features in each of the CP, TMI, and TCI feature sets (Supplementary Data 2–4). Among CP features, immune cells, including type 2 macrophages (CD68+/vimentin low) and T & B immune cell clusters, have the highest relative importance scores (1.000 and 0.692 respectively) compared to those (0.507–0.631) of different subtypes of epithelial cells (Supplementary Data 2), highlighting their importance in patient prognosis. In TMI features, two types of immune-epithelial interactions (T and B cells-CK<sup>low</sup>HR<sup>low</sup> epithelial cells, Macrophage<sub>1</sub>-CK<sup>low</sup>HR<sup>hi</sup>p53<sup>+</sup>) and a type of fibroblast-epithelial interaction (small elongated fibroblast-CK<sup>7</sup>+CK<sup>+</sup> epithelial cells) and are top 3 dominant features (Supplementary Data 3). Among the TCI features, the interactions among most proliferative epithelial cells are the strongest, as expected (Supplementary Data 4).

**Identifying survival subtypes among breast cancer patients.** Our next goal was to identify patient subpopulations associated with survival, using the top survival features selected by importance scores from the best model using Set XII (the combination of CP, TMI, and TCI features). We performed Non-negative Matrix Factorization (NMF) based consensus clustering on the top 50 features according to the importance scores ("Methods"). As a result, we identified seven optimal patient subpopulation clusters, indexed by 1–7 from the best to the worst survival risks (Fig. 3a). We confirmed that seven clusters are the optimum value, based on Cophenetic score and Silhouette coefficient, two metrics of clustering accuracy (Fig. 3b). The Kaplan-Meier plot for the overall survival of the seven survival subtypes yields a much higher C-index of 0.80 and a more significant log-rank *p*-value  $p = 8.6e^{-0.6}$  (Fig. 3c) for  $n = 259$  independent patients, compared to the C-index of 0.63 and the log-rank *p*-value of 0.001 from the stratification of the four molecular subtypes (Fig. 3d). These results demonstrate that the single-cell level CP, TMI, and TCI features yield more informative subtypes that better reveal the heterogeneity in survivorship among patients.

We further describe the survival subtypes based on their top features and their relationships with other clinical information (Fig. 3e, Supplementary Table 2). As expected, the better survival subtypes tend to have more ER+ and PR+ cases, and the worst survival subtypes tend to have higher tumor grade (Grade III) and HER2- cases. The best survival subtype 1 is enriched with a subtype of epithelial cells with high levels of cytokeratin (CK) 7 and 19 but low hormone receptors (CK7, CK19, low HR). It also has high levels of interaction scores between this epithelial cell subtype and several subtypes of fibroblast cells, highlighting the importance of the interaction between these two cell types for patient outcomes. The next best subtype 2 also has enriched scores on a subtype of epithelial cells that express pan-cytokeratins and hormone receptors (pan-CK, HR), as well as their interactions with certain fibroblast cells. On the other hand, the subtype 7 with the worst survival is characterized by high degrees of interactions among proliferative epithelial cells, as well as interactions between macrophages/T-cells expressing high

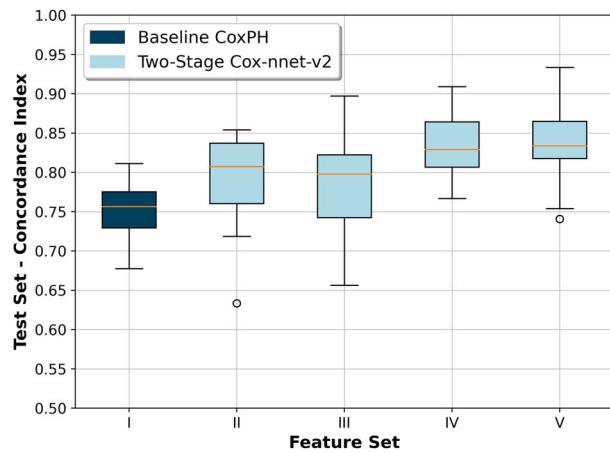


**Fig. 1 Overview of different feature sets in the data. a** Methodology for extracting the cell-cell interaction features. We utilize the cellular neighborhood information obtained from the phenograph. The phenograph result contains different locally connected cellular communities. For each cellular community (the neighborhood graph), we iterate through each cell (node) and count the number of interactions between the particular cell and its neighbors. We repeat this process for all the different cellular communities to assess the 378 different pairwise cell-cell interactions. **b** Heatmaps illustrate the different feature types for all the patients. The patients are arranged in the order defined by the hierarchical clustering based on the epithelial-epithelial interactions.

**a****b**

Legend	
Set	Feature Set
I	Clinical
II	Phenotypic* + TMI*
III	Phenotypic* + TCI*
IV	TMI* + TCI*
V	Phenotypic* + TMI* + TCI*

\*Denotes hidden layer features obtained from cox-nnet-v2



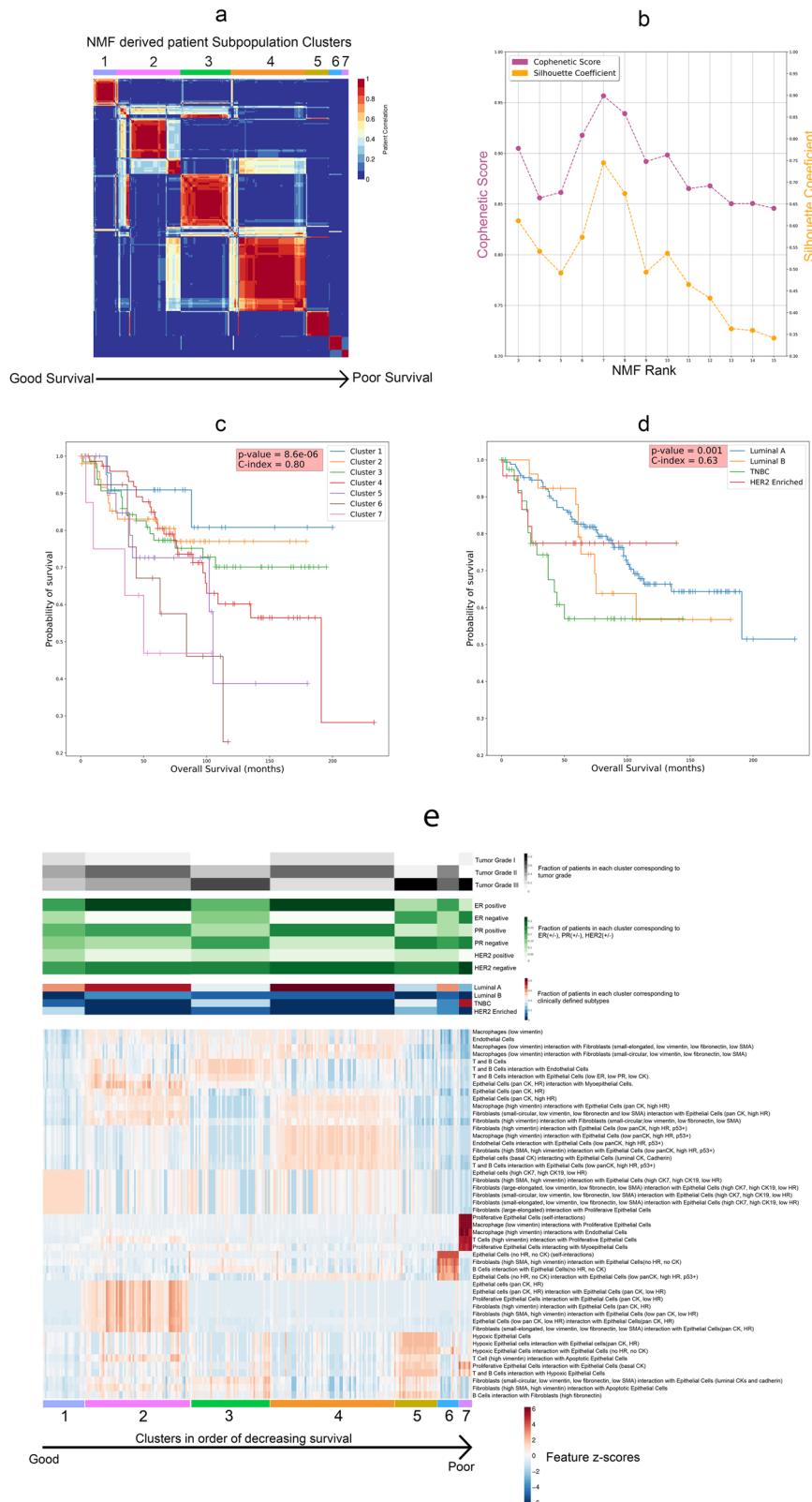
**Fig. 2 Two-stage Cox-nnet model comparison with baseline and Cox-nnet-v2. a** Model architecture for the two-stage Cox-nnet model. Three individual Cox-nnet models were built for each data type. The hidden nodes from the first-stage Cox-nnet models were combined to form the input to construct a Cox-nnet model in the second stage. **b** Comparison of Concordance index across different Feature Set-Model pairs through boxplot with  $n = 20$  independent experiments. The box displays the quartiles of the dataset, and the whiskers show the rest of the distribution, except for points that are determined to be outliers, using a method that is a function of the inter-quartile range.

vimentin and the proliferative epithelial cells. The second worst survival subtype 6 has a high level of epithelial cells lacking CK and HR expression, as well as strong interactions between fibroblast/B cells and these epithelial cells. In summary, the survival subtypes show distinct profiles of epithelial cell subtypes and interactions between the epithelial cells and adjacent immune and fibroblast cell subtypes.

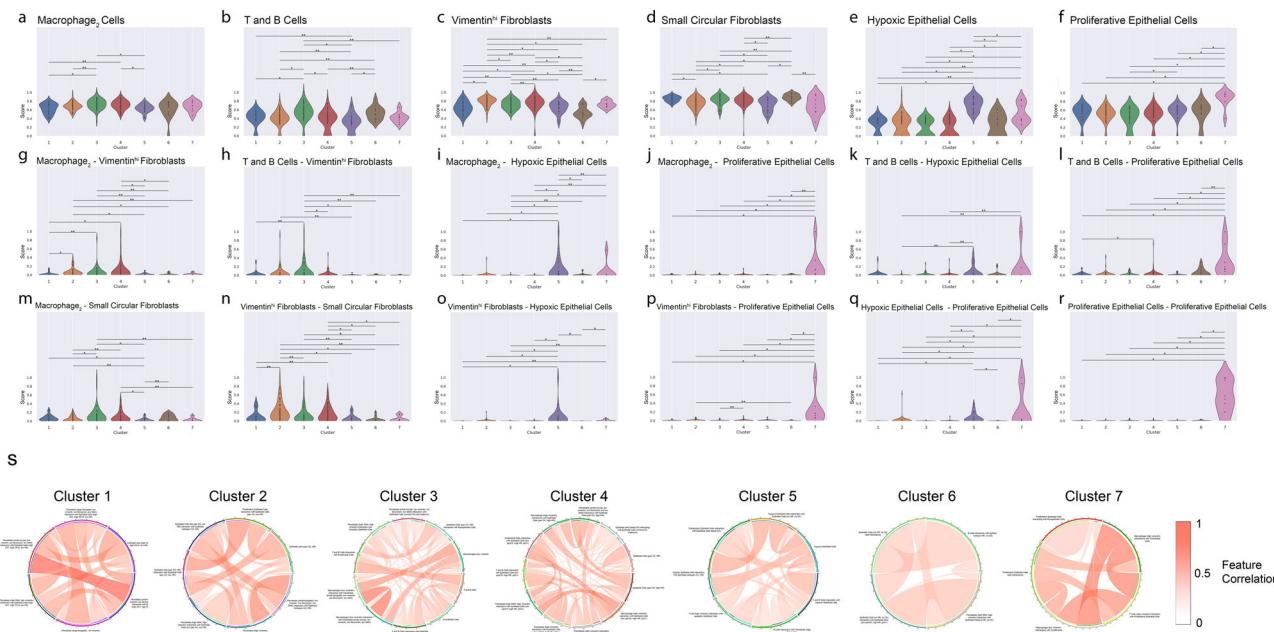
We also compared the seven survival clusters with the 18 single-cell pathology (SCP) subgroups based on the 27 cellular phenotype features, as identified from the original study<sup>16</sup>. As shown in Supplementary Fig. 1, cluster 1 has a strong association (90%) with SCP\_11 (CK7 and pan-CK) subgroup. Cluster 5 is enriched with SCP\_17 (Hypoxic) subgroup. Cluster 6 is equally distributed in SCP\_7 (Epithelial\_low) and SCP\_10 (Epithelial\_low mixed) subgroups. Cluster 7 shows a connection with the SCP\_14

(Proliferative) subgroup. Cluster 2, Cluster 3 and Cluster 4 do not show a clear association with some particular SCP subgroups.

**Characterization of the seven survival subtypes.** We next directly compared the enrichment of different types of immune, fibroblast and epithelial cells, as well as their interaction scores in the seven survival subtypes (Fig. 4). Subtypes 5–7 show quite distinct patterns from subtypes 1–4. In particular, subtype 5 has a high level of hypoxic epithelial cells (Fig. 4e), as well as high levels of interactions between these hypoxic epithelial cells and macrophage<sub>2</sub>, T/B cells and vimentin-expressing fibroblasts (Fig. 4i, k and o). Subtype 6 has the second highest level of proliferative epithelial cells (Fig. 4f) but the lowest levels of vimentin expressing fibroblast cells (Fig. 4c), and accordingly, the second highest



**Fig. 3 Non-negative Matrix Factorization (NMF)-based subpopulation detection associated with survival.** **a** NMF heatmap for 259 patients in our cohort illustrating the seven subpopulations arranged in order of decreasing survival. **b** Cophenetic Score and Silhouette Coefficient versus the NMF Rank. **c** Kaplan-Meier (KM) plots illustrate the Overall Survival for patients in the seven NMF-derived subpopulation clusters. **d** Kaplan-Meier plots illustrate the Overall Survival for patients based on clinicopathological classification. **e** Heatmap of top-ranked features from best performing two-stage Cox-nnet model, in associations with tumor grade, clinical features and molecular subtypes.



**Fig. 4 Characterization of the single image-based survival subtypes.** **a–r** Scoring and profiling for the seven survival subtypes based on various cellular phenotypes and cell-cell interaction features. Cell counts and interactions were normalized between 0 and 1 to make comparison possible on the same scale. **a** Macrophage<sub>2</sub> Cells, **b** T and B Cells, **c** Vimentin<sup>hi</sup> Fibroblasts, **d** Small Circular Fibroblasts, **e** Hypoxic Epithelial Cells, **f** Proliferative Epithelial Cells, **g** Macrophage<sub>2</sub>-Vimentin<sup>hi</sup> Fibroblasts, **h** T and B Cells-Vimentin<sup>hi</sup> Fibroblasts, **i** Macrophage<sub>2</sub>-Hypoxic Epithelial Cells, **j** Macrophage<sub>2</sub>-Proliferative Epithelial Cells, **k** T and B Cells-Hypoxic Epithelial Cells, **l** T and B Cells-Proliferative Epithelial Cells, **m** Macrophage<sub>2</sub>-Circular Fibroblasts, **n** Vimentin<sup>hi</sup> Fibroblasts-Small Circular Fibroblasts, **o** Vimentin<sup>hi</sup> Fibroblasts-Hypoxic Epithelial Cells, **p** Vimentin<sup>hi</sup> Fibroblasts-Proliferative Epithelial Cells, **q** Hypoxic Epithelial Cells-Proliferative Epithelial Cells, **r** Proliferative Epithelial Cells-Proliferative Epithelial Cells. Statistical testing was done for  $n = 259$  patients using the Mann-Whitney U Test with Benjamini-Hochberg-based false discovery rate (FDR) adjustment. The significant pairs are marked as follows: \*:  $p$ -value  $< 0.01$ . \*\*:  $0.01 < p$ -values  $< 0.05$ . Each violin is drawn using the kernel density estimate of the underlying distribution. **s** Circos plots demonstrate the correlation between feature pairs associated with each subpopulation.

scores in interactions between T/B cells and proliferative epithelial cells (Fig. 4l). Subtype 7 stands out with the highest level of proliferative epithelial cells (Fig. 4f) and the second highest level of hypoxic epithelial cells (Fig. 4e). As a result, subtype 7 has the highest scores in the greatest number of TMI-tumor interaction categories (Fig. 4j–l, p–r). In summary, the subtypes 5–7 with some of the worst survival outcomes are highly enriched with hypoxic (subtype 5) or proliferative epithelial cells (subtype 6 and 7), as well as the corresponding interactions between immune cells and these epithelial cells; however, they have much lower interactions between fibroblast and immune cells in the tumor microenvironment (Fig. 4g, h, m, n).

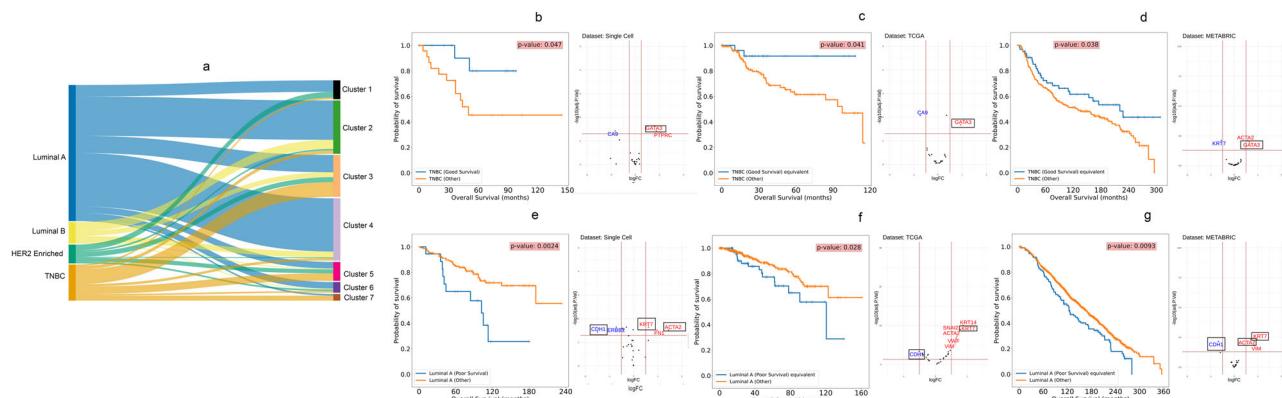
Next, we investigated the higher-order of correlation relationship among the cell-cell interactions for each survival subtype. We calculated the correlations among all the CP, TMI and TCI features and studied the pairs of cell-cell interaction with correlations greater than 0.5 (Fig. 4s). For subtype 1, a high correlation exists between epithelial cell-small, elongated fibroblast interaction and epithelial cell-small circular fibroblast interaction, indicating that they share similar modes of interactions. Similarly, subtype 3 shows a high correlation between macrophage-small elongated fibroblast interaction and macrophage-small circular fibroblast interaction. In subtype 7, proliferative epithelial-macrophage/T-Cell (vimentin-expressing) interaction is highly correlated with interactions between vimentin-expressing macrophage and endothelial cells, further confirming the detrimental and synergistic effect of certain immune cells on the proliferative tumor.

#### Discovery of subpopulations of TNBC and luminal A subtypes with atypical survival outcomes.

To uncover new insights into

this survival subtyping, we compared the classification results using molecular subtyping vs. the survival subtyping approach (Fig. 5a). Each molecular subtype is split into multiple survival subtypes, demonstrating the widespread heterogeneity in terms of survival. We dichotomized the 7 survival subgroups into good survival (subgroups 1–4) and bad survival (subgroups 5–7) groups and noticed that for luminal A and TNBC cancers, these subgroups show significantly different survival outcomes (log-ranked  $p$ -value  $< 0.05$  for  $n = 44$  independent TNBC patients and  $n = 166$  independent luminal A patients) from their counterparts (Fig. 5b, e). Despite being a molecular subtype regarded as having the worst prognosis, TNBC actually has a large proportion (61.36%) of atypical good survival patients from clusters 1–4, where cluster 3 is the major cluster among the four (66.66%). On the other hand, while the luminal A subtype generally has the best prognosis among the four molecular subtypes, around 11% of the luminal A breast cancers still belong to the relatively bad survival group. We next investigated the signatures of these atypical subpopulations by differential expression analysis. The atypical good survival subpopulation in TNBC has over-expression of *GATA3* (Fig. 5b, Supplementary Data 5). On the other hand, the atypical bad survival subpopulation in the luminal A subtype has an over-expression of *KRT7* and *ACTA2* and an under-expression of *CDH1* (Fig. 5e, Supplementary Data 6).

We next tested if such atypical subpopulations can be validated in general. We used the UNION-COM (Cao, Bai, Hong, & Wan, 2020) method to perform label-transfer learning based on the pseudo-bulk protein expression in this single-cell dataset and applied the model to the transcripts of the same genes in TCGA-BRCA and METABRIC data. The results show that TCGA-BRCA and METABRIC datasets indeed have such good survival subpopulations in TNBC (Fig. 5c, d)



**Fig. 5 Discovery and validation of atypical subpopulations in triple-negative breast cancer (TNBC) and luminal A patients.** **a** Sankey plot showing the distribution of patients in the seven subpopulations vs. clinical subtypes. **b** Kaplan-Meier (KM) survival plot for the two TNBC subpopulations in the single-cell dataset and the differentially expressed protein biomarkers for the two subpopulations. **c** Validation of the two TNBC subpopulations in TCGA breast cancer data corresponding to those in **(b)** by KM plot and differentially expressed genes. **d** Validation of the two TNBC subpopulations in METABRIC breast cancer data corresponding to those in **(b)** by KM plot and differentially expressed genes. **e** KM survival plot for the two luminal A subpopulations in the single-cell dataset and the differentially expressed protein biomarkers for the two subpopulations. **f** Validation of the two luminal A subpopulations in TCGA breast cancer data corresponding to those in **(e)** by KM plot and differentially expressed genes. **g** Validation of the two luminal A subpopulations in METABRIC breast cancer data corresponding to those in **(e)** by KM plot and differentially expressed genes. (For each subplot, the blue curve represents the subpopulation in context).

and bad survival subpopulations in luminal A patients (Fig. 5f, g). The atypical subpopulations in both validation cohorts have significantly different survival curves (log-rank  $p$ -value  $< 0.05$  for  $n = 159$  independent TCGA-BRCA TNBC samples,  $n = 599$  independent TCGA-BRCA luminal A patients,  $n = 299$  independent Metabric TNBC patients and  $n = 1369$  independent Metabric luminal A patients) compared to their counterparts. Moreover, *GATA3* is also consistently overexpressed in the atypical good survival subpopulations in both the TCGA-BRCA and METABRIC datasets (Fig. 5c, d, Supplementary Data 7 and 8). Additionally, *CA9* shows a common under-expression in the TCGA-BRCA dataset (Fig. 5b, c). Similarly, the matching poor survival luminal A subpopulations in TCGA-BRCA and METABRIC datasets show the same patterns of over-expression of *KRT7* and *ACTA2* and under-expression of *CDH1* (Fig. 5f, g, Supplementary Data 9 and 10). Loss of *CDH1* is a characteristic of invasive lobular breast cancer<sup>33</sup>. Confirming this, we found that lobular type increases from 20.72 to 50.7%, comparing the good vs. bad survival subgroups in TCGA luminal A breast cancers. It also increases from 8.34 to 12.7% for the good vs bad survival subgroups in the luminal A cancers in the METABRIC dataset.

TNBC was previously reported to have six molecular subtypes<sup>29</sup>, including 2 basal-like (BL1 and BL2), an immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem-like (MSL), and a luminal androgen receptor (LAR) subtype, as well as an extra unstable (UNS) subtype. We further cross-checked the survival-based TNBC subtypes here with those reported earlier (Supplementary Fig. 2). We selected six markers (*MKI67*, *MYC*, *EGFR*, *CDH1*, *SNAI2*, *TWIST1*) that are in common with the Supplementary Fig. 6 of the original study and our single-cell imaging dataset and relied on the pixels in this heatmap as the normalized gene expression values for the comparative study. We scaled the marker gene intensities from the single-cell imaging data by the RGB values of the pixels. The good TNBC survival group of the image data shows a lower expression in the proliferation markers and myoepithelial markers and a mixed expression in EMT markers (Supplementary Fig. 2b). And the bad TNBC survival group of the image data shows a combination of low and high expressions in Proliferation markers and EMT markers (Supplementary Fig. 2c). From the clustering result of averaged gene expression in two types of subtype systems, the

good TNBC survival subgroup is the closest to MSL and LAR subgroups, whereas the bad TNBC survival group is closest to BL2 and M subgroups (Supplementary Fig. 2d).

## Discussion

Breast cancer is a highly heterogeneous disease, and molecular subtypes based on ER, PR and HER2 statuses are currently the mainstream classification system. In this study, we leverage the strengths of detailed single-cell pathology images and neural-network-based prognosis modeling and define a class of survival-related subtypes for breast cancer. We argue that cellular phenotypes and their interactions provide additional valuable information to predict survival, leading to improved clinical impacts.

The uniqueness of this study lies in several aspects. From the analytical aspect, we explicitly computed hundreds of features describing cell-cell interactions (TMI and TCI) on single-cell imaging data. We used these features as inputs for a neural-network-based survival prediction method called 2-stage Cox-nnet, which highly accurately fits patient survival and is much better than clinical data-based Cox-PH regression. From the biomedical aspect, we have defined seven survival subtypes with distinct profiles of epithelial, immune, and fibroblast cells, as well as their interaction patterns. These survival subtypes amend the molecular subtypes originally based on tumor cell signatures, as each of the molecular subtypes is highly heterogeneous in terms of the patient's prognosis. It also extends beyond previous work that classified breast cancers by immuno-subtypes<sup>34</sup> by considering additionally the interactions among tumor, immune cells, etc.

These survival subpopulations are well characterized by protein markers and cellular phenotypes. The relatively good survival subtypes 1–4 are enriched with CK or HR expression epithelial cells as well as interactions between fibroblasts and other immune cell types, which are lacking in the poorer survival subtypes 5–7. The best survival subtype 1 has a high expression level of luminal CK. Corresponding to this observation, CKs were reported to be associated with better overall survival in breast cancer before<sup>35</sup>. Subtype 2, which also has good survival, has high levels of CK8/18 and HR, which was also reported to be associated with good overall survival<sup>36</sup>. Subtype 5 has a dominant hypoxia phenotype,

demonstrated by the high presence of hypoxia epithelial cells. Hypoxia is associated with resistance against therapies and poor outcomes<sup>37,38</sup>. Despite most of the patients being luminal A subtype, patients in this subgroup show the lowest level of Vimentin<sup>hi</sup> Fibroblasts cells (high vimentin expression, low smooth muscle actin (SMA) expression, and low fibronectin expression)<sup>39</sup>. It also has the 2nd highest level of proliferative epithelial cells, both of which may contribute to poor survival. Subtype 7 has the highest level of proliferative epithelial cells marked by *KI-67* expression, associated with poor survival<sup>40,41</sup>. It also has some of the strongest interactions with cells in EMT, suggesting the highest degree of tumor infiltration by immune cells<sup>42</sup>. The highly active TMI observed in hypoxic subtype 5 and proliferative subtype 7 are quite interesting. Subtype 5 has lower proportions of immune cells (e.g., macrophages) but the strongest interactions between macrophages and the hypoxic epithelial cells, the signature epithelial cells of this subtype. It also has the highest interaction between Vimentin<sup>hi</sup> fibroblasts and hypoxic epithelial cells. The apparent paradox between low macrophage concentration vs high interaction with epithelial cells may be explained by the inflammatory cytokines produced by macrophages at the hypoxic site, which decreases tumoricidal activity<sup>43</sup>. Subtype 7 has the highest proportion of proliferative epithelial cells, and 75% of the patients are TNBC. This subtype has the strongest interactions between various cells in the tumor microenvironment (vimentin expression T/B cells, fibroblasts, and macrophages<sup>2</sup>) and proliferative epithelial cells, as well as interactions between vimentin-expressing macrophages and hypoxic cells. A high degree of immune infiltration was reported in TNBC patients with exhausted T Cell populations<sup>44</sup>, which is consistent with our results. This subtype also shows the highest interaction between hypoxic epithelial cells and proliferative epithelial cells. The hypoxic condition may be a result of the increased aggressiveness of proliferating epithelial cells<sup>45</sup>, consistent with prior observation of a close relationship between proliferation-hypoxia in other solid cancers such as squamous cell carcinoma<sup>46</sup>.

The subtyping system enabled the discovery of atypical survival subpopulations. A good survival subpopulation among TNBC patients (mostly from survival subtype 3) is identified with over-expression of *GATA3*; and a small population of bad survival subtype among luminal A patients are found, with over-expression of *KRT7* and *ACTA2* but under-expression of *CDH1*. These results are robustly verified in TCGA-BRCA and METABRIC datasets. *GATA3* is a zinc finger transcription factor involved in cell type differentiation and proliferation. Corresponding to our observations, higher levels of *GATA3* were also shown to be correlated with better survival in breast cancer patients<sup>47</sup>. *KRT7* encodes cytokeratin-7 and *ACTA2* encodes smooth muscle alpha-2 actin, both are cytoskeleton components. *KRT7* plays a role in cell migration and epithelial-mesenchymal transition (EMT) pathways and is associated with poor survival subpopulations for ovarian cancer<sup>48,49</sup>. Similarly, *ACTA2* was reported to be a marker for poor prognosis in lung cancer<sup>50</sup>. *CDH1* is a transmembrane glycoprotein that is primarily responsible for cell adhesion, and its downregulation in tumors led to increased invasiveness in breast cancer and lung cancer<sup>51</sup>. Thus, the work here paved the foundation for improving the subtyping of TNBC and luminal A cancers using the above-mentioned biomarkers. Moreover, therapeutics targeting these molecules may be effective at improving TNBC and luminal A cancer patients' survival.

We further investigated the marker genes that have high correlations (Pearson's correlation coefficient  $>0.5$  or  $<-0.5$ ) with the cell-cell interactions, using the two extreme subtypes 1 and 7 (Supplementary Fig. 3). Many above-described relationships

with regard to prognosis are confirmed again. For subgroup 1 (Supplementary Fig. 3a), there exist positive correlations between *CDH1* and epithelial-epithelial cell interactions. Since the loss of *CDH1* is a characteristic of invasive lobular breast cancer<sup>33</sup>, a strong positive correlation between *CDH1* and the epithelial-epithelial interactions does indicate good survival. On the other hand, positive correlations of *MKI67* with epithelial-epithelial interaction and T-cell-epithelial cell interaction in group 7 (Supplementary Fig. 3b) indicate poor survival, as expected<sup>52</sup>. Interestingly, *GATA3* is negatively correlated with T-cell-epithelial cell interaction in group 7, a cell-cell interaction signature for poor survival. However, it is positively correlated with epithelial-epithelial interaction in group 1 with good survival.

While it is true that the cost and time of the multiplex imaging platform currently do not allow live-time assistance to clinicians with precision treatment, this technology is highly reproducible and futuristic. It has already been used for the diagnosis of other diseases like leukemia<sup>53</sup>. This technology has provided information on disseminated tumor cell phenotypes that frequently deviate from the clinical disease subtype<sup>54</sup>. It also advanced our understanding of the intra-tumor heterogeneity and how the interactions among different cell types in tumors and tumor microenvironments affect patient outcomes<sup>15,16,55</sup>. It is expected that in the near future, with the reduction of cost and time and assistance of AI tools and user-friendly computational tools, such technologies will eventually be adopted by pathology and be more informative to clinicians.

In summary, the imaging cytometry data at the single-cell resolution has enabled an unprecedented opportunity to explicitly study cells in the tumor and tumor microenvironment, as well as their interactions. By taking advantage of CP, TMI and TCI features, novel survival subtypes are identified in breast cancers, each with distinct profiles and more molecular and survival homogeneity. Moreover, this subtyping system allows us to identify good survival subpopulations in TNBC and bad survival subpopulations in luminal A cancers, with robust biomarkers in multiple population cohorts. The work lays down the foundation to integrate single-cell resolution information for survival prediction at a large population scale. It has the potential to improve breast cancer prognosis prediction and patient-stratified treatment in the future.

## Data availability

The breast cancer imaging data supporting this study (image mass cytometry data, phenograph neighborhood information, clinical features and patient prognosis outcome) are available online at Zenodo per the original study (<https://doi.org/10.5281/zenodo.3518284>)<sup>56</sup>. The source data for generating the main and supplementary figures are available at the Zenodo repository (<https://doi.org/10.5281/zenodo.10038601>)<sup>57</sup>. The data for validation of the results are available at the TCGA portal (<https://www.cancer.gov/ccg/research/genome-sequencing/tcga>) for TCGA-TNBC data and at cBioPortal ([https://www.cbioportal.org/study/summary?id=brca\\_metabric](https://www.cbioportal.org/study/summary?id=brca_metabric)) for METABRIC data.

## Code availability

The source code to generate the figures and feature datasets for this work is available at [https://github.com/lanagarmire/BC\\_imaging](https://github.com/lanagarmire/BC_imaging)<sup>57</sup>.

Received: 18 May 2023; Accepted: 23 November 2023;

Published online: 19 December 2023

## References

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).

2. Mackillop, W. J. The importance of prognosis in cancer medicine. *TNM Online* <https://doi.org/10.1002/0471463736.tnmp01.pub2> (2006).

3. McAllister, S. S. & Weinberg, R. A. Tumor-host interactions: a far-reaching relationship. *J. Clin. Oncol.* **28**, 4022–4028 (2010).

4. Turashvili, G. & Brogi, E. Tumor heterogeneity in breast cancer. *Front. Med.* **4**, 227 (2017).

5. Lüönd, E., Tiede, S. & Christofori, G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br. J. Cancer* **125**, 164–175 (2021).

6. Dagozo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).

7. Casasent, A. K. et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell* **172**, 205–217.e12 (2018).

8. Yan, X. et al. Comprehensive description of the current breast cancer microenvironment advancements via single-cell analysis. *J. Exp. Clin. Cancer Res.* **40**, 142 (2021).

9. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).

10. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308.e36 (2018).

11. Karaayvaz, M. et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, 3588 (2018).

12. Ding, S., Chen, X. & Shen, K. Single-cell RNA sequencing in breast cancer: understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Commun.* **40**, 329–344 (2020).

13. Saviano, A., Henderson, N. C. & Baumert, T. F. Single-cell genomics and spatial transcriptomics: discovery of novel cell states and cellular interactions in liver physiology and disease biology. *J. Hepatol.* **73**, 1219–1230 (2020).

14. Baharou, H., Canete, N. P., Cunningham, A. L., Harman, A. N. & Patrick, E. Mass cytometry imaging for the study of human diseases—applications and data analysis strategies. *Front. Immunol.* **10**, 2657 (2019).

15. Ali, H. R. et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 (2020).

16. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).

17. Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S. & Garmire, L. X. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* **13**, 112 (2021).

18. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* **14**, e1006076 (2018).

19. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

20. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).

21. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).

22. Ma, X. et al. Identification of a molecular subtyping system associated with the prognosis of Asian hepatocellular carcinoma patients receiving liver resection. *Sci. Rep.* **9**, 1–8 (2019).

23. Frigyesi, A. & Höglund, M. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer Inform.* **6**, 275–292 (2008).

24. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

25. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

26. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

27. Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).

28. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

29. Lehmann, B. D. et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).

30. Zhan, Z. et al. Two-stage Cox-nnet: biologically interpretable neural-network model for prognosis prediction and its application in liver cancer survival using histopathology and transcriptomic data. *NAR Genom. Bioinform.* **3**, lqab015 (2021).

31. Wang, D., Jing, Z., He, K. & Garmire, L. X. Cox-nnet v2.0: improved neural-network-based survival prediction extended to large-scale EMR data. *Bioinformatics* **37**, 2772–2774 (2021).

32. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. B Methodol.* **34**, 187–220 (1972).

33. Dossus, L. & Benusiglio, P. R. Lobular breast cancer: incidence and genetic and non-genetic risk factors. *Breast Cancer Res.* **17**, 37 (2015).

34. Tekpli, X. et al. An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat. Commun.* **10**, 5499 (2019).

35. Lu, S., Yakirevich, E., Wang, L. J., Resnick, M. B. & Wang, Y. Cytokeratin 7-negative and GATA binding protein 3-negative breast cancers: clinicopathological features and prognostic significance. *BMC Cancer* **19**, 1085 (2019).

36. Menz, A. et al. Diagnostic and prognostic impact of cytokeratin 18 expression in human tumors: a tissue microarray study on 11,952 tumors. *Mol. Med.* **27**, 1–16 (2021).

37. Jögi, A., Ehinger, A., Hartman, L. & Alkner, S. Expression of HIF-1 $\alpha$  is related to a poor prognosis and tamoxifen resistance in contralateral breast cancer. *PLoS ONE* **14**, e0226150 (2019).

38. Mimeaull, M. & Batra, S. K. Hypoxia-inducing factors as master regulators of stemness properties and altered metabolism of cancer- and metastasis-initiating cells. *J. Cell. Mol. Med.* **17**, 30–54 (2013).

39. Pellegrino, B. et al. Luminal breast cancer: risk of recurrence and tumor-associated immune suppression. *Mol. Diagn. Ther.* **25**, 409–424 (2021).

40. Kanyilmaz, G. et al. Prognostic importance of Ki-67 in breast cancer and its relationship with other prognostic factors. *Eur. J. Breast Health* **15**, 256–261 (2019).

41. Yerushalmi, R., Woods, R., Ravdin, P. M., Hayes, M. M. & Gelmon, K. A. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol.* **11**, 174–183 (2010).

42. Liu, Z., Li, M., Jiang, Z. & Wang, X. A comprehensive immunologic portrait of triple-negative breast cancer. *Transl. Oncol.* **11**, 311–329 (2018).

43. Saccani, A. et al. p50 nuclear factor-kappaB overexpression in tumor-associated macrophages inhibits M1 inflammatory responses and antitumor resistance. *Cancer Res.* **66**, 11432–11440 (2006).

44. Zhang, H., Qin, G., Yu, H., Han, X. & Zhu, S. Comprehensive genomic and immunophenotypic analysis of CD4 T cell infiltrating human triple-negative breast cancer. *Cancer Immunol. Immunother.* **70**, 1649–1665 (2021).

45. Keith, B. & Simon, M. C. Hypoxia-inducible factors, stem cells, and cancer. *Cell* **129**, 465–472 (2007).

46. Kennedy, A. S. et al. Proliferation and hypoxia in human squamous cell carcinoma of the cervix: first report of combined immunohistochemical assays. *Int. J. Radiat. Oncol. Biol. Phys.* **37**, 897–905 (1997).

47. Mehra, R. et al. Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res.* **65**, 11259–11264 (2005).

48. An, Q. et al. KRT7 promotes epithelial-mesenchymal transition in ovarian cancer via the TGF- $\beta$ /Smad2/3 signaling pathway. *Oncol. Rep.* **45**, 481–492 (2021).

49. Communal, L. et al. A keratin 7 and E-cadherin signature is highly predictive of tubo-ovarian high-grade serous carcinoma prognosis. *Int. J. Mol. Sci.* **22**, 5325 (2021).

50. Lee, H. W. et al. Alpha-smooth muscle actin (ACTA2) is required for metastatic potential of human lung adenocarcinoma. *Clin. Cancer Res.* **19**, 5879–5889 (2013).

51. Oka, H. et al. Expression of E-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis. *Cancer Res.* **53**, 1696–1701 (1993).

52. Soliman, N. A. & Yussif, S. M. Ki-67 as a prognostic marker according to breast cancer molecular subtype. *Cancer Biol. Med.* **13**, 496–504 (2016).

53. Grimwade, L. F., Fuller, K. A. & Erber, W. N. Applications of imaging flow cytometry in the diagnostic assessment of acute leukaemia. *Methods* **112**, 39–45 (2017).

54. Fischer, J. R. et al. Multiplex imaging of breast cancer lymph node metastases identifies prognostic single-cell populations independent of clinical classifiers. *Cell Rep. Med.* **4**, 100977 (2023).

55. Danenberger, E. et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nat. Genet.* **54**, 660–669 (2022).

56. Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. *Zenodo* <https://doi.org/10.5281/zenodo.3518284> (2019).

57. Yadav, S., Zhou, S., He, B., Du, Y. & Garmire, L. Deep-learning and transfer learning identify breast cancer survival subtypes from single-cell imaging data. *Zenodo* <https://doi.org/10.5281/zenodo.10038601> (2023).

## Acknowledgements

We thank Dr. Hartland Jackson for providing data for this project. We thank Yuheng Du in the Garmire group for valuable discussions. This research was supported by grants R01 LM012373 and R01 LM012907 awarded by NLM and R01 HD084633 awarded by NICHD to L.X.G.

**Author contributions**

L.X.G. envisioned this project and supervised the study. S.Y. performed the initial data analysis, generated the main figures, and wrote the manuscript. S.Z. characterized the atypical molecular subtypes, revised the manuscript and figures, tested the code used in this study and made it available on the Github repository. B.H. helped with subgroup selection and interpretation. Y.D. helped test the atypical molecular subtypes with regard to the cell proportions.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43856-023-00414-6>.

**Correspondence** and requests for materials should be addressed to Lana X. Garmire.

**Peer review information** *Communications Medicine* thanks Ke-Da Yu and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permission 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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2023