nature portfolio

Peer Review File

Explainable machine learning–guided integrated multiomics analysis reveals macrophage-driven immune suppression in breast cancer

Corresponding Author: Dr Youness Azimzade

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The is a well written manuscript which further illuminates the role of cells in the TME on survival. The TME has become an increasing important for breast cancer growth, immune escape and targeting. This paper will contribute to our analysis and understanding on several levels.
First the PI using deconvolution of existing databased to understand the cells in the TME and their association with patient outcome. These tools can be used by others in the research community to better understand the roles of cells in the TME on cancer growth.
Second, one the significant cells are identified, they are able to obtain a more granular understanding of their role and the influence of surround cells i.e. the immediate neighborhood on cell function. Again, while yielding insights these types of methodology can be broadly used to improve our understanding.
Finally, they are able to identify specific neighborhood effects on macrophages which influences their gene expression and function in specific breast cancer subtypes. Moreover, their findings are consistent with other reports using orthogonal techniques.
From their data, the team presents hypothesis generating data for further pre-clinical and clinal testing.
I think that this manuscript will contribute to methdology and understanding in the breast as well as broader reserash community.

(Remarks on code availability)
I superficially reviewed the code, however, I did not install it.

Reviewer #3

(Remarks to the Author)
The study by Azimzade and colleagues presents an integrated multi-omics analysis to elucidate the breast tumor microenvironment (TME) and the influence of macrophages on clinical outcomes. The authors performed creative computational analyses of human breast cancer datasets from published single-cell RNA-Sequencing, bulk transcriptomics, and spatial IMC studies, to propose the role of macrophage polarity in driving poor prognostic outcomes and immune suppression in the luminal A subtype of breast cancer.

The authors first applied a single-cell atlas to estimate cellular fractions in bulk transcriptomics datasets from patient tumors using the CIBERSortX deconvolution method. Cell types were inferred across multiple clinical cohorts including TCGA, METABRIC and other neoadjuvant chemotherapy (NAC) trials. The authors developed a new machine-learning based model (SHAP and SSVM/RSF) to identify the associations of inferred cellular fractions with prognostic factors, such as relapse-free survival (RFS) and pathological complete response (pCR). Cell types with opposing survival scores, where RFS and pCR trends were discordant, were identified for different macrophage populations in hormone receptor-positive breast cancers, possibly explained by additional heterogeneity within this disease subtype. Re-stratification of patient

cohorts identified RFS and pCR associations based on macrophage infiltration. Imaging IMC datasets revealed that macrophages exhibited distinct spatial heterogeneity, where subsets in closer proximity to MHC-I & MHC-II epithelial cells associated with worse survival in the Luminal A subtype. Furthermore, these macrophages resided in niches co-localizing with T-cells subsets implicated in immune suppression, including T-regs and exhausted T-cells.

The interplay between macrophages and adaptive immune-responses in the TME are of particular interest to the wider field of cancer immunology. This study presents a new approach to investigating prognostic outcomes from inferred cell fractions, which is commonly used to assess large-scale patient cohorts where bulk RNA-seq is only available. While this method could be applicable to other settings, I do have several concerns about the results and the overall novelty of the findings that should be addressed.

1. This study proposes that HLA-ABC-high Macrophages are exclusively associated with Tregs and Tex in the LumA subtype of breast cancer. This is puzzling as LumA tumors are considered to be immune 'cold' with very low TIL infiltration (PMID: 32621251). Tregs and Tex cells are likely not high in abundance in LumA compared to other subtypes, which raises concerns about the clinical relevance of this proposed axis. The Danenberg et al 2022 study [22], which the spatial IMC data originated from, also reported that "suppressed expansion structures", defined by Tregs and Tex cells, are actually depleted in LumA tumors compared to other subtypes.

2. The clinical association of macrophages in breast cancer is well-established (PMID: 27959923). Similar macrophage-related findings were also reported in Danenberg et al 2022 study [22]. Spatial niches containing macrophages in 'APC-enriched' and T-regs/Exhausted T-cells in the 'Suppressed expansion' structures were reported with increased hazard ratios. The authors should address how their findings, in particular to the re-analysis of spatial IMC data, provide novelty and built-on this prior work.

3. The authors should acknowledge that another potential explanation for the opposing clinical association scores, such as those identified for macrophage subsets (Figure 3, Figure 4b), could be from technical limitations of bulk-deconvolution methods (CIBERSortX). Estimating cellular fractions from bulk data is difficult for highly plastic cell types like myeloid-cells in tumors.

4. This study would greatly benefit from better resolving how previously identified macrophage subsets/clusters from single-cell studies (Fig. 4a) relate to the spatially-defined macrophages described here. Figure 5f shows no enrichment of HLA-ABC expression in different macrophage clusters, however, HLA-ABC-high macrophages are identified to have a high IFN response signature (Figure 5h). Have the authors explored which macrophage subset/cluster (from scRNA-Seq) are enriched for this IFN-signature? The authors should re-consider their analyses to address this point.

5. How is overall Macrophage content determined for high and low groups with the CIBERSortX method (Figure 5c-f)? Is this a signature derived from all macrophages in scRNA-Seq data? The authors should comment on this and highlight why is this able to identify similar pCR and survival trends compared to the subset-level data (Figure 5b).

6. It appears that the NAC cohort is an aggregate of patient data from 15 different studies, as described in the methods. This can contribute to significant technical batch effects in downstream analyses, such as the deconvolution steps using CIBERSORTx. In addition, the authors should demonstrate the consistency of the survival and pCR results across different cohorts used.

7. Have the authors explored spatial IMC data from other subtypes such as LumB, Her2, Basal, to investigate if this Macrophage polarity in MHC-I & II epithelial niches, and co-localization with Tregs/Tex is unique to LumA tumors?

8. The authors should consider analyzing CD4 T-regs independently from Exhausted CD8+ T-cells when investigating spatial co-localization with Macrophages (Fig. 5). These are very distinct T-cell subsets that secrete distinct factors and hold unique functions and fates. What is the proportion each cell type in these Macrophage niches?

9. In Figure 4e, a very small number of patients with Mac=0 (n=9) is compared to a large number of patients with Mac>0 (n=334). The authors should consider appropriate statistical analyses to look at this cellular correlation, that might include stratification based on a mean or median, or some continuous metric.

10. Many different patient cohorts and types of omics data are being re-analyzed in this study. It is hard to interpolate which cohorts are being used from the current manuscript text, figures, and figure legends.

(Remarks on code availability)
The github repository contains excellent and detailed documentation of the various R and python code used in this study, including the machine-learning model developed by the authors.

Reviewer #4

(Remarks to the Author)
We want to provide a technical review of the machine learning and computational aspects of the study. Addressing these points will significantly strengthen the study's rigor and reproducibility.
Expansion of Predictive Models:
The current study evaluates the performance of Random Survival Forests (RSF) and Survival Support Vector Machines

(SSVM). However, comparing these models with additional survival analysis methods, such as the Cox Proportional Hazards Model and Gradient-Boosted Models is essential. This comparison would provide a more thorough assessment of the proposed approach.
The authors utilized two machine learning models (RSF and SSVM) to predict relapse risk. Although the distinctive characteristics of RSF and SSVM may help minimize model selection bias, concerns persist regarding whether this bias was adequately addressed. As illustrated in Figure 1, Step 2, certain clinical variables were excluded from survival score calculations due to inconsistent slope associations between the two models, which could lead to information loss. The purple clinical variable lacks a calculated survival score because of the differing slope associations between the two models. Additional analyses and discussions should be included.
Hyperparameter Optimization Details:
The manuscript mentions the use of GridSearchCV for hyperparameter tuning. However, it is unclear which hyperparameters were considered for each model. Please provide a detailed list of the hyperparameters included in the search space and their respective ranges. Additionally, report the optimal hyperparameter values selected for each model to enhance transparency and reproducibility.
Mathematical Formulation of SHAP Values and Other Metrics:
The explanation of SHAP values is currently qualitative. Please include the mathematical formulation used to compute SHAP values. Similarly, providing explicit mathematical definitions or references to standard formulations would be beneficial for other performance metrics used in the study, such as the concordance index (C-index).
While SHAP value-based model interpretation is valuable for assessing the independent contributions of each cell type, it may miss the complex interactions among cell types within the tumor microenvironment. For instance, interactions between macrophages and T cells play a crucial role in regulating immunosuppression and anti-tumor immune responses. Thus, it is necessary to directly analyze the effects of interactions between cell types or enhance the analysis by including interaction terms in the model. Furthermore, it is essential to compare the analysis results that account for cell-to-cell interactions with those based on existing SHAP values to evaluate the impact of these interactions on the results quantitatively.
The validation of SHAP values used to represent the association of each cell type with survival probability appears inadequate. An ML model was created to predict the risk of relapse using clinical variables for each cell type (Figure 1, Step 1). This ML model achieved a C-index performance of approximately 0.7, which is difficult to regard as robust predictive performance compared to other models with similar objectives (https://pmc.ncbi.nlm.nih.gov/articles/PMC11483452/), given that 0.5 indicates random prediction in C-index characteristics. Furthermore, cross-dataset validation was not conducted. When an ML model's performance is lacking, it becomes hard to trust the explanatory power of features in relation to the output, as there may be noise or critical signals that are absent. Additionally, the justification for using relapse as a criterion for feature selection to investigate the association with the survival probability-related survival rate (RFS) and chemotherapy response (pCR) is still weak.
Reproducibility and Code Execution Requirements:
To ensure reproducibility, please provide details on the Python version and the specific versions of key libraries used in the experiments (e.g., scikit-learn, etc.). This information can be included in the manuscript or made available in the corresponding GitHub repository. Additionally, providing a requirements.txt or environment.yml file in the repository would facilitate code execution in a consistent environment.
Data preprocessing and batch effects:
For each figure in the results, it's essential to specify the number of data points used and their sources. This helps verify whether the results derive from biased outcomes in a specific cohort or a small dataset. The number of patients per data point will significantly decrease when segmented by subtypes. If the results are based on limited data points, limitations may arise in drawing general inferences. Specifying the number of data points is crucial to ensure the reliability of the results.
Neglecting to account for batch effects among cohorts can lead to biased outcomes that favor a specific cohort when merging data from different cohorts in large-scale datasets. The study does not indicate how batch effects were addressed. If batch effects were minimized, this should be elaborated in the methods section. Typically, ComBat or SVA are employed to adjust for batch effects.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #4

(Remarks to the Author)
The author has addressed all the referees' requests, and the revised version effectively incorporates their feedback.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
PART I
Evaluation of the authors' response to Reviewer 3's comments

Q1
Reviewer 3 notes that LumA, an immune "cold" subtype, exhibits low tumor-infiltrating lymphocytes (including Tregs and Tex cells). This casts doubt on the authors' reported correlation between Treg-Tex-macrophage co-localization and clinical outcomes.
Although the authors have presented detailed data confirming the presence of Tregs and Tex cells in LumA tumors (Figure S11), this fails to address the core concern. Notably, the authors propose that Tregs and Tex cells exert substantial effects in LumA tumors despite their low abundance, which is markedly lower than in other subtypes. This raises a critical question: why are more pronounced immunosuppressive effects not observed in other subtypes with higher abundance of these cells? Given the low abundance of Tregs and Tex cells in LumA, their attribution of the formation of the immunosuppressive microenvironment to this mechanism is questionable. Consequently, additional evidence (supplemental analyses or literature citations) is recommended to support this mechanism.

Q2
Reviewer 3 posits that Danenberg et al. (2022) have previously demonstrated that the spatial architecture formed by 'APC-enriched' macrophages or Tregs and Tex cells is linked to increased hazard ratios. Consequently, the novelty of the insights reported by the authors need to be clarified.
In their response, the authors emphasize the innovative contributions of their study in two key respects: firstly, they have clearly identified the HLA-ABC-high macrophage subset as the functionally relevant population; secondly, they have elucidated the specific impact of the spatial distribution of these three cell types on adverse outcomes. In my assessment, the authors' response effectively underscores the innovative value of their research. Building upon prior work which indicates that spatial niches consist of 'APC-enriched' macrophages or Tregs and Tex cells can form a immunosuppressive microenviroment respectively, their study further elucidates that the spatial colocalization of HLA-ABC-high macrophages, regulatory T cells (Tregs), and exhausted T cells (Tex cells) orchestrates the formation of an immunosuppressive microenvironment, which in turn contributes to polarized clinical outcomes. This work offers a more refined and detailed mechanistic dissection of the underlying processes.

Q3
Reviewer 3 raised concerns regarding the technical limitations from bulk deconvolution methods (CIBERSortX).
The authors have acknowledged this technical issue and incorporated an additional statement in the manuscript. Given the widespread application of similar deconvolution approaches within the field, I consider the authors' response to this query to be adequate.

Q4
Reviewer 3 argues that the study needs to further clarify the association between the macrophage subsets identified in previous single-cell studies and the spatially defined macrophages.
The authors have supplemented a volcano plot (Figure 5g) indicating that the CXCL10 subset shows enrichment in the IFN response. In my assessment, the authors' response has partially addressed the reviewer's query. However, the authors should provide a more detailed explanation of the definition of HLA-ABC-high/low macrophages. For example: Do "HLA-ABC-high macrophages" refer to macrophages with relatively high expression of all HLA-A, B, and C, or those with high expression of one or two of these genes?

Q5
Reviewer 3 inquired about the methodology by which the authors stratified patients into high and low macrophage groups.
The authors have supplemented violin plots (Figure 4c) and explanations to clarify the stratification criteria. In my view, the responses provided by the authors are rationale-based.

Q6
Reviewer 3 argues that, given the use of multi-source datasets, the authors should correct batch effects and verify the consistency of efficacy and prognostic assessments across different cohorts.
The authors have presented UMAP plots (Figure S1) before and after correction, along with justifications for their decision not to perform batch effects correction. In my assessment, the authors' response is insufficient for the following reasons:
1. In the UMAP plot (Figure S1) provided by the authors, we found that some datasets exhibit significant batch effects regardless of whether batch correction is performed (e.g., GSE22358). The authors should implement measures to tackle this issue such as excluding certain genes in the dataset that induce significant batch effects because of the different detection methodologys (e.g., specific non-coding RNAs) or considering the exclusion of a particular cohort.
2. The authors should provide the details of how pathological complete response (pCR) was evaluated in the different neoadjuvant chemotherapy (NAC) cohorts they used, such as specifying whether the Miller-Payne (MP) grading system or the residual cancer burden (RCB) classification was employed.

Q7
Reviewer 3 suggested that the authors should verify whether the co-localization of macrophages, Tregs, and Tex cells is universally present across various PAM50 tumor subtypes.
The authors have provided evidence (Figure S13) in response to this comment. This response is deemed reasonable and adequate.

Q8
Reviewer 3 argues that, given the significant functional differences between Tregs and Tex cells, these two cell subsets should be analyzed separately.

The authors responded that these two subsets were grouped together in the annotations from Dannenberg et al.'s work. In my assessment, this response is highly illogical: due to the substantial differences between these two cell subsets, they clearly warrant separate analysis. The authors should consider using a dataset which can analyze Tregs and Tex cells separately.

Q9
Reviewer 3 contends that the authors employed inappropriate statistical analyses to demonstrate the differences in Tregs and Tex cells between the MAC0 and MAC > 0 groups.
The authors maintain that their chosen analytical methods and presentation format are reasonable. I align with the authors' perspective, as the current analysis and presentation format more effectively convey the intended meaning of the authors' work.

PART Ⅱ
Other issues in this manuscript

Q1
It is generally recognized that Luminal breast cancer, particularly LumA subtype, exhibits a low pathological complete response (pCR) rate to neoadjuvant chemotherapy but demonstrates favorable overall prognosis following standard adjuvant endocrine therapy. This dichotomous clinical outcome may be attributed to the fact that tumors with higher differentiation exhibit immune-cold characteristics, resulting in a lower pathological complete response (pCR) rate. Meanwhile, the higher degree of differentiation is also associated with reduced tumor malignancy and more favorable prognosis. Therefore, this dichotomous clinical phenomenon may not necessarily be driven by the tumor microenvironment (TME) but could also stem from the intrinsic biological properties of the tumor itself.
The authors have presented the efficacy or prognostic value of different immune cell types across various tumor subtypes in Figure 2; however, many cell types were not displayed due to failure to meet the threshold. We would like to see the complete edition to verify the specificity of the association between macrophages and this dichotomous clinical phenotype, as other types of immune cells may also exhibit similar clinical correlations to macrophages.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #5

(Remarks to the Author)
I have substantial concerns about this manuscript. Its overall scope is limited, the conceptual advances feel incremental, and the rationale for several analyses is not sufficiently robust. The idea that MHC^+ macrophages drive T-cell exhaustion (T_ex) has already been mechanistically explored in at least two solid immunology studies (PMID: 39724910, Immunity, 2025; PMID: 35623342, Cancer Cell, 2022), which significantly reduces the novelty of the main analytical conclusions here. Overall, compared with other works of a similar type currently published in Nature Communications, the competitiveness and reader appeal of this study are inadequate.

Q1. In Fig. S9, the prognostic association of macrophages is significant in only one cohort, and is even reversed in TCGA. This suggests the phenomenon may not be broadly robust or generalizable.
Moreover, macrophages are highly heterogeneous, so their functions should not be analyzed as a single lumped population. In Fig. 4, however, the authors stratify by overall macrophage abundance rather than by subset abundance, calling into question the true clinical and biological meaning of the analyses presented in Fig. 4.

Q2. In Fig. 3, why are CD8 T cells associated with non-pCR? As the canonical cytotoxic effector cells in anti-tumor immunity, how is this result explained? Does this finding call into question the validity of the pCR score?

Q3. The authors still do not clearly show which single-cell–defined macrophage subset corresponds to the MHC^+ macrophages observed in the IMC data, nor how those subsets relate to the groups in Fig. 4. At present, the analyses of IMC, single-cell, and bulk RNA-seq data remain poorly integrated.

Q4. Please clarify the precise definition of Tex used here. Is it closer to Tex_prog, Tex_int, or Tex_term?

Q5. Please verify whether the group labels in Fig. 4g are incorrect.

Q6. In Fig. S5, the point colors and line colors should not be the same; as currently drawn they are hard to distinguish.

(Remarks on code availability)

Version 3:

Reviewer comments:

Reviewer #5

(Remarks to the Author)
The authors have addressed most of the concerns, but several issues remain to be discussed:

1. In Fig. 5g, the MHC-lo group shows upregulation of MHC genes. Please verify whether the group labels may have been incorrectly assigned. In addition, if the labels are indeed incorrect, the MHC-hi group shows upregulation of CXCL9. In PMID: 37535729 (2023, Science), a high CXCL9/SPP1 ratio was associated with better prognosis and superior immunotherapy pCR across multiple cancer types. Please present the CXCL9/SPP1 ratio and their individual expression levels in the MHC-lo/high groups, as well as the correlations between MHC gene expression and CXCL9, SPP1 individual expression, and the CXCL9/SPP1 ratio in macrophages. Please discuss how the above analytical results may align with or contradict the conclusions of the article PMID: 37535729 and the conclusions of the present study.

2. Please replace Fig. 5b and Fig. 5c with versions in which Treg and Tex are labeled separately.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
The authors have addressed the comment of Reviewer 5 in an appropriate way, and have provided justified interpretation of their results.
Figures 5 was corrected, as requested.
The explanation for the link between the CXCL9/SPP1 ratio and HLA-ABC levels is fully satisfactory. The correlations can be cancer -specific , cohort-specific and model-specific.
There is current gap of knowledge in our understanding what is the driving mechanism in macrophage-mediated immunosuppression, and what is just a biomarker manifestation, however, not reflecting the key regulatory pro-or anti-tumoral functions.
The data provided by the submission are valuable, even if the results are controversial to some data published by others.

Evidence provided by various research groups analysing of various cancer patient cohorts is need to discriminate between “drivee or passanger”, and to generate clinically useful knowledge. In this context, the results of the submission are relevant.

(Remarks on code availability)

Response to Reviewers

We sincerely appreciate the reviewers' constructive feedback. Their valuable insights have significantly enhanced our manuscript and their professional approach made the revision process a pleasure. Below, we address each comment in detail. Revisions to both the main text and the Supplementary Information are in blue for clarity. For each point raised, we also provide corresponding changes here as well (please note that Figure names dont follow that of main text or SI).

COMMENTS FROM REVIEWER #1

Reviewer: *The is a well written manuscript which further illuminates the role of cells in the TME on survival. The TME has become an increasing important for breast cancer growth, immune escape and targeting. This paper will contribute to our analysis and understanding on several levels. First the PI using deconvolution of existing databased to understand the cells in the TME and their association with patient outcome. These tools can be used by others in the research community to better understand the roles of cells in the TME on cancer growth. Second, one the significant cells are identified, they are able to obtain a more granular understanding of their role and the influence of surround cells i.e. the immediate neighborhood on cell function. Again, while yielding insights these types of methodology can be broadly used to improve our understanding. Finally, they are able to identify specific neighborhood effects on macrophages which influences their gene expression and function in specific breast cancer subtypes. Moreover, their findings are consistent with other reports using orthogonal techniques. From their data, the team presents hypothesis generating data for further pre-clinical and clinal testing. I think that this manuscript will contribute to methdology and understanding in the breast as well as broader reserash community.*

Reply: We appreciate the positive feedback on our manuscript and its contributions.

COMMENTS FROM REVIEWER #3

Reviewer: *The study by Azimzade and colleagues presents an integrated multi-omics analysis to elucidate the breast tumor microenvironment (TME) and the influence of macrophages on clinical outcomes. The authors performed creative computational analyses of human breast cancer datasets from published single-cell RNA-Sequencing, bulk transcriptomics, and spatial IMC studies, to propose the role of macrophage polarity in driving poor prognostic outcomes and immune suppression in the luminal A subtype of breast cancer.*

The authors first applied a single-cell atlas to estimate cellular fractions in bulk transcriptomics datasets from patient tumors using the CIBERSortX deconvolution method. Cell types were inferred across multiple clinical cohorts including TCGA, METABRIC and other neoadjuvant chemotherapy (NAC) trials. The authors developed a new machine-learning based model (SHAP and SSVM/RSF) to identify the associations of inferred cellular fractions with prognostic factors, such as relapse-free survival (RFS) and pathological complete response (pCR). Cell types with opposing survival scores, where RFS and pCR trends were discordant, were identified for different macrophage populations in hormone receptor-positive breast cancers, possibly explained by additional heterogeneity within this disease subtype. Re-stratification of patient cohorts identified RFS and pCR associations based on macrophage infiltration. Imaging IMC datasets revealed that macrophages exhibited distinct spatial heterogeneity, where subsets in closer proximity to MHC-I & MHC-II epithelial cells associated with worse survival in the Luminal A subtype. Furthermore, these macrophages resided in niches co-localizing with T-cells subsets implicated in immune suppression, including T-regs and exhausted T-cells.

*The interplay between macrophages and adaptive immune-responses in the TME are of particular interest to the wider field of cancer immunology. This study presents a new approach to investigating prognostic outcomes from inferred cell fractions, which is commonly used to assess large-scale patient cohorts where bulk RNA-seq is only available. While this method could be applicable to other settings, I do have several concerns about the results and the overall novelty of the findings that should be addressed.- **1**. This study proposes that HLA-ABC-high Macrophages are exclusively associated with Tregs and Tex in the LumA subtype of breast cancer. This is puzzling as LumA tumors are considered to be immune 'cold' with very low TIL infiltration (PMID: 32621251). Tregs and Tex cells are likely not high in abundance in LumA compared to other subtypes, which raises concerns about the clinical relevance of this*

proposed axis. The Danenberg et al 2022 study [22], which the spatial IMC data originated from, also reported that "suppressed expansion structures", defined by Tregs and Tex cells, are actually depleted in LumA tumors compared to other subtypes.

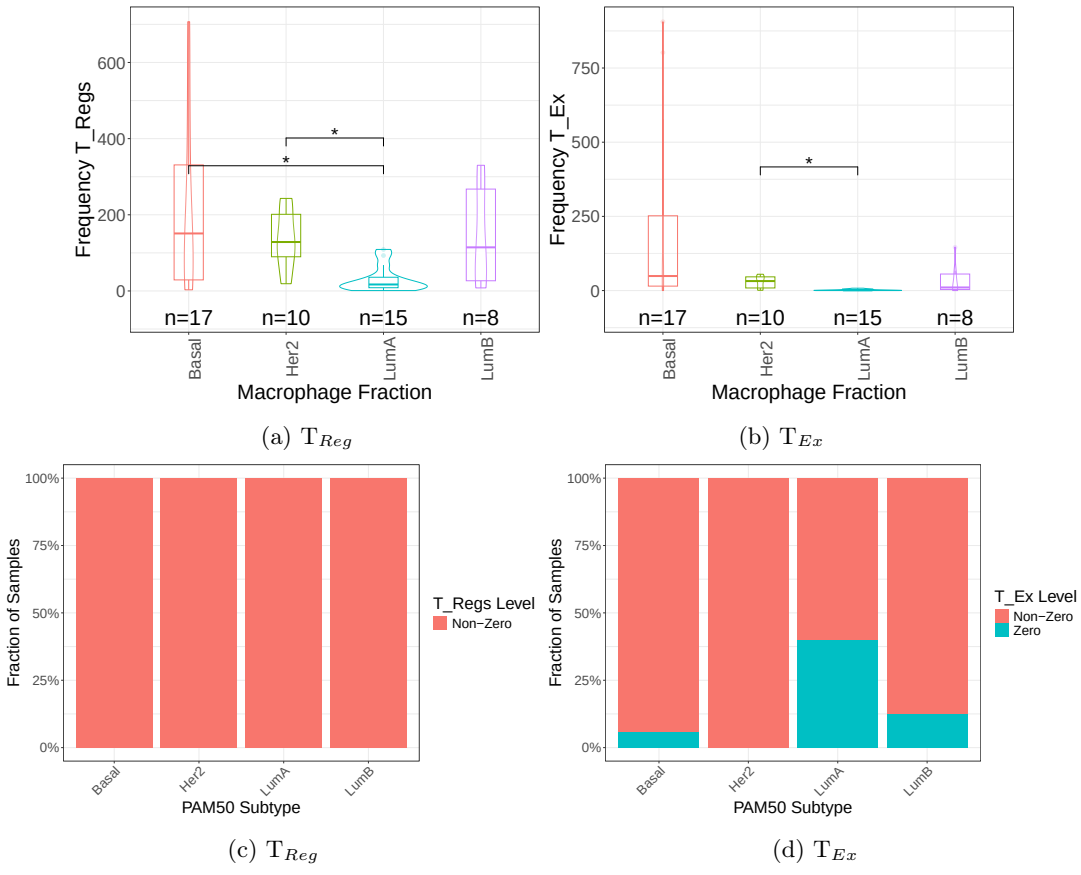
Reply: Thank you for pointing this out. To address this concern, we conducted additional analyses in scRNA-seq and IMC data. First, we compared the frequency of Tex and Tregs across PAM50 subtypes. Then we assessed whether these cells were present in each subtype. As the reviewer noted, the frequency of these cells is lower in LumA. However, in the scRNA-seq data, all LumA samples contained Tregs, and the majority of them contained Tex cells. In the IMC data, approximately 25% of LumA samples showed the presence of Tex and Tregs. Given that these cells appeared in only about 50% of Basal and Her2 samples of IMC data (while they exist in almost all scRNA-seq samples), the relatively low numbers in LumA may reflect technical limitations of such data. While the frequency of Tex and Treg is lower in LumA and ER+ samples, we believe these tumors still contain considerable amounts of these cells to be clinically relevant.

A new dedicated section was added to SI discussing these results and the main text was changed accordingly:

T_{Reg} AND T_{Ex} FREQUENCY IN TUMOR SUBTYPES.

scRNA-seq Analysis

In the scRNA-seq dataset, we analyzed the frequency of T_{Reg} and T_{Ex} cells separately across the PAM50 subtypes. Our results indicate that while the overall frequency of these cells is significantly lower in LumA compared to other subtypes, all LumA samples contained Tregs, and the majority also included Tex cells. This suggests that while Tex and Tregs may be less abundant in LumA tumors, they are still present in most cases.



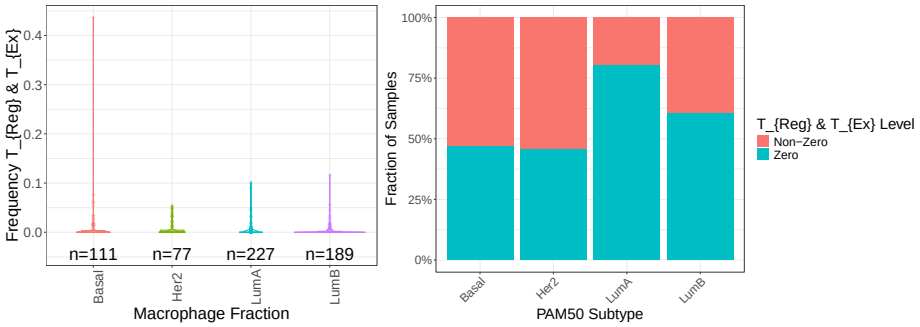
(a) T_{Reg} (b) T_{Ex}

(c) T_{Reg} (d) T_{Ex}

Figure 1: Frequency of T_{Reg} (a) and T_{Ex} (b) across PAM50 subtypes. Fraction of samples including T_{Reg} (c) and T_{Ex} (d) across PAM50 subtypes.

IMC Analysis

In the IMC dataset, we examined the presence of Tex and Tregs together across PAM50 subtypes. We observed that approximately 25% of LumA samples contained both Tex and Tregs. This proportion is lower than in Basal and Her2 subtypes, where these cells were detected in about 50% of the samples. Given the spatial nature of IMC data, the lower detection rate in LumA may be influenced by technical limitations, such as lower sensitivity for rare cell types in highly heterogeneous tumor environments.



(a) (b)

Figure 2: a) Frequency of T_{Reg}&T_{Ex} across PAM50 subtypes. b) Fraction of samples including T_{Reg}&T_{Ex} across PAM50 subtypes.

Reviewer: *2. The clinical association of macrophages in breast cancer is well-established (PMID: 27959923). Similar macrophage-related findings were also reported in Danenberg et al 2022 study [22]. Spatial niches containing macrophages in 'APC-enriched' and T-regs/Exhausted T-cells in the 'Suppressed expansion' structures were reported with increased hazard ratios. The authors should address how their findings, in particular to the re-analysis of spatial IMC data, provide novelty and built-on this prior work.*

Reply: Thank you for this suggestion. We agree it is important to more clearly acknowledge prior work in the field and clarify how our findings built-on it by delving deeper into the functional heterogeneity and spatial context of macrophages to explain their complex roles in treatment response and survival. The following text was added to the updated version of the discussion:

The Danenberg et al. study provided an important framework linking macrophage-rich niches to adverse outcomes [1]. Through re-analysis of spatial IMC data from this work, we corroborate and extend the earlier findings by revealing a dichotomy by which specific phenotypic and spatial profiles may drive opposing clinical outcomes in breast cancer. Danenberg et al. identified spatial niches, such as 'APC-enriched' and 'Suppressed expansion' structures with increased hazard ratios driven by macrophages. Our data delineates further subdivision of macrophages based on HLA-ABC expression, in which HLA-ABC^{hi} macrophages are particularly associated with immunosuppressive niches. Furthermore, our findings suggest that the precise localization of macrophage subtypes within the TME significantly contributes to their functional duality. We further demonstrate that with increasing macrophage infiltration, there is a reorganization of their spatial context, including colocalization with T_{Reg} and T_{Ex} cells, particularly in LumA samples, providing novel insight into the dynamic interplay between these immune cells. In summary, our study further advances knowledge in this field by offering a refined classification of macrophages based on both their spatial distribution and marker expression.

Reviewer: *3. The authors should acknowledge that another potential explanation for the opposing clinical association scores, such as those identified for macrophage subsets (Figure 3, Figure 4b), could be from technical limitations of bulk-deconvolution methods (CIBERSortX). Estimating cellular fractions from bulk data is difficult for highly plastic cell types like myeloid-cells in tumors.*

Reply: We recognize the importance of this point and have included a dedicated section to the text discussing the limitations of the study. We have also adapted the main text accordingly. Text was updated as below:

... subset of macrophages (here LAM1) can exhibit opposite pCR and Survival scores. It should be noted that using in-silico methods to infer frequency of highly plastic cells is challenging and these results are subject to limitations of

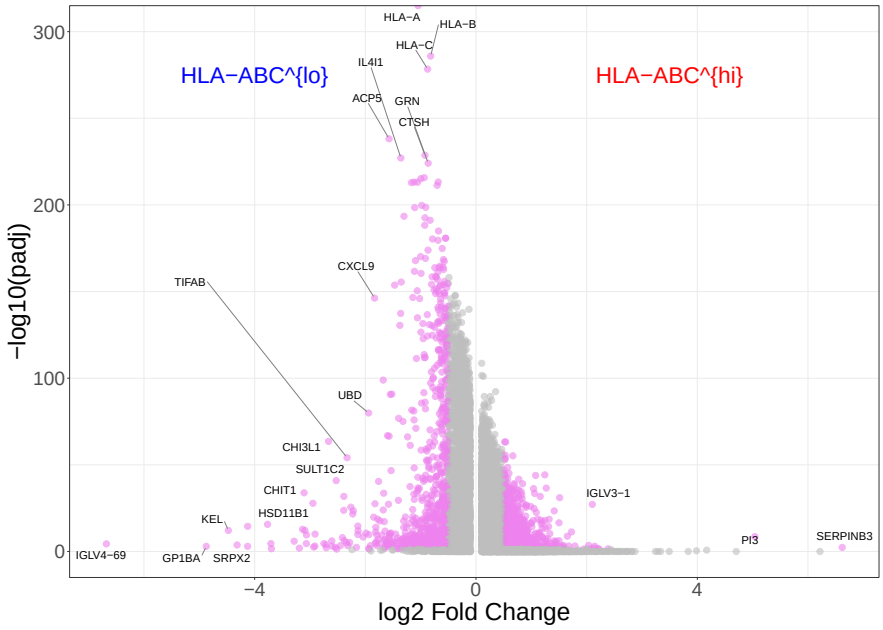
our method as well.

And a new section entitled "limitations of this study" was added that includes:

This study relies on in silico estimation of the cellular content of the TME using CIBERSORTx. Such methods inherently face limitations in capturing detailed expression profiles and discriminating between low-abundance cell types due to noise and variability in RNA data, as well as similarities between the expression profiles of different cells. Although these methods are widely used for estimating macrophages [**?**] and their subsets [2], these limitations should be taken into account when interpreting the presented results.

Reviewer: *4. This study would greatly benefit from better resolving how previously identified macrophage subsets/clusters from single-cell studies (Fig. 4a) relate to the spatially defined macrophages described here. Figure 5f shows no enrichment of HLA-ABC expression in different macrophage clusters, however, HLA-ABC-high macrophages are identified to have a high IFN response signature (Figure 5h). Have the authors explored which macrophage subset/cluster (from scRNA-Seq) are enriched for this IFN-signature? The authors should re-consider their analyses to address this point.*

Reply: We fully agree with this suggestion. In our analysis, while HLA-A, B, and C showed the most significant p values, they were dropped because of their lower log-fold change values ($\sim$1). We have updated the plot, and these genes appear as the most significant points. See below the updated volcano plot:



(a)

Figure 3: DEA (g) for HLA-ABC$^{\text{lo}}$ vs. HLA-ABC$^{\text{hi}}$ macrophages.

In addition, we added a new analysis where each major macrophage subtype was compared to the rest of macrophages and results were added to the SI. The CXCL10 subset shows enrichment in IFN response.

DEA AND PEA FOR MACROPHAGES SUBTYPES

We carried out differential expression analysis (DEA) and pathway enrichment analysis (PEA) on macrophage subsets from our single-cell RNA sequencing data. These analyses aimed to determine which macrophage cluster is enriched for the interferon (IFN) response signature observed in HLA-ABC$^{\text{hi}}$ macrophages.

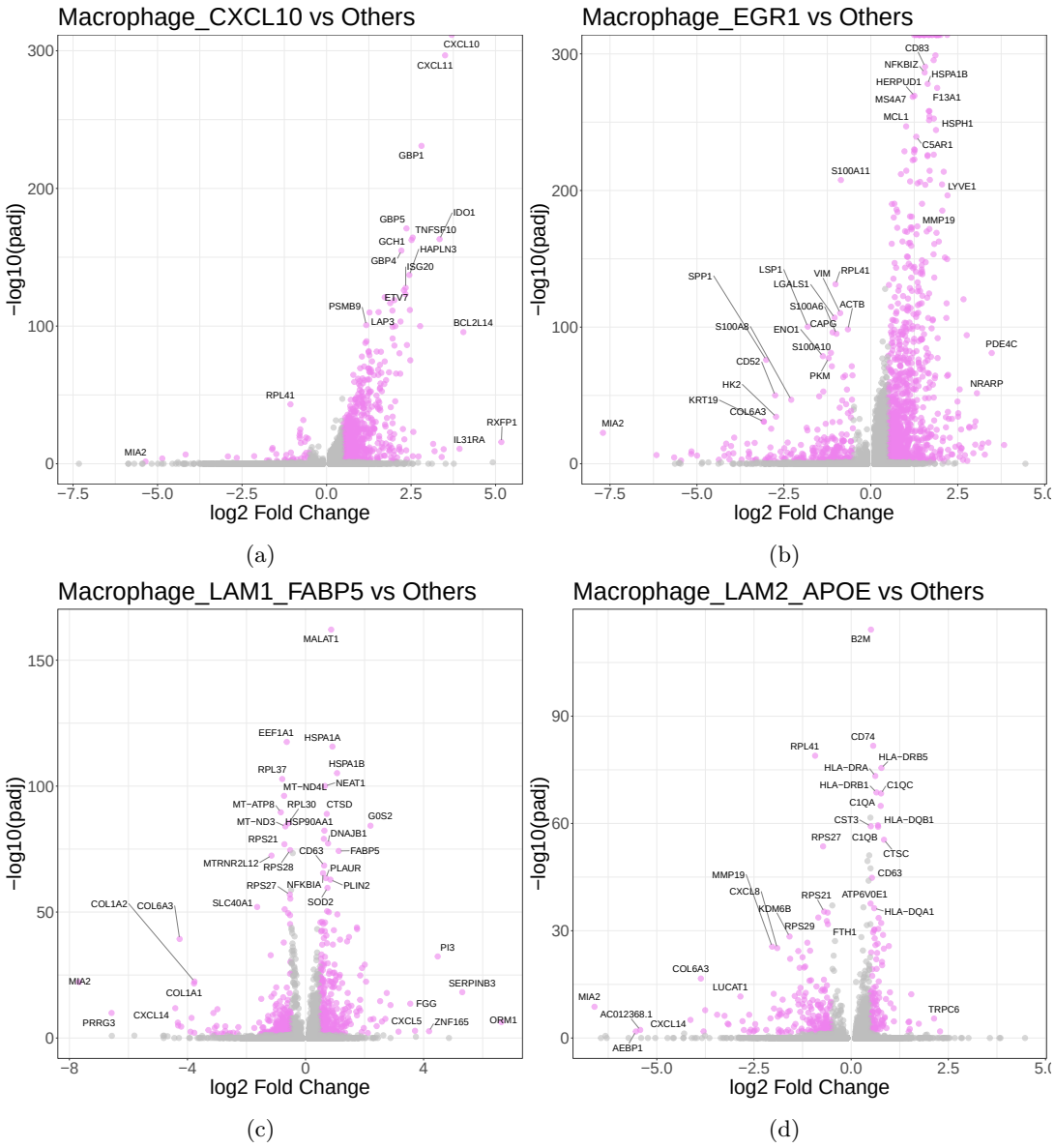


Figure 4: Volcano plots for each subset of macrophages compared with all other subsets.

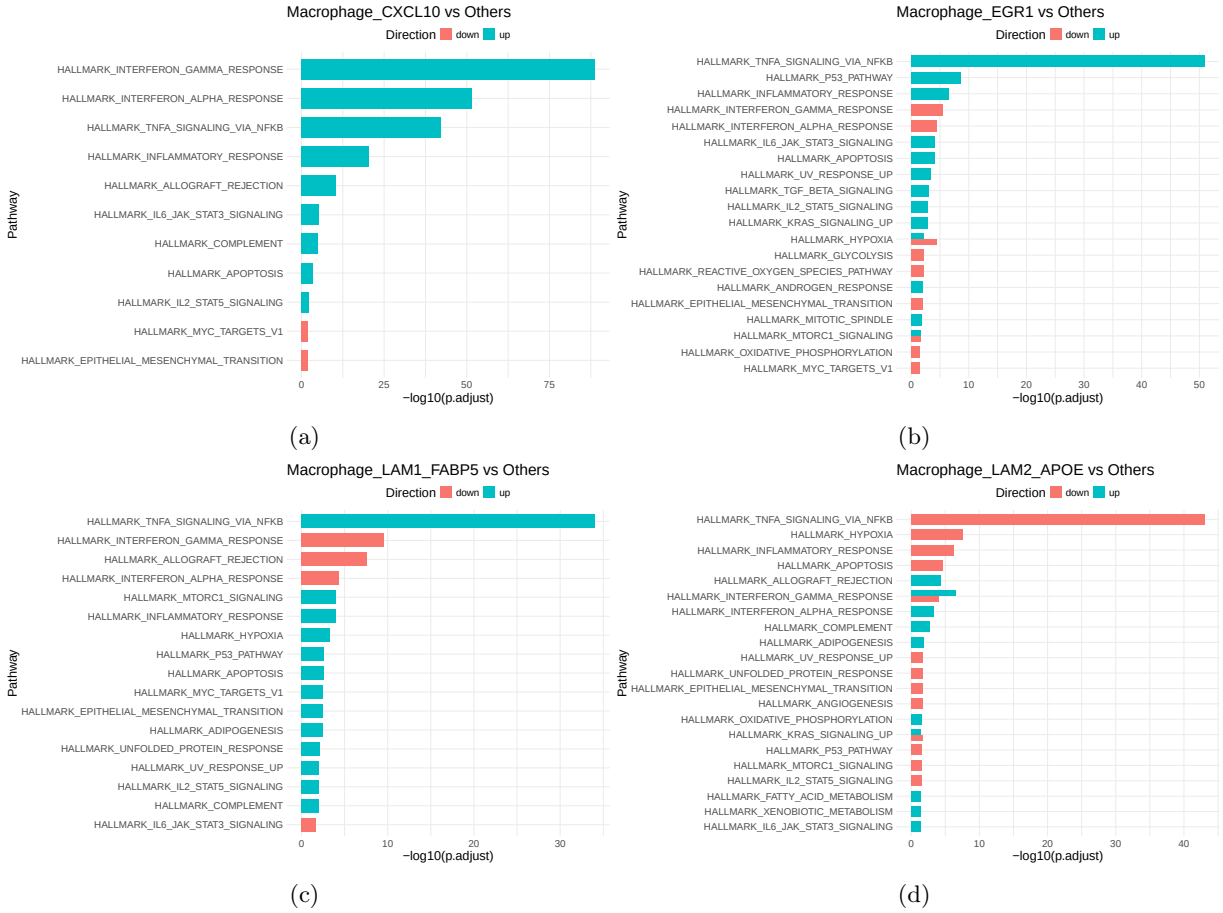
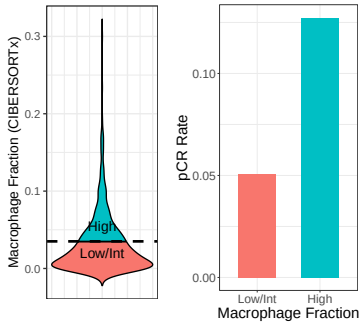


Figure 5: PEA for each subset of macrophages vs other subsets.

Reviewer: ***5****. How is overall Macrophage content determined for high and low groups with the CIBERSortX method (Figure 5c-f)? Is this a signature derived from all macrophages in scRNA-Seq data? The authors should comment on this and highlight why is this able to identify similar pCR and survival trends compared to the subset-level data (Figure 5b).*

Reply: Thank you, we acknowledge the need for clarification. We added a new subplot to Fig4c and the text below explaining how the groups were identified.



(a)

Figure 6: LumA samples from NAC cohort (n=589) were divided into two groups (Low/Int and High) based on fraction of macrophages estimated through deconvolution.

In our NAC cohort, we categorized LumA samples (n=589) into two groups based on macrophage content: one

with a low to intermediate fraction (67 % of samples) and another with a high fraction (33% of samples). In other words, based on estimated fraction of macrophages using CIBERSORTx, we divide LumA samples into groups with different levels of macrophage content (see Fig. 6(c)).

Reviewer: ***6****. It appears that the NAC cohort is an aggregate of patient data from 15 different studies, as described in the methods. This can contribute to significant technical batch effects in downstream analyses, such as the deconvolution steps using CIBERSORTx. In addition, the authors should demonstrate the consistency of the survival and pCR results across different cohorts used.*

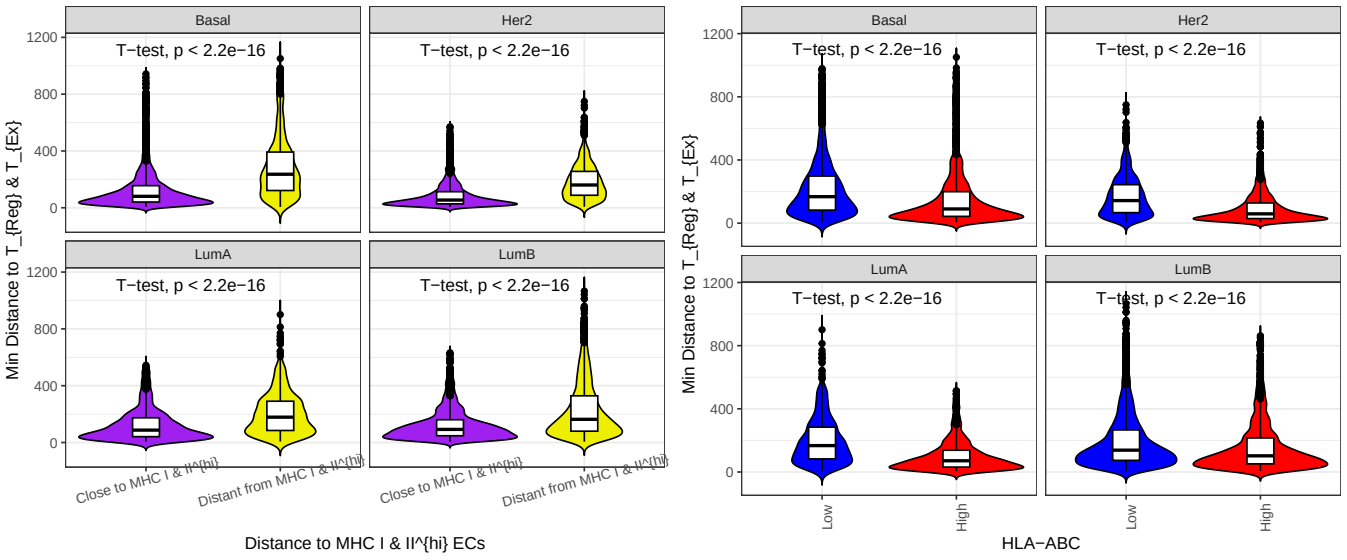
Reply: We appreciate this concern and have included a discussion on batch effect correction in the revised manuscript. To ensure a reliable measure and avoid trial specific biases, the NAC cohort itself was randomly divided into two sets (9 trials with 1009 samples as discovery set and 6 trials with 998 as validation set) and results were cross validated there. A section was added to SI to discuss this issue, see .

Reviewer: ***7****. Have the authors explored spatial IMC data from other subtypes such as LumB, Her2, Basal, to investigate if this Macrophage polarity in MHC-I & II epithelial niches, and co-localization with Tregs/Tex is unique to LumA tumors?*

Reply: We thank the reviewer for this suggestion. We have added two new series of analyses to the SI to address this comment:

MACROPHAGES AND T_{Reg}&T_{Ex} ACROSS SUBTYPES

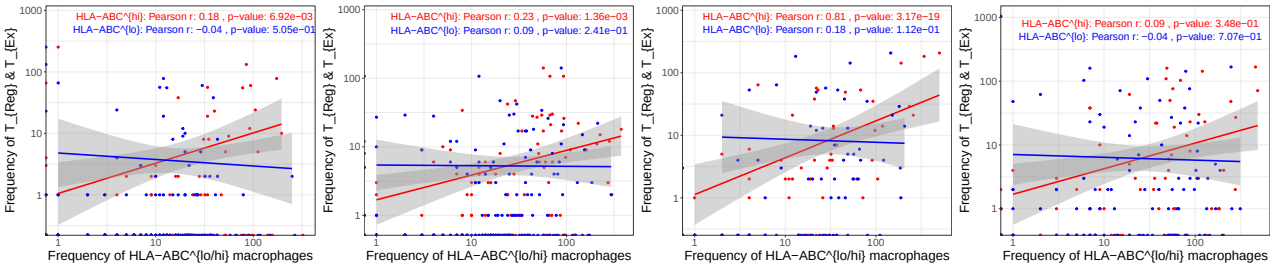
Next, we compare the distance between subsets of macrophages and T_{Reg}&T_{Ex} across PAM50 subtypes. First, we divide macrophages into two groups based on their minimum distance from MHC I & II^{hi} and then compare the distance to T_{Reg}&T_{Ex} (see Fig. 7 (a)). In addition, we divide macrophages into two groups with high or low levels of HLA-ABC and compare their minimum distance from T_{Reg}&T_{Ex}, as shown in Fig. 7 (b). Both analyses suggests that colocalization of close to MHC I & IIhi ECs and HLA-ABC^{hi} with T_{Reg}&T_{Ex} is not limited to an specific subset of breast tumors.



(a) (b)

Figure 7: Comparison of minimum distance of T_{Reg}&T_{Ex} with two subsets of macrophages that are divided based on their distance to MHC I & II^{hi} (a) and HLA-ABC level (b) in PAM50 subtypes.

As a supplementary analysis, and similar to Fig. 5 (d), we compare correlation of HLA-ABC^{hi} macrophages to frequency of T_{Reg}&T_{Ex}. The results, shown in Fig. 8, further confirms that HLA-ABC^{hi} macrophages are exclusively correlated with T_{Reg}&T_{Ex}.



(a) LumA (b) LumB (c) Her2 (d) Basal

Figure 8: Correlation of frequency of HLA-ABC$^{\text{hi}}$ and HLA-ABC$^{\text{lo}}$ macrophages and T_{Reg}&T_{Ex} in all PAM50 subtypes.

The main text also was edited to include these results:

Further analysis of T_{Reg}&T_{Ex} frequency near two macrophage groups across LumA samples supports this finding (refer to Fig. 5(c)). This pattern consistently appears across other PAM50 subtypes as well, with T_{Reg}&T_{Ex} being located closer to macrophages that are close to MHC I & II$^{\text{hi}}$ and to macrophages with high HLA-ABC levels (see Suppl. Fig. S13)

.... Interestingly, the HLA-ABC$^{\text{hi}}$ group is significantly associated with T_{Reg}&T_{Ex}. The same observation holds for subtypes of breast cancer analyzed separately, see Suppl. Fig. S14). It is worth ...

Reviewer: ***8****. The authors should consider analyzing CD4 T-regs independently from Exhausted CD8+ T-cells when investigating spatial co-localization with Macrophages (Fig. 5). These are very distinct T-cell subsets that secrete distinct factors and hold unique functions and fates. What is the proportion each cell type in these Macrophage niches?*

Reply: We thank the reviewer for this suggestion and we agree that these two cell types differ. Unfortunately, in the annotation by Dannenberg et al, these cells are put together.

Reviewer: ***9****. In Figure 4e, a very small number of patients with Mac=0 (n=9) is compared to a large number of patients with Mac>0 (n=334). The authors should consider appropriate statistical analyses to look at this cellular correlation, that might include stratification based on a mean or median, or some continuous metric.*

Reply: We appreciate the concern regarding the imbalance in the data. However, our primary objective was to investigate whether the mere presence of these cells influences immune suppression, and thus we chose this threshold intentionally. We believe that changing it would alter the conceptual framework of our analysis and compromise our effort to examine the specific causal relationship under investigation.

Reviewer: ***10****. Many different patient cohorts and types of omics data are being re-analyzed in this study. It is hard to interpolate which cohorts are being used from the current manuscript text, figures, and figure legends.*

Reply: Thanks for this suggestions. We have clarified cohort usage throughout the manuscriptand a more descriptive text was added to legends and manuscript.

Reviewer: *(Remarks on Code Availability)* The GitHub repository contains excellent and detailed documentation of the various R and Python code used in this study.

Reply: We are pleased that the documentation was found useful.

COMMENTS FROM REVIEWER #4

Reviewer: *We want to provide a technical review of the machine learning and computational aspects of the study. Addressing these points will significantly strengthen the study's rigor and reproducibility. Expansion of Predictive Models: The current study evaluates the performance of Random Survival Forests (RSF) and Survival Support Vector Machines (SSVM). However, comparing these models with additional survival analysis methods, such as the Cox Proportional Hazards Model and Gradient-Boosted Models is essential. This comparison would provide a more thorough assessment of the proposed approach.*

Reply: We deeply appreciate the detailed feedback on our work. We acknowledge this suggestion and have expanded our analysis accordingly. It should be noted that Gradient Boost Machine (GBM) for survival was initially included in our models but was dropped later on. The updated version, includes both comparison to GBM and Cox Proportional Hazards Model. The following sections were added to SI:

Model Selection

In this study, three popular modeling options were selected for survival analysis: Gradient Boost Machine (GBM) for survival, Random Survival Forest (RSF), and Survival Support Vector Machine (SSVM). Hyperparameter optimization was carried out using GridSearchCV for each of the models.

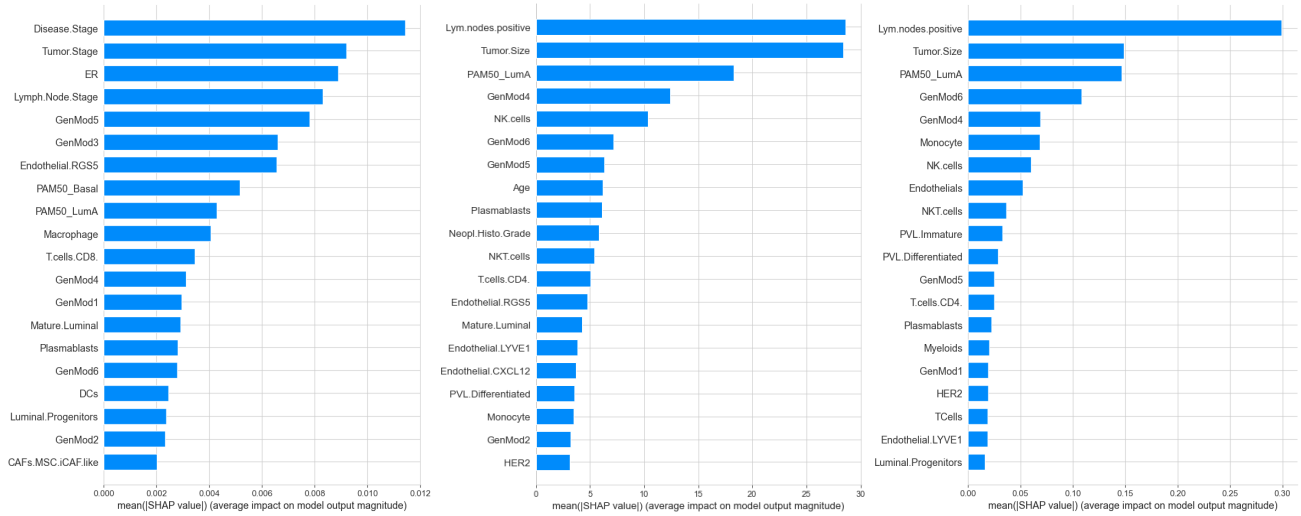
Comparison to Classical Cox Proportional Hazards Model

In order to assess the performance of the machine learning models, a classical Cox Proportional Hazards model was also considered. The Cox model was tuned using GridSearchCV with a dynamic parameter grid for the regularization parameter `alpha`, defined as 1.5^j for $j = -10, \ldots, 9$. For MBRC, the optimal value obtained was $\alpha = 38.4$, which resulted in a best cross-validated concordance index of 0.65. When the final model was evaluated on the test set, it achieved a C-index of 0.64. In TCGA, the same optimal value was obtained for α, resulting in a best cross-validated C-index of 0.7 and C-index of 0.67 on test set.

Model specific bias

SHAP analysis was conducted on the three models to assess the contribution of each feature to the predictions. As anticipated, the SHAP values differ among models, with the same feature sometimes receiving different contributions depending on the model. This variation is reflective of model-specific biases inherent to each algorithm. In other words, each model captures the relationships in the data differently, owing to differences in their assumptions, structure, and capacity to model non-linear interactions—which in turn affects how feature importance is calculated. These differences underscore that while SHAP values offer valuable insights into feature contributions, their interpretation must account for the context of the specific model, as the inherent bias in each machine learning method can influence the resulting values.

To account for the variation in SHAP values across different models, SHAP analyses were performed separately on each dataset (TCGA and MBRC) using models that provided meaningful feature contributions. Initially, it was observed that GBM tended to assign very small contributions to many features (see Fig. 9), which made its SHAP values impractical for our objectives. Consequently, GBM was excluded from the cross-validation process. Instead, the remaining two models were utilized for cross-validation to mitigate model-specific bias and to ensure a robust assessment of feature importance across datasets.



(a) SSVM (b) RSF (c) GBM

Figure 9: SHAP values from different models for MBRC dataset.

The main text was also amended as follows:

In each group of samples, existing data include clinical features such as age, tumor stage, tumor size and cell fractions calculated using deconvolution. To predict the risk of relapse, we considered Random Survival Forests (RSF) [3], Survival Support Vector Machines (SSVM) [4] and Gradient Boost Machine (GBM) for survival [**?**] were considered. To assess the association of each cell type to survival probability, we first calculated SHAP values [5] for all cell types in each model. GBM was dropped due to its tendency to set very small SHAP values for a majority of features (see SI and Suppl. Fig. S3 for more details). For cell types that show a clear association between SHAP values and cell fractions (judged based on a linear fit; see Fig. **??**), we checked if the association was similar across the two remaining models, RSF and SSVM. For such cell types, we calculated the average slope of the SHAP value versus cell fraction, normalized it and obtained the Survival Score. Finally, cell types were ranked based on their Survival Scores. As the final validation, we compared Survival Scores for MBRC and TCGA and excluded cell types that did not show similar scores. In other words, although all cell types and available clinical features are included in the ML model training, a cell type is assigned a Survival Score only if it shows a consistent association with the risk of relapse in both models. (See Suppl. Fig. S4 for an example of how cell types are excluded at different steps.) A cell type is then included in the discussion only if its Survival Scores show the same sign across both TCGA and MBRC.

Reviewer: *The authors utilized two machine learning models (RSF and SSVM) to predict relapse risk. Although the distinctive characteristics of RSF and SSVM may help minimize model selection bias, concerns persist regarding whether this bias was adequately addressed. As illustrated in Figure 1, Step 2, certain clinical variables were excluded from survival score calculations due to inconsistent slope associations between the two models, which could lead to information loss. The purple clinical variable lacks a calculated survival score because of the differing slope associations between the two models. Additional analyses and discussions should be included.*

Reply: We appreciate the reviewer's concern. Our approach was designed to deliberately mitigate model-specific bias through cross-validation between RSF and SSVM. By only assigning a survival score to cell types whose slope associations were consistent across both models, we ensured that only robust and reproducible features were carried forward in the analysis. While we agree this criterion did result in some loss of information (e.g., the purple clinical variable), we believe that this trade-off is acceptable to prevent unreliable predictions. In our view, focusing on features with consistent behavior across models strengthens the reliability of Survival scores. We have added a new section to SI (see below) that showcases how many cell types are dropped, clarifying the role each filtering plays in our pipeline.

Implementation of our pipeline

Using MBRC as an example and the same color scheme as Fig. 1, we show how the pipeline works in Fig. 10. Out of 28 cell types, 4 do not show a clear correlation (green). When comparing the two models, 2 more cell types have opposite associations and are dropped (purple). The remaining 22 cell types, which show either a positive (blue) or negative (red) association, are used to calculate survival scores.

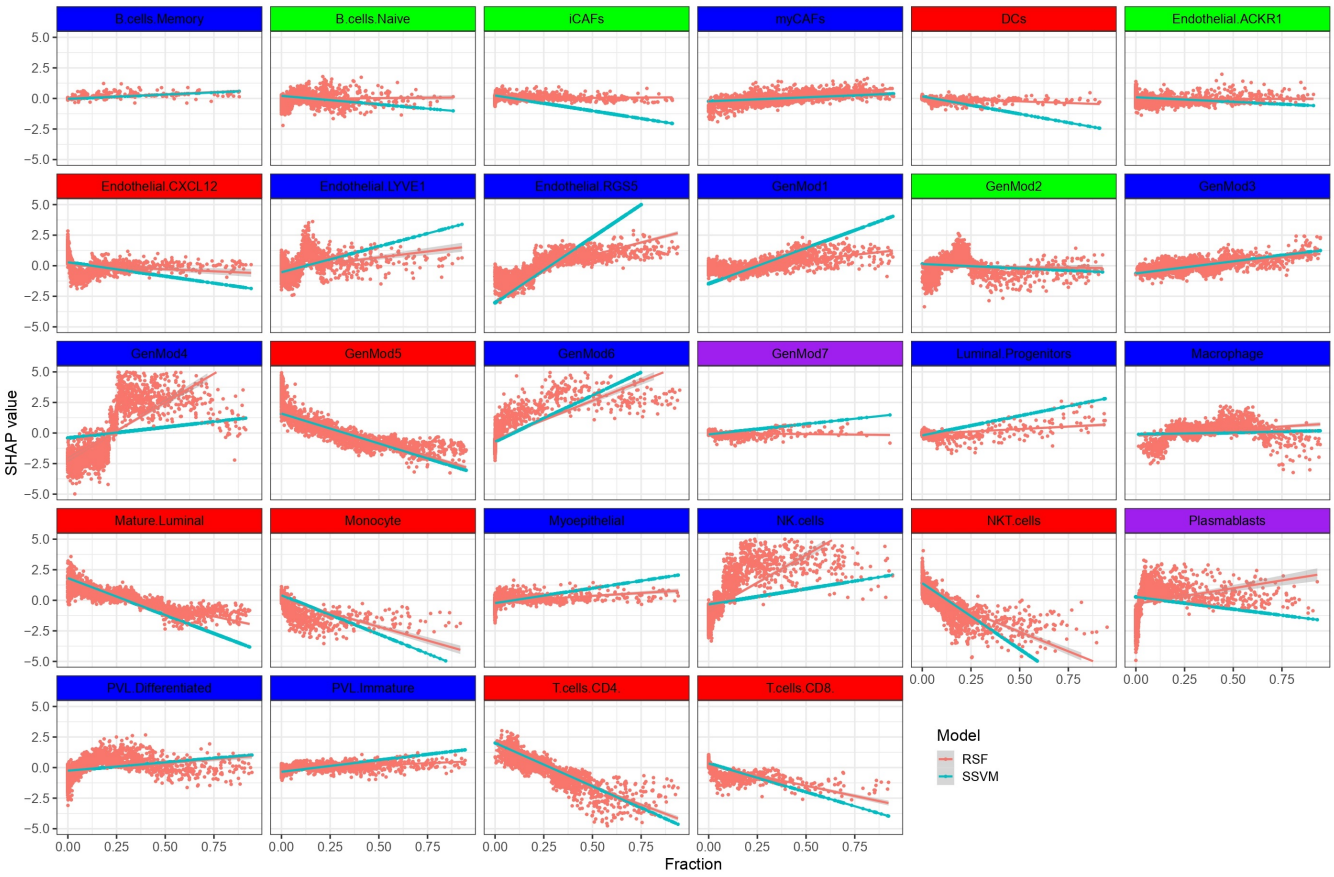


Figure 10: SHAP values for SSVM and RSF vs normalized cell type fractions, following the same color scheme introduced in Fig. 1. Some cell types are filtered out (green and purple) and for the remaining cells with positive (blue) or negative (red) association we calculate survival scores.

Reviewer: *Hyperparameter Optimization Details: The manuscript mentions the use of GridSearchCV for hyperparameter tuning. However, it is unclear which hyperparameters were considered for each model. Please provide a detailed list of the hyperparameters included in the search space and their respective ranges. Additionally, report the optimal hyperparameter values selected for each model to enhance transparency and reproducibility.*

Reply: We thank the referee for highlighting this aspect in our manuscript. We have expanded the discussion on hyperparameter tuning, specifying the chosen parameters and their respective ranges. The text was changed accordingly:

Hyperparameter Tuning using GridSerachCV

MBRC

For RSF, the grid search was performed over `n_estimators` in the range 50 to 450 (increments of 50), `min_samples_split` as 5, 10, 15, and 20, and `min_samples_leaf` from 3 to 27 (increments of 3). The optimal performance was achieved

consistently with 50 estimators and a `min_samples_leaf` of 18, while the `min_samples_split` varied among 5, 10, 15, and 20. All these configurations resulted in a C-Index of 0.69 ($\pm$ 0.01).

For SSVM, the parameter grid was defined as `alpha` as 2.2^i (with $i = -20, \ldots, 0$), `max iter`$=j$ as (with $j = 0, \ldots, 6$), and `tol` as 2.2^k with k (with $k = -10, \ldots, 0$). The optimal parameters were an `alpha` of 7.8×10^{-5}, `max_iter` equal to 2, and a `tol` of 3.8×10^{-4}, yielding a C-Index of 0.683.

For GBM, the hyperparameters were tuned by varying `n_estimators` as $10+30i$ (with $i = 1, \ldots, 5$), `learning_rate` as 1.8^j (with $j = -8, \ldots, -1$), and `max_depth` as k (with $k = 1, \ldots, 10$). The best model obtained had 100 estimators, a learning rate of 0.095, and a maximum depth of 2, resulting in a C-Index of 0.70 ($\pm$0.02).

TCGA

For RSF, the grid search was performed over `n_estimators` in the range 50 to 450 (increments of 50), `min_samples_split` as 5, 10, 15, and 20, and `min_samples_leaf` from 3 to 27 (increments of 3). The optimal performance was achieved with 150 estimators and a `min_samples_leaf` of 18, while the `min_samples_split` could be any of 5, 10, 15, or 20. All of these configurations resulted in a C-Index of 0.676 ($\pm$0.041).

For SSVM, the parameter grid was defined as `alpha` in 3^i (with $i = -23, \ldots, -3$), `max_iter` in $\{0, 1, 2, 3\}$, and `tol` in 4^j (with $j = -5, \ldots, -1$). The optimal parameters were `alpha` $= 6.27 \times 10^{-7}$, `max_iter` $= 2$, and `tol` $= 9.76 \times 10^{-4}$, yielding a C-Index of 0.731.

For GBM, the hyperparameters were tuned by varying `n_estimators` from 50 to 300 (in increments of 50), `learning_rate` in 1.9^k (with $k = -6, \ldots, -1$), and `max_depth` in $[1, 10]$. The best model obtained had 100 estimators, a learning rate of 0.02, and a maximum depth of 3 leading to C-index of 0.69 ($\pm$ 0.06).

Reviewer: *Mathematical Formulation of SHAP Values and Other Metrics: The explanation of SHAP values is currently qualitative. Please include the mathematical formulation used to compute SHAP values. Similarly, providing explicit mathematical definitions or references to standard formulations would be beneficial for other performance metrics used in the study, such as the concordance index (C-index).*

Reply: Thanks for this suggestion. We have added the following mathematical details to improve clarity:

Evaluation Metric

The concordance index (C-index), also referred to as Harrell's C [6], is a widely used performance measure in survival analysis and risk prediction models. It assesses how well a model's predicted risk scores agree with the actual ordering of events (such as time to failure or death). Intuitively, the C-index represents the probability that, for a randomly selected pair of subjects, the subject with the higher predicted risk experiences the event earlier than the subject with the lower predicted risk.

Consider a dataset with n subjects, where each subject i is characterized by $(t_i, \delta_i, \hat{f}(x_i))$ where: t_i is the observed time (which could be either the event time or the censoring time), δ_i is an indicator variable such that $\delta_i = 1$ if the event is observed and $\delta_i = 0$ if the observation is censored, $\hat{f}(x_i)$ is the predicted risk score for subject i. For such a configuration, the C-index is defined mathematically as [7]:

$$C = \frac{\sum_{i,j} \mathbb{1}\left\{\hat{f}(x_i) > \hat{f}(x_j)\right\} \cdot \mathbb{1}\{t_i < t_j\}}{\sum_{i,j} \mathbb{1}\{t_i < t_j\}} \tag{1}$$

where: $\mathbb{1}\{\cdot\}$ is the indicator function, which equals 1 if the condition is true and 0 otherwise. The summation is over all comparable pairs (i, j) such that subject i experiences the event before subject j (i.e., $t_i < t_j$). The numerator counts the number of pairs for which the predicted risk scores are correctly ordered (i.e., the subject with the earlier event has a higher risk score). Denominator represents the total number of comparable pairs. A C-index value of 1 indicates perfect concordance, meaning the model perfectly ranks all subjects according to their risk. A value of 0.5 suggests that the model's predictions are no better than random guessing.

MODEL INTERPRETATION AND SHAP ANALYSIS

After developing models predicting risk of relapse, our focus to understanding the specific roles of different cell types in driving these predictions. To evaluate the contribution of individual features, we applied SHAP (SHapley Additive exPlanations) [5], a post-hoc explanation method that computes the marginal contribution of each feature to the model's output on a per-sample basis. In our analysis, a positive SHAP value for a given feature indicates its propensity to elevate the likelihood of relapse, while the absolute magnitude of the SHAP value serves as an indicator of the feature's overall importance.

Theoretical Foundation

SHAP is one of the most widely used explainable machine learning (XML) approaches, grounded in the concept of Shapley values from cooperative game theory [8]. Named after Lloyd Shapley, who was awarded the Nobel Memorial Prize in Economic Sciences in 2012, the Shapley value fairly distributes the total "gain" (or contribution) among all participating features. For a given feature i, the SHAP value is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right) \tag{2}$$

In this equation, $f_{S \cup \{i\}}$ denotes the model trained with feature i included, while f_S is the model trained with feature i withheld. The vector x_S represents the values of the input features in the subset S. By computing the contribution of feature i across all possible subsets $S \subseteq F \setminus \{i\}$, SHAP assigns a fair share of importance to each feature based on its influence on the prediction.

Reviewer: *While SHAP value-based model interpretation is valuable for assessing the independent contributions of each cell type, it may miss the complex interactions among cell types within the tumor microenvironment. For instance, interactions between macrophages and T cells play a crucial role in regulating immunosuppression and anti-tumor immune responses. Thus, it is necessary to directly analyze the effects of interactions between cell types or enhance the analysis by including interaction terms in the model. Furthermore, it is essential to compare the analysis results that account for cell-to-cell interactions with those based on existing SHAP values to evaluate the impact of these interactions on the results quantitatively.*

Reply: We appreciate the reviewer's valuable suggestions regarding the potential limitations of SHAP value-based interpretations when it comes to capturing higher-order or synergistic interactions among cell types in the tumor microenvironment. Indeed, interactions between macrophages and T cells, among others, are crucial to shaping immunosuppressive and anti-tumor responses. We agree that these complex relationships may not be fully captured by methods focusing primarily on individual feature contributions.

SHAP offers inference of pairwise interactions and we initially explored such interactions to disentangle these effects; however, correlations among specific cell types complicated attempts to ascertain truly independent interaction effects. Additionally, resulting interaction-effect estimates often lacked sufficient stability and reproducibility to confidently support strong biological conclusions.

Despite these challenges, we fully acknowledge that analyzing cell-to-cell interactions is essential to a complete understanding of tumor immunology. Accordingly, we have supplemented our analyses by leveraging cell-cell interaction assays in scRNA-seq data and conducting conditional survival analyses in IMC data.

Reviewer: *The validation of SHAP values used to represent the association of each cell type with survival probability appears inadequate. An ML model was created to predict the risk of relapse using clinical variables for each cell type (Figure 1, Step 1). This ML model achieved a C-index performance of approximately 0.7, which is difficult to regard as robust predictive performance compared to other models with similar objectives (https://pmc.ncbi.nlm.nih.gov/articles/PMC11483452/), given that 0.5 indicates random prediction in C-index characteristics. Furthermore, cross-dataset validation was not conducted. When an ML model's performance is lacking, it becomes hard to trust the explanatory power of features in relation to the output, as there may be noise or critical signals that are absent.*

Reply: We appreciate the reviewer's concerns regarding the use of SHAP values in conjunction with a model that has a C-index of 0.7, as well as the lack of cross-dataset validation. We agree that a robust predictive model is essential for trustworthy interpretation and feature attribution.

We would like to emphasize that a C-index of approximately 0.7 remains competitive within the context of relapse-free survival prediction. For reference, similar studies employing Cox proportional hazards models, RSF or other ensemble methods often report C-indices in the range of 0.65–0.73. Therefore, while we acknowledge that 0.7 does not constitute a perfect prediction rate, it is sufficiently strong—and comparable to published work—to suggest that our model captures meaningful signals associated with relapse risk.

Additionally, while the absolute performance metric is indeed a factor in trusting feature importance explanations, we have ensured that the SHAP values reflect consistent signals across two cross validation steps. In other words, we confirm that the same cell-type features are repeatedly identified as influential for relapse risk for different models in different datasets. For macrophages in particular, we validate the results for survival on IMC data using classical statistics. This consistency, combined with the reasonable predictive performance of our model, lends credibility to our SHAP-based insights. Nonetheless, we concur that additional external validation and direct comparison to other established approaches on the same dataset would further strengthen confidence in the results.

A comparison with models on the same data set shows that our models are equally good in ranking risk of relapse and the following section was added to SI:

Comparison to existing ML models

The obtained C-index of approximately 0.7 indicates a robust discriminative ability in predicting relapse-free survival. This level of performance is in line with—and in some cases exceeds—that of many existing models [9]. Classical methods such as the Cox Proportional Hazards model typically yield lower C-index values, while ensemble approaches like RSF and other modern techniques have reported C-index values in a similar range (e.g., 0.65–0.73) in comparable studies. Thus, a C-index of 0.7 is considered reasonably good, reflecting a strong capability to differentiate between patients with varying risks of relapse. It should be noted that prediction of overall survival can lead to better performance [10], mainly due to inclusion of relapse-free survival as a predictive feature in such models.

Reviewer: *Additionally, the justification for using relapse as a criterion for feature selection to investigate the association with the survival probability-related survival rate (RFS) and chemotherapy response (pCR) is still weak.*

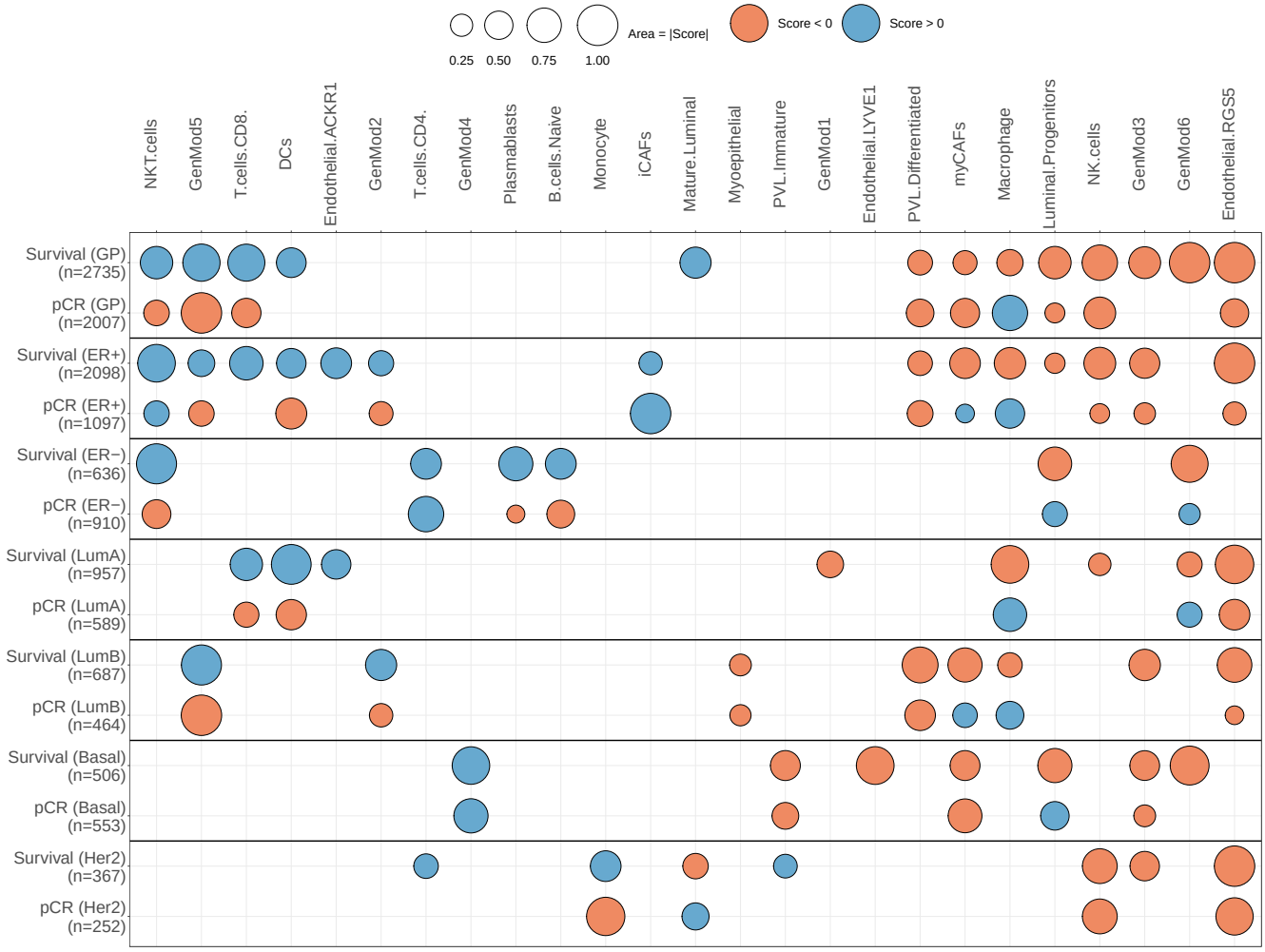
Reply: Relapse free survival is a key clinical outcome. In our work, we used the data provided for local and non local relapse free survival (in MBRC) and disease free survival (in TCGA) as surrogate for relapse free survival. We also use terms such as "risk of relapse" because the RFS data inherently capture time to recurrence (i.e., relapse). By modeling RFS as our endpoint, we essentially estimate how each feature influences the hazard (or likelihood) of experiencing a relapse over time. Thus, while we do not label the event explicitly as "relapse" in the raw data—because it is stored as RFS or DFS—the statistical interpretation treats it as time to relapse, justifying our discussion of "risk of relapse" throughout the manuscript. We now clarify this connection more explicitly to avoid confusion.

Reviewer: *Reproducibility and Code Execution Requirements: To ensure reproducibility, please provide details on the Python version and the specific versions of key libraries used in the experiments (e.g., scikit-learn, etc.). This information can be included in the manuscript or made available in the corresponding GitHub repository. Additionally, providing a requirements.txt or environment.yml file in the repository would facilitate code execution in a consistent environment.*

Reply: We thank the reviewer for their suggestion and have now added the requested information regarding software versions and dependencies in the GitHub repository in a requirements.txt file.

Reviewer: *Data preprocessing and batch effects: For each figure in the results, it's essential to specify the number of data points used and their sources. This helps verify whether the results derive from biased outcomes in a specific cohort or a small dataset. The number of patients per data point will significantly decrease when segmented by subtypes. If the results are based on limited data points, limitations may arise in drawing general inferences. Specifying the number of data points is crucial to ensure the reliability of the results.*

Reply: We appreciate this suggestion and have added more information on number of data samples used and provided source of data being used for each figure by updating the legends. Fig. 3 was updated as follows for increased clarity:



(a)

Figure 11: Comparison of pCR (obtained from NAC cohort) and Survival Scores (obtained from survival cohorts) in GP, ER and PAM50 Subtypes: While pCR and survival scores have the same sign for some cell types, for GenMod5 and macrophages, they consistently show opposite signs.

A section was added to SI that includes number of samples across all data used in this study:

BREAST TUMOR SUBTYPES

Breast cancer is a heterogeneous disease with several subtypes defined by distinct histological and molecular characteristics and prognosis. Main subtypes characterized by the status of Estrogen Receptor (ER), Progesterone Receptor (PR) and Human Epidermal Growth Factor Receptor 2 (HER2) [11], or by their expression profile using the PAM50 method [12]. Most ER+ tumors include the Luminal A (LumA) and Luminal B (LumB) subtypes; these generally respond well to endocrine therapies and have a relatively favorable prognosis. In contrast, ER- tumors, which includes most of Basal-like and HER2-enriched subtypes, tend to be more aggressive, with limited targeted treatment options available. Table S1 includes frequency of these subtypes across all data used in this study (PAM50 Normal-like subtype was not explored in this study).

Data	MBRC	TCGA	NAC	IMC	scRNA-seq
ER+	1347	751	1097	539	26
ER-	417	219	910	142	31
LumA	597	360	589	227	15
LumB	466	221	464	189	8
Basal	310	196	553	111	17
Her2	244	123	252	77	10
GP	1764	971	2007	681	57

Table S1: Number of samples in GP and subtypes of MBRC, TCGA, NAC, IMC and scRNA-seq data.

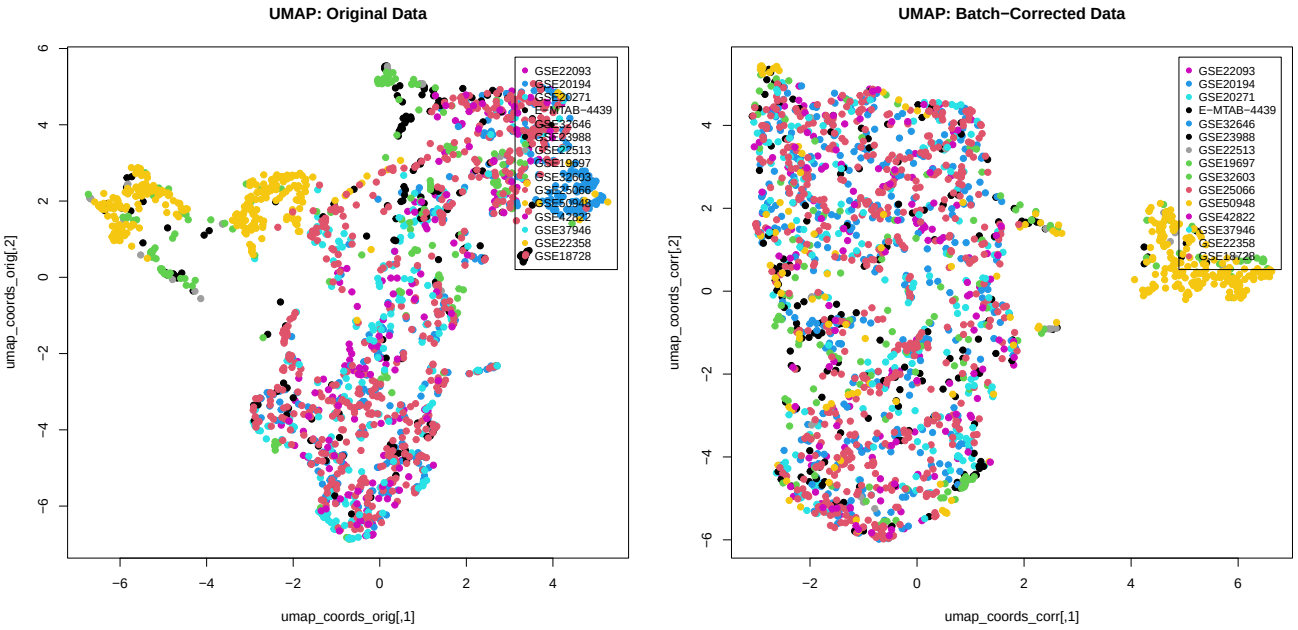
Reviewer: *Neglecting to account for batch effects among cohorts can lead to biased outcomes that favor a specific cohort when merging data from different cohorts in large-scale datasets. The study does not indicate how batch effects were addressed. If batch effects were minimized, this should be elaborated in the methods section. Typically, ComBat or SVA are employed to adjust for batch effects.*

Reply: We appreciate this suggestion and have revised the SI section to clarify our efforts on batch effect correction. The following sections were added to the text:

BATCH CORRECTION

To assess the potential impact of batch effects on NAC cohort, we initially attempted to correct for batch variation using the ComBat algorithm [13] (see for discussion on TCGA and MBRC). For batch correction with ComBat, the fractions matrix was transposed so that rows corresponded to features and columns to samples. A null model design matrix was constructed, and ComBat was applied to adjust for batch effects. The corrected data were then transposed back, ensuring that each row represented a sample.

Uniform Manifold Approximation and Projection (UMAP) was performed on both the original and the batch-corrected data to visually inspect batch-related clustering, as shown in Fig. 12 (a). Additionally, kBET (k-nearest neighbor Batch Effect Test) was used to quantitatively assess batch mixing, with rejection rates compared between the original and corrected datasets (see Fig. 12 (b).



(a)

(b) kBET

Figure 12: UMAP for NAC cohort before and after batch correction using ComBat. and kBET.

Although batch correction with ComBat resulted in a modest decrease in batch-associated metrics (e.g., kBET observed rejection rate decreased from 100% to 80% the improvement was limited. Furthermore, our data represent fractions or proportions, and negative values produced by some batch correction methods would be biologically implausible. Given the limited benefit and the need to preserve the inherent non-negativity and biological interpretability of the data, we opted not to apply batch correction to the final analyses.

To avoid any bias to data sets, NAC cohort was randomly divided the data into discovery (9 datasets, 1009 samples) and validation (6 datasets, 998 samples) subsets. Then, pCR scores from the two subsets were cross validated against each others, keeping only cell types with consistent scores across two subsets. The same approach was applied to all tumor subtypes as well (see [14] for more details).

Dataset specific bias

TCGA and MBRC are two major cohorts in breast cancer studies, and they differ in several ways such as sampling, population composition, and follow-up protocols. In addition, TCGA data comes from RNA sequencing, while MBRC data is generated using microarrays, so some differences are expected. Moreover, the characteristics of the data vary significantly: TCGA has a censoring rate of 90% and MBRC has a censoring rate of 55%, based on data from cBioPortal.

Due to these variations and our aspiration to provide measures that are reliable and work across different data sets, in our analyses we only keep features that show consistent scores. We believe that this filtering is reliable and rigorous as it does compare the internal structures of the models and cross validate the associations captured by models for the two datasets.

[1] E. Danenberg, H. Bardwell, V. R. Zanotelli, E. Provenzano, S.-F. Chin, O. M. Rueda, A. Green, E. Rakha, S. Aparicio, I. O. Ellis, *et al.*, Breast tumor microenvironment structures are associated with genomic features and clinical outcome, Nature Genetics **54**, 660 (2022).

[2] S. Z. Wu, G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy, N. Bartonicek, *et al.*, A single-cell and spatially resolved atlas of human breast cancers, Nature Genetics **53**, 1334 (2021).

[3] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, Random survival forests, (2008).

[4] V. Van Belle, K. Pelckmans, J. A. Suykens, and S. Van Huffel, Survival svm: a practical scalable algorithm., in *ESANN* (2008) pp. 89–94.

[5] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems **30** (2017).

[6] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, Evaluating the yield of medical tests, Jama **247**, 2543 (1982).

[7] M. Schmid, M. N. Wright, and A. Ziegler, On the use of harrell's c for clinical risk prediction via random survival forests, Expert Systems with Applications **63**, 450 (2016).

[8] L. S. Shapley, Notes on the n-person game—ii: The value of an n-person game, (1951).

[9] G. Buyrukoğlu, Survival analysis in breast cancer: evaluating ensemble learning techniques for prediction, PeerJ Computer Science **10**, e2147 (2024).

[10] K. Evangeline I, S. A. Kirubha, and J. G. Precious, Survival analysis of breast cancer patients using machine learning models, Multimedia Tools and Applications **82**, 30909 (2023).

[11] H. G. Russnes, O. C. Lingjærde, A.-L. Børresen-Dale, and C. Caldas, Breast cancer molecular stratification: from intrinsic subtypes to integrative clusters, The American Journal of Pathology **187**, 2152 (2017).

[12] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, Molecular portraits of human breast tumours, Nature **406**, 747 (2000).

[13] W. E. Johnson, C. Li, and A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, Biostatistics **8**, 118 (2007).

[14] Y. Azimzade, M. Haugen, X. Tekpli, C. B. Steen, T. Fleischer, D. Kilburn, H. Ma, E. V. Egeland, G. Mills, O. Engebraaten, *et al.*, Explainable machine learning reveals the role of the breast tumor microenvironment in neoadjuvant chemotherapy outcome, bioRxiv , 2023 (2023).

Response to Reviewers

We are grateful to the reviewers for their constructive feedback. Their valuable insights have enhanced our manuscript and their professional approach made the revision process both productive and pleasant. Below, we address each comment in detail. Changes are in blue for clarity. Corresponding changes are provided here as well (Figure names dont follow that of SI).

COMMENTS FROM REVIEWER #4

Reviewer: The author has addressed all the referees' requests, and the revised version effectively incorporates their feedback.

Reply: We sincerely appreciate the positive feedback and are glad that the revisions have successfully addressed the referees' requests.

COMMENTS FROM REVIEWER #5

PART I: Evaluation of the authors' response to Reviewer 3's comments

PART I, Q1: *Reviewer 3 notes that LumA, an immune "cold" subtype, exhibits low tumor-infiltrating lymphocytes (including Tregs and Tex cells). This casts doubt on the authors' reported correlation between Treg-Tex-macrophage co-localization and clinical outcomes. Although the authors have presented detailed data confirming the presence of Tregs and Tex cells in LumA tumors (Figure S11), this fails to address the core concern. Notably, the authors propose that Tregs and Tex cells exert substantial effects in LumA tumors despite their low abundance, which is markedly lower than in other subtypes. This raises a critical question: why are more pronounced immunosuppressive effects not observed in other subtypes with higher abundance of these cells? Given the low abundance of Tregs and Tex cells in LumA, their attribution of the formation of the immunosuppressive microenvironment to this mechanism is questionable. Consequently, additional evidence (supplemental analyses or literature citations) is recommended to support this mechanism.*

Reply: This is indeed an open question that our manuscript does not fully address. We have, however, now run DEA/PEA on Tregs and Tex cells in luminal vs Basal/Her2 subtypes. These results, combined with existing literature, allow for several hypotheses. The LumA subtype is a well-established immune-cold tumor, typically having a lower abundance of immune cells, including both Tregs and Tex cells. In such a context, the relative ratio of Tregs and Tex to other immune cells may confer a significant immunosuppressive effect. In contrast, in more inflamed subtypes like TNBC, a higher absolute number of immunosuppresive cells would be required to achieve a similar effect (https://pmc.ncbi.nlm.nih.gov/articles/PMC7881184/). Furthermore, functional heterogeneity within immune subpopulations might play a critical role. For instance, the literature shows that individual Tregs can act as potent sources of immunosuppressive cytokines such as TGF- (https://pubmed.ncbi.nlm.nih.gov/31911451), and mechanisms like the CCR6-CCL20 axis can enhance Treg glycolisis and suppressive capacity, enabeling limited number of Tregs to exert disproportionate effects (https://pubmed.ncbi.nlm.nih.gov/39133127). Similarly, T cell exhaustion seems to exist on a continuum rather than a binary state. Even partially exhausted (pre-exhausted) T cells can contribute to immunosuppression (https://pubmed.ncbi.nlm.nih.gov/32659053 and https://pubmed.ncbi.nlm.nih.gov/39257319). Thus, T cells at various stages of dysfunction may collectively impair anti-tumor immunity, even when their overall abundance is low. Finally, Tregs may form immunosuppressive niches around lymphatic vessels in the peripheral tumor stroma, creating localized zones of suppression (https://pubmed.ncbi.nlm.nih.gov/39029466). Nevertheless, further research is required to provide a more comprehensive characterisation of functional differences in the immune tumour microenvironment across these breast cancer subtypes.

We added these parts to SI and main text:

DEA and PEA

We run DEA and PEA across Luminal (LumA + LumB) vs Basal/Her2 to find differentially expressed genes and to reveal pathway-level differences in T_{Reg} and T_{Ex} as shown in Fig. 1.

For T_{Ex} cells (Figs. 1a and b), luminal tumours showed transcriptional signatures consistent with deeper exhaustion. Up-regulated pathways included TNFα/NF-κB, TGF-β signalling, mTORC1, and IL-2/STAT5, alongside increased expression of co-stimulatory molecules (CD81), proteasomal components (UBE2M/UBE2S), and extracellular matrix–associated genes (COL1A2, TNFSF9). Conversely, interferon responses, antigen-presentation genes, and certain TCR variable segments were down-regulated. Together, this pattern suggests that luminal T_{Ex} adopt a metabolically active but functionally impaired state, characterised by reduced effector capacity and enhanced stress adaptation.

For T_{Reg} cells (Figs. 1c and d), luminal tumours displayed marked up-regulation of pathways associated with stress adaptation and suppressive function, including TNFα via NF-κB, IL-2/STAT5 signalling, hypoxia, and the unfolded protein response. Notably, heat-shock proteins (e.g., HSPA1A/B, HSP90AA1, DNAJB1) and metabolic enzymes (e.g., GLS, UBE2F/UBE2S) were over-expressed, suggesting adaptation to a nutrient- and oxygen-deprived microenvironment that may enhance suppressive potency. In contrast, pathways linked to interferon-α/γ responses and antigen presentation were down-regulated, consistent with a shift away from classical immune activation toward a suppressive phenotype. These findings suggest that luminal T_{Reg} exert qualitatively stronger suppressive activity despite being numerically less abundant.

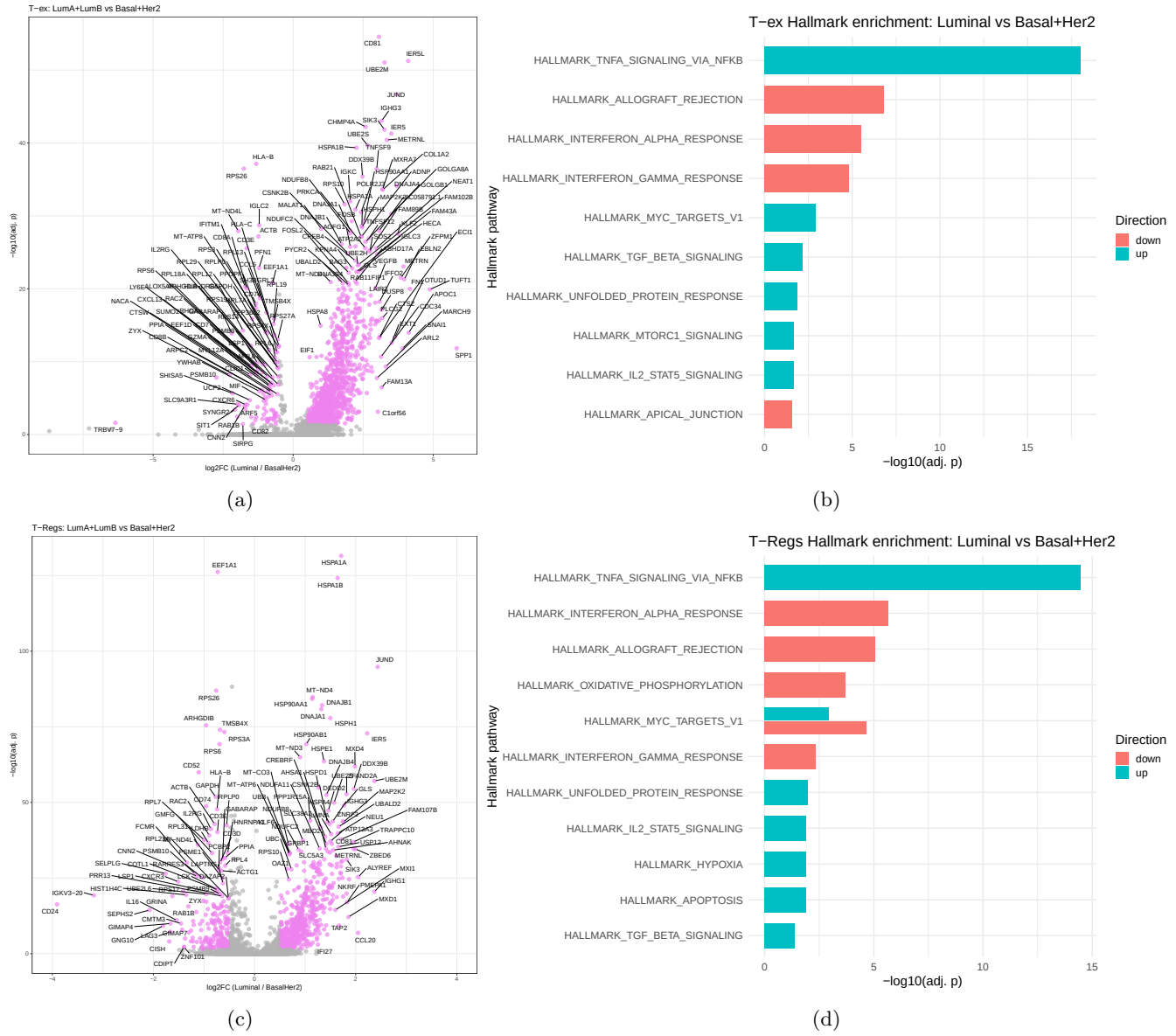


Figure 1: Differential expression (volcano plots) and pathway enrichment (barplots) of T_{Ex} (panels a–b) and T_{Reg} (panels c–d) in luminal versus basal/HER2 tumours.

RESULTS

... Although the cell type proportions differ, LumA samples still contain both T_{Reg} and T_{Ex} cells (see Suppl. Figs S13 and S14). Furthermore, the lower frequency of immune cells—particularly T cells—in luminal tumours does not translate into reduced immune activity. Our DEA and PEA analyses of T_{Reg} and T_{Ex} cells in luminal versus basal/HER2 tumours reveal qualitative differences in function. Specifically, luminal T_{Reg} exhibit up-regulation of NF-κB, IL-2/STAT5, and hypoxia pathways, while luminal T_{Ex} display signatures of deeper exhaustion and stress adaptation (Suppl. Fig. S15).

DISCUSSION

Macrophages: Macrophages in tumor tissues, also referred to as tumor-associated macrophages (TAMs), play a critical role in tumor progression, angiogenesis and metastasis and their association with clinical outcomes is well-stablished [1, 2]. Although luminal tumours are characterised by lower overall immune infiltration, this does not equate to reduced immune activity. Several studies have shown that T_{Reg} and macrophages are relatively enriched in luminal subtypes despite fewer effector T cells, resulting in a disproportionately suppressive microenvironment [3, 4]. Importantly, functional and transcriptional analyses indicate that T_{Reg} and T_{Ex} cells in luminal tumours adopt stress-adapted, metabolically active states with heightened suppressive capacity [5]. This is also consistent with our DEA of T-cell subsets, where both T_{Reg} and T_{Ex} from luminal tumours showed enrichment of suppressive and stress-adapted pathways compared with basal/HER2 tumours. Together, these findings support the view that functional capacity, rather than absolute abundance, shapes immunosuppressive activity. Further research is required to provide a more comprehensive characterisation of functional differences in the immune tumour microenvironment across breast cancer subtypes.

PART I, Q2: *Reviewer 3 posits that Danenberg et al. (2022) have previously demonstrated that the spatial architecture formed by 'APC-enriched' macrophages or Tregs and Tex cells is linked to increased hazard ratios. Consequently, the novelty of the insights reported by the authors need to be clarified. In their response, the authors emphasize the innovative contributions of their study in two key respects: firstly, they have clearly identified the HLA-ABC-high macrophage subset as the functionally relevant population; secondly, they have elucidated the specific impact of the spatial distribution of these three cell types on adverse outcomes. In my assessment, the authors' response effectively underscores the innovative value of their research. Building upon prior work which indicates that spatial niches consist of 'APC-enriched' macrophages or Tregs and Tex cells can form a immunosuppressive microenviroment respectively, their study further elucidates that the spatial colocalization of HLA-ABC-high macrophages, regulatory T cells (Tregs), and exhausted T cells (Tex cells) orchestrates the formation of an immunosuppressive microenvironment, which in turn contributes to polarized clinical outcomes. This work offers a more refined and detailed mechanistic dissection of the underlying processes.*

Reply: We thank the reviewer for acknowledging the novelty of our study.

PART I, Q3: *Reviewer 3 raised concerns regarding the technical limitations from bulk deconvolution methods (CIBERSortX). The authors have acknowledged this technical issue and incorporated an additional statement in the manuscript. Given the widespread application of similar deconvolution approaches within the field, I consider the authors' response to this query to be adequate.*

Reply: Thank you for your understanding. We appreciate your acknowledgment of our response and the broader context of deconvolution approaches in the field.

PART I, Q4: *Reviewer 3 argues that the study needs to further clarify the association between the macrophage subsets identified in previous single-cell studies and the spatially defined macrophages. The authors have supplemented a volcano plot (Figure 5g) indicating that the CXCL10 subset shows enrichment in the IFN response. In my assessment, the authors' response has partially addressed the reviewer's query. However, the authors should provide a more detailed explanation of the definition of HLA-ABC-high/low macrophages. For example: Do "HLA-ABC-high macrophages" refer to macrophages with relatively high expression of all HLA-A, B, and C, or those with high expression of one or two of these genes?*

Reply: Thank you. We have clarified how the division was performed in the revised text as follows:

... In scRNA-seq data, we have three genes for HLA class I: HLA-A, HLA-B, and HLA-C. We define a single value representing HLA-ABC as follows: HLA-ABC = log2(HLA-A + 1) + log2(HLA-B + 1) + log2(HLA-C + 1). Then, we divided all macrophages (10,700 cells across 57 samples) into two groups using this single value ($HLA\text{-}ABC^{lo}$ and $HLA\text{-}ABC^{hi}$).

PART I, Q5: *Reviewer 3 inquired about the methodology by which the authors stratified patients into high and low macrophage groups. The authors have supplemented violin plots (Figure 4c) and explanations to clarify the*

stratification criteria. In my view, the responses provided by the authors are rationale-based.

Reply: Thank you for your feedback. We appreciate your recognition of the rationale behind our stratification approach.

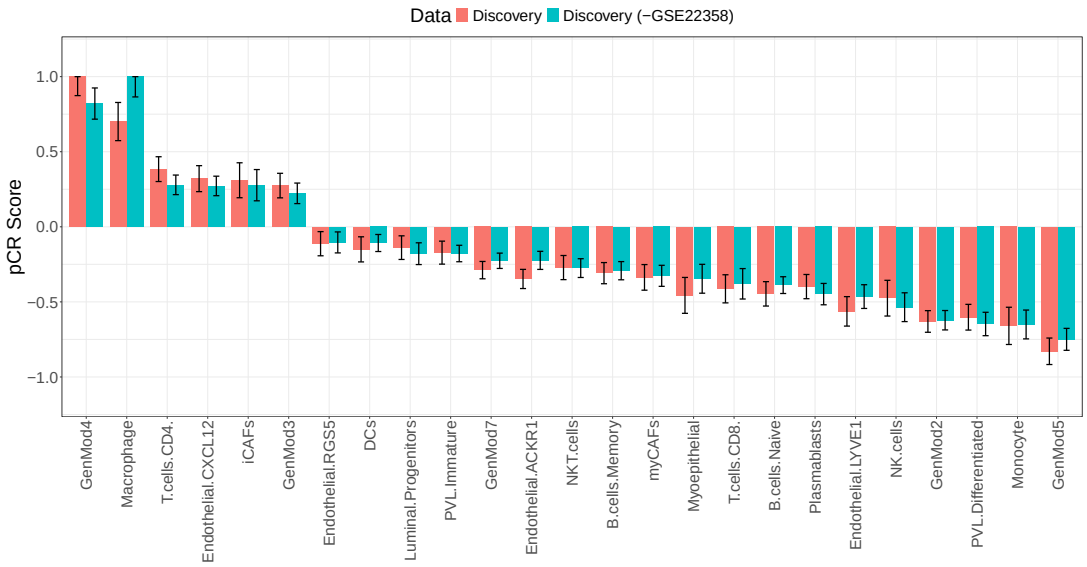
PART I, Q6.1 *Reviewer 3 argues that, given the use of multi-source datasets, the authors should correct batch effects and verify the consistency of efficacy and prognostic assessments across different cohorts. The authors have presented UMAP plots (Figure S1) before and after correction, along with justifications for their decision not to perform batch effects correction. In my assessment, the authors' response is insufficient for the following reasons: 1. In the UMAP plot (Figure S1) provided by the authors, we found that some datasets exhibit significant batch effects regardless of whether batch correction is performed (e.g., GSE22358). The authors should implement measures to tackle this issue such as excluding certain genes in the dataset that induce significant batch effects because of the different detection methodologys (e.g., specific non-coding RNAs) or considering the exclusion of a particular cohort.*

Reply: We thank the reviewer for raising this important point. To directly address the concern regarding GSE22358, which appeared to show residual batch-related clustering in the UMAP plots, we conducted an additional sensitivity analysis. Specifically, we excluded GSE22358 from the NAC discovery cohort, retrained our model, and recalculated the pCR scores both within the discovery set and across the entire NAC cohort. The results obtained after exclusion of this dataset were highly consistent with those from the full analysis, with no drastic changes to pCR scores. . A dedicated section was added to SI as follows:

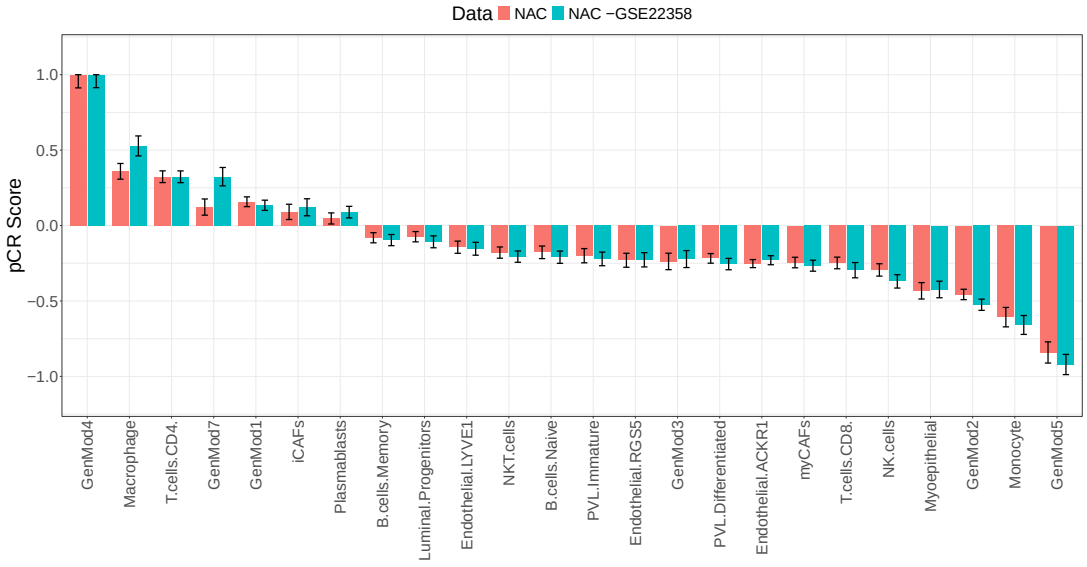
Excluding a dataset (GSE22358)

To further assess the potential impact of batch effects, we performed a sensitivity analysis in which the dataset GSE22358, due to its distincs separation from the rest of data in UMAP plots, was excluded from the NAC discovery cohort. We retrained the model without this dataset and recalculated pCR scores for both the discovery set and the full NAC cohort.

As shown in Fig. 2, the results with and without GSE22358 were highly consistent. These findings suggest that while some individual cohorts may display stronger batch-related signatures, our overall framework—based on discovery/validation partitioning and cross-validation across independent datasets—remains robust. It also should be noted that batch effects may sometimes be driven by specific features (e.g., non-coding RNAs measured with differing methodologies) and that excluding such genes or, in certain cases, excluding problematic cohorts could be a reasonable strategy.



(a) Discovery cohort



(b) Full NAC cohort

Figure 2: Comparison of pCR scores with and without GSE22358. (a) Discovery cohort vs. discovery cohort excluding GSE22358. (b) Full NAC cohort vs. NAC cohort excluding GSE22358.

PART I, Q6.2 *The authors should provide the details of how pathological complete response (pCR) was evaluated in the different neoadjuvant chemotherapy (NAC) cohorts they used, such as specifying whether the Miller-Payne (MP) grading system or the residual cancer burden (RCB) classification was employed.*

Reply: We thank the referee for raising this important point. To clarify, we have added a new supplementary table (Table S1) summarizing for each NAC cohort how pathological complete response (pCR) or treatment response was defined and evaluated, alongside other features. The following section was added to SI:

NAC COHORT

Inclusion criteria were breast tumors that underwent NAC, with available bGEP microarray data and information on treatment response. We identified a total of 15 datasets, which are presented in Table S1, along with treatment regimen, subtypes, and outcome evaluation. The majority of datasets relied on either the Residual Cancer Burden (RCB) classification or the standard binary definition of pCR (absence of residual invasive cancer in breast and axillary lymph nodes, with carcinoma in situ permitted). In some cases, studies applied a study-specific responder vs non-responder classification that combined clinical examination (tumor shrinkage or RECIST criteria) with pathologic findings at surgery. Notably, none of the included cohorts employed the Miller–Payne (MP) grading system.

Datasets								
Expression Data Access Number /Reference Paper for Results	Information / Name and Clinical Trial Number	Number of Samples	Stage	Treatment	Her2	ER	Outcome Measurement	Cohort
E-MTAB-4439 silwal2017 [6]	Randomized, Phase II clinical trial for Avastin in combination with neoadjuvant treatment (NeoAva) NCT00773695	131	< IV	Fluorouracil, Epirubicin and Cyclophosphamide, Taxanes (Docetaxel or Paclitaxel) + Bevacizumab	Neg	Mix	RCB	Disc
GSE18728/ korde2010 [7]	Phase II trial with protocol available at lebowitz2004 [8]	61	II/III	Docetaxel and Capecitabine	Neg	Mix	study-specific responder / non-responder classification	Disc
GSE19697 lin2010 [9]	Prospective clinical trial at Washington University (WASH2) with protocol available at lin2010 [9]	24	-	Epirubicin and Docetaxel+ Zoledronic Acid	Neg	Neg	pCR vs RD	Disc
GSE20194 popovici2010 [10]	Prospective pharmacogenomic marker discovery study with corresponding protocol available at shi2010 [11]	278	I-III	Paclitaxel, 5-fluorouracil, Cyclophosphamide and Doxorubicin	Mix	Mix	pCR vs RD	Disc
GSE20271 tabchy2010 [12]	Prospective, randomized, international clinical trial with protocol available at tabchy2010 [12]	178	I-III	Paclitaxel, 5-fluorouracil, Cyclophosphamide and Doxorubicin	Mix	Mix	pCR vs RD	Disc
GSE22093 iwamoto2011 [13]	MDACC/IGR dataset with protocol available at iwamoto2011 [13]	97	I-III	Fluorouracil, Adriamycin, and Cytoxan	Nor	Mix	pCR vs RD	Disc
GSE22358 gluck2012[14]	A multicenter, open-label study on treating operable early-stage breast cancer with further information available at gluck2012[14]	122	I-III	Capecitabine, Docetaxel and Trastuzumab	Mix	Mix	RCB	Disc
GSE42822 shen2012 [15]	Phase II trial with protocol available at shen2012cell [15]	90	II-III	5-fluorouracil, Epirubicin, Cyclophosphamide; Docetaxel Capecitabine	Mix	Mix	pCR vs RD	Disc
GSE22513 (VAN) bauer2010 [16]	High-risk, operable breast cancers treated at Vanderbilt University or New York University with the protocol available at: bauer2010 [16]	28	II -III	Paclitaxel + Radiation	Mix	Neg	pCR vs RD	Disc
GSE25066 hatzis2011 [17]	Prospective multicenter study with protocols according to hatzis2011 [17]	482	II -III	Paclitaxel, Docetaxel with Capecitabine	Mix	Mix	RCB	Valid
GSE32603 magbanua2015 [18]	Clinical trial to study magnetic resonance imaging (I-SPY1) with clinical trial number of NCT00033397	138	< IV	Taxane	Mix	Mix	RCB	Valid
GSE32646 miyake2012 [19]	Primary breast cancer patients at Osaka University Hospital (Osaka) miyake2012 [19]	115	II -III	Paclitaxel, 5-fluorouracil, Epirubicin and Cyclophosphamide	Mix	Mix	pCR vs RD	Valid
GSE37946 liu2012 [20]	HER2+ patients received treatment at the MDACC (details here: liu2012 [20]).	50	II -III	Fluorouracil/ Epirubicin or Adriamycin/ Cyclophosphamide - Taxol plus Trastuzumab	Pos	Mix	pCR vs RD	Valid
GSE50948 prat2014 [21]	NOAH trial (randomized clinical trial) ISRCTN86043495 and control cohort of 42 HER2- tumors	156	< IV	Doxorubicin/Paclitaxel, Cyclophosphamide Methotrexate, Fluorouracil+ Trastuzumab, AT-CMF for control cohort	Mix	Mix	pCR vs RD	Valid
GSE23988 iwamoto2011 [13]	Phase II trial with details available at iwamoto2011 [13]	61	II -III	5-fluorouracil, Doxorubicin, Cyclophosphamide, Docetaxel and Capecitabine	Mix	Mix	pCR vs RD	Valid

Table S1: NAC cohort trials and their corresponding bGEPs datasets.

PART I, Q7: *Reviewer 3 suggested that the authors should verify whether the co-localization of macrophages, Tregs, and Tex cells is universally present across various PAM50 tumor subtypes. The authors have provided evidence (Figure S13) in response to this comment. This response is deemed reasonable and adequate.*

Reply: Thank you for your comment. We appreciate you finding our response and the evidence provided to be adequate.

PART I, Q8: *Reviewer 3 argues that, given the significant functional differences between Tregs and Tex cells, these two cell subsets should be analyzed separately. The authors responded that these two subsets were grouped together in the annotations from Dannenberg et al.'s work. In my assessment, this response is highly illogical: due to the substantial differences between these two cell subsets, they clearly warrant separate analysis. The authors should consider using a dataset which can analyze Tregs and Tex cells separately.*

Reply: Following your recommendation, we have now conducted a separate analysis for T_{Reg} and T_{Ex}. In brief, separating the two populations does not change our main conclusions regarding the the exclusive correlation of HLA-ABC^{hi} and T_{Reg} and T_{Ex}. The follwoing parts were added to the main text and SI.

MAIN TEXT

Additionally, in samples with MHC I & II^{hi} present, only those with macrophages also have T_{Reg}&T_{Ex}, as shown in Fig. **??** (e). It should be noted that T_{Reg} and T_{Ex} belong to different lineages but given the annotation provided by the source paper [22], they were used together. We have separated these cells based on their distinct markers and performed the analysis again and showed that HLA-ABC^{hi} macrophages are exclusively associated with T_{Reg} and with T_{Ex}, see Suppl. Fig. S18 .

SI

Further stratification of T_{Reg}&T_{Ex}

In Danenberg *et al.* [22], T_{Reg} and T_{Ex} were not analyzed separately. Here, we have now re-examined the data by explicitly separating the two populations. We stratified the combined T-cell compartment by FOXP3 and PD-1 expressions: among T_{Reg}&T_{Ex} cells, top 33% with FOXP3 labeled as T_{Reg}. Among the remaining 67%, cells with PD-1 > median were treated as T_{Ex}. This annotation led to FOXP3^{hi} as T_{Reg}, and FOXP3^{lo} PD-1^{hi} as T_{Ex} (FOXP3^{lo} PD-1^{lo} remain unlabeled). This is an effort to annotate cells with high confidence without complicating the analysis and should be seen as support for our main results. We then re-computed, the correlations between the frequencies of HLA-ABC^{hi}/HLA-ABC^{lo} macrophages and each T-cell subset separately in all subsets of tumors. As Fig. 3 shows, HLA-ABC^{hi} macrophages are exclusively correlated with both T_{Ex} and T_{Reg}.

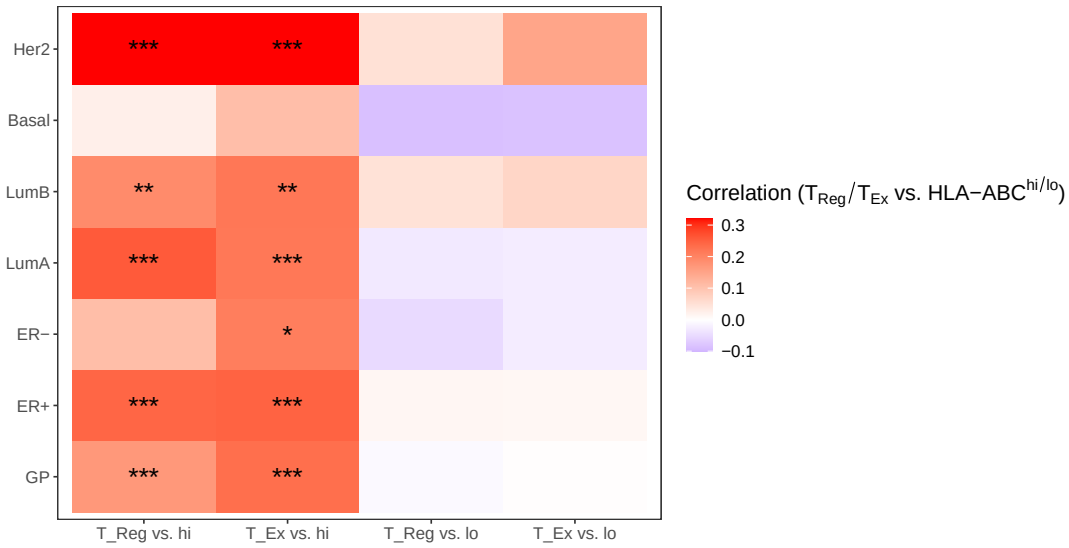


Figure 3: Correlations between the frequencies of T_{Ex} or T_{Reg} and HLA-ABC$^{hi/lo}$ macrophages across GP and tumor subtypes.

It should be noted that because cell-state annotation is inherently complex and context-dependent—and FOXP3/PD-1 can be transiently expressed—this FOXP3/PD-1-based separation is operational rather than definitive. We applied it to a limited set of cells to obtain an initial view of how these T-cell subsets associate with macrophage phenotypes.

PART I, Q9: *Reviewer 3 contends that the authors employed inappropriate statistical analyses to demonstrate the differences in Tregs and Tex cells between the MAC0 and MAC>0 groups. The authors maintain that their chosen analytical methods and presentation format are reasonable. I align with the authors' perspective, as the current analysis and presentation format more effectively convey the intended meaning of the authors' work.*

Reply: Thank you for your feedback. We appreciate your support of our chosen analytical approach and presentation format in addressing this aspect of the analysis.

PART II, Other issues in this manuscript,

Q1: *It is generally recognized that Luminal breast cancer, particularly LumA subtype, exhibits a low pathological complete response (pCR) rate to neoadjuvant chemotherapy but demonstrates favorable overall prognosis following standard adjuvant endocrine therapy. This dichotomous clinical outcome may be attributed to the fact that tumors with higher differentiation exhibit immune-cold characteristics, resulting in a lower pathological complete response (pCR) rate. Meanwhile, the higher degree of differentiation is also associated with reduced tumor malignancy and more favorable prognosis. Therefore, this dichotomous clinical phenomenon may not necessarily be driven by the tumor microenvironment (TME) but could also stem from the intrinsic biological properties of the tumor itself. The authors have presented the efficacy or prognostic value of different immune cell types across various tumor subtypes in Figure 2; however, many cell types were not displayed due to failure to meet the threshold. We would like to see the complete edition to verify the specificity of the association between macrophages and this dichotomous clinical phenotype, as other types of immune cells may also exhibit similar clinical correlations to macrophages.*

Reply: We thank the referee for this very helpful comments. We have done the followings to address it:

a) We have updated the Suppl. Fig. 6 (then 5) to include all scores, regardless of the conditions. This leads to plot including all pCR and Survival scores calculated in this study.

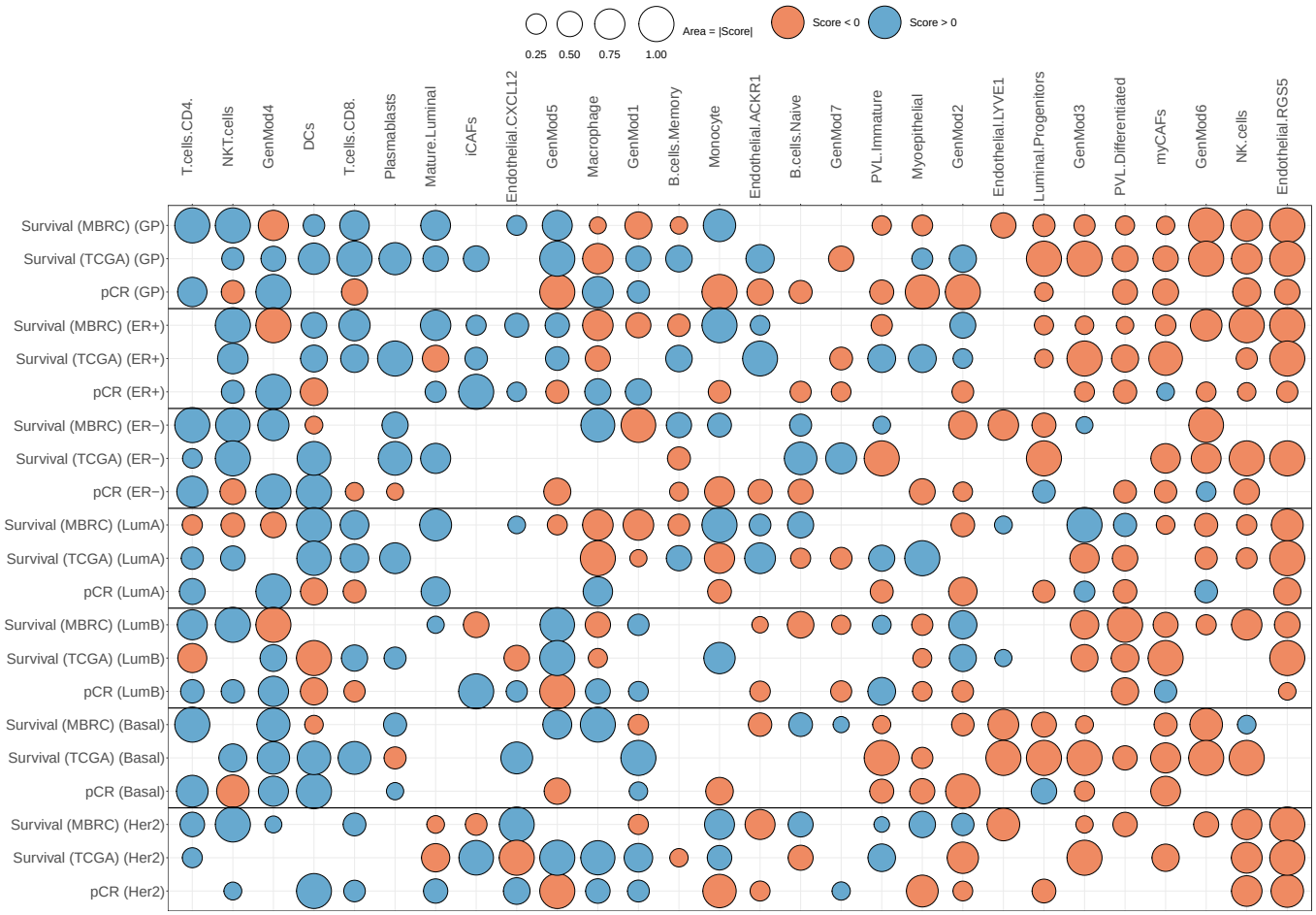


Figure 4: Survival Scores for subtypes in MBRC and TCGA datasets, alongside pCR scores.

b) We agree that while TME is reflective of tumor biology, its note the sole driver factor in tumor response to treatment and progression. As such, we added a new paragraph describing the limitations of TME.

LIMITATIONS OF THIS STUDY:

...

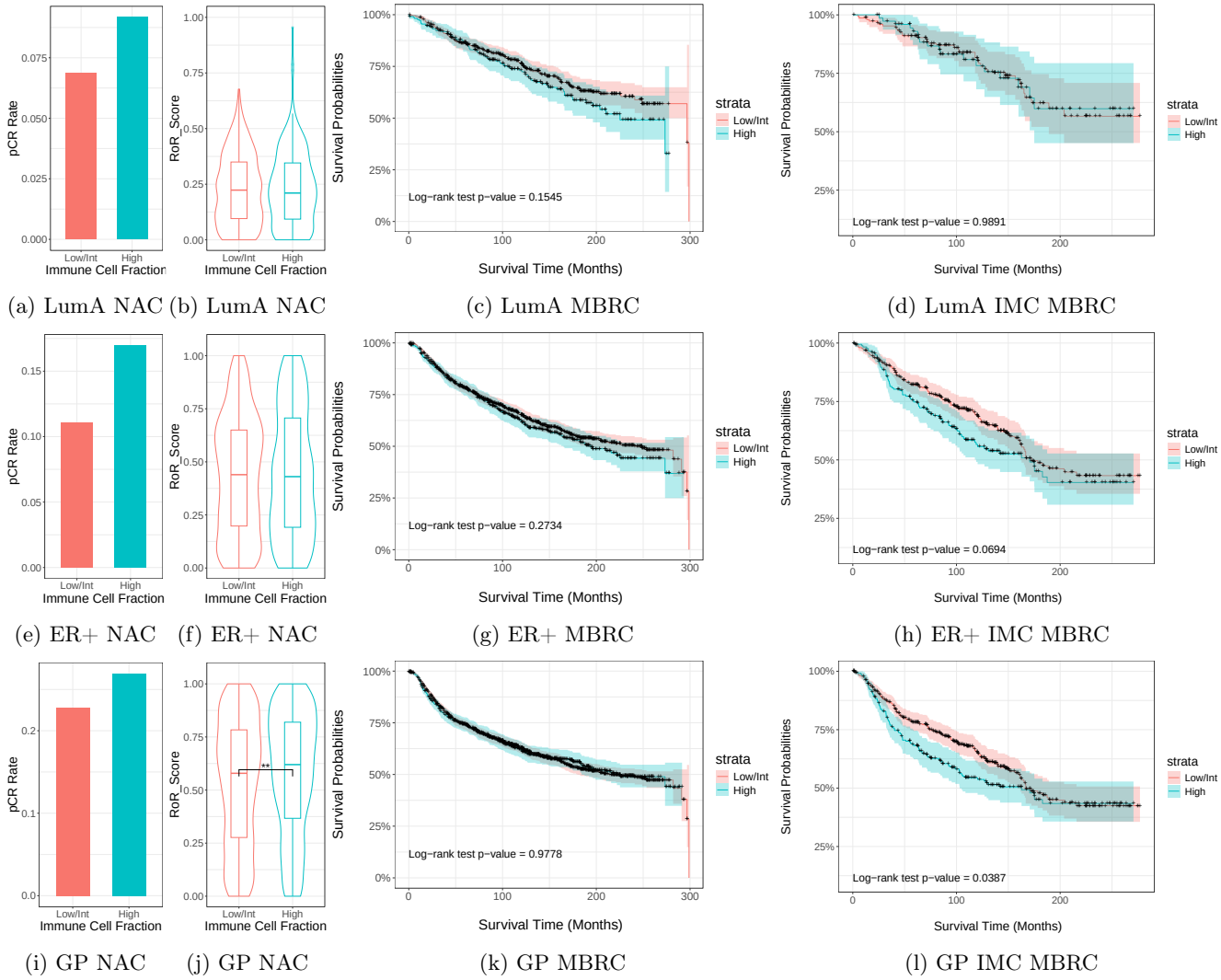
It is important to acknowledge that TME is not the sole determinant of clinical outcomes in breast cancer. Dichotomous phenomenon presented in this paper may not be exclusively TME-driven, but could also reflect intrinsic tumor characteristics such as cellular differentiation status, proliferative capacity and genomic features. We did not explicitly investigate the contribution of genomic alterations, transcriptomic heterogeneity or other tumor-intrinsic biological properties, which may independently or synergistically influence both immune infiltration and patient outcomes. Future integrative studies will be necessary to disentangle these interconnected determinants and clarify the specificity of macrophage-related effects relative to tumor-intrinsic factors.

c) We have explored if immune infiltration (presence of all immune cells and not just macrophages) stratifies breast cancer patients. Thus we added a dedicated section to SI analyzing this hypothesis. In both IMC data and deconvolution data, we show that immune infiltration does not lead to different clinical outcome in LumA and ER+ sunbsets of data. The following section was added to the SI:

IMMUNE INFILTRATION AND CLINICAL OUTCOMES

While we showed that Macrophage infiltration levels leads to distinct clinical outcomes, here we explore how immune infiltration, defined by fraction of all immune cells affects clinical outcomes. In deconvolution, immune cells include: DCs, B.cells.Memory, B.cells.Naive, T.cells.CD4., T.cells.CD8., Monocyte, NKT.cells, NK.cells and Plasmablasts. We define the summation of all these cell fractions as immune cells fraction and use that to stratify LumA, ER+ and GP samples. In NAC cohort, we compared pCR rates and RoR scores and in MBRC data we compare their survival probabilities. As Fig. 5 shows, LumA and ER+ samples dont show significant difference. In GP, difference shows up in RoR score.

In IMC data, we have Macrophages, B cells, $CD38^+$ lymphocytes, Macrophages & granulocytes, T_{Reg}&T_{Ex}, $CD4^+$ T cells & APCs, $CD38^+$ lymphocytes, $CD4^+$, T cells, $CD8^+$ T cells and Granulocytes. Similarly, we put the summation of these cell fractions as Immune Cells Fraction and use it to stratify the samples in LumA, ER+ and GP. As the Fig. 5 shows, immune cells fraction does not stratify the patients in LumA and ER+ samples.



(a) LumA NAC (b) LumA NAC (c) LumA MBRC (d) LumA IMC MBRC

(e) ER+ NAC (f) ER+ NAC (g) ER+ MBRC (h) ER+ IMC MBRC

(i) GP NAC (j) GP NAC (k) GP MBRC (l) GP IMC MBRC

Figure 5: Immune cells fraction used to stratify LumA, ER+ and GP samples in NAC, MBRC and IMC MBRC data. In both LumA and ER+, immune cell fraction levels don't lead to different clinical outcomes.

[1] H. R. Ali, L. Chlon, P. D. Pharoah, F. Markowetz, and C. Caldas, Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study, PLoS medicine **13**, e1002194 (2016).

[2] M. Stavrou and A. Constantinidou, Tumor associated macrophages in breast cancer progression: implications and clinical relevance, Frontiers in Immunology **15**, 1441820 (2024).

[3] J. Zhao, W. Yu, M. Zhou, *et al.*, Immune microenvironment and intrinsic subtyping in hormone receptor-positive/her2-negative breast cancer, Breast Cancer Research **23**, 10.1186/s13058-021-01438-0 (2021).

[4] L. Buisseret, C. Desmedt, S. Garaud, *et al.*, The composition of t cell infiltrates varies in primary invasive breast cancer of different molecular subtypes as well as according to tumor size and nodal status, Breast Cancer Research **21**, 10.1186/s13058-018-1076-6 (2019).

[5] W. Chen, J. Jiang, Y. Xu, *et al.*, Research progress on the role of regulatory t cell in tumor microenvironment in the treatment of breast cancer, Frontiers in Oncology **11**, 749120 (2021).

[6] L. Silwal-Pandit, S. Nord, H. von der Lippe Gythfeldt, E. K. Møller, T. Fleischer, E. Rødland, M. Krohn, E. Borgen, Ø. Garred, T. Olsen, *et al.*, The longitudinal transcriptional response to neoadjuvant chemotherapy with and without bevacizumab in breast cancermolecular response in bevacizumab-treated breast tumors, Clinical Cancer Research **23**, 4662 (2017).

[7] L. A. Korde, L. Lusa, L. McShane, P. F. Lebowitz, L. Lukes, K. Camphausen, J. S. Parker, S. M. Swain, K. Hunter, and J. A. Zujewski, Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer, Breast Cancer Research and Treatment **119**, 685 (2010).

[8] P. F. Lebowitz, J. Eng-Wong, S. M. Swain, A. Berman, M. J. Merino, C. K. Chow, D. Venzon, F. Zia, D. Danforth, E. Liu, *et al.*, A phase ii trial of neoadjuvant docetaxel and capecitabine for locally advanced breast cancer, Clinical cancer research **10**, 6764 (2004).

[9] Y. Lin, S. Lin, M. Watson, K. M. Trinkaus, S. Kuo, M. J. Naughton, K. Weilbaecher, T. P. Fleming, and R. L. Aft, A gene expression signature that predicts the therapeutic response of the basal-like breast cancer to neoadjuvant chemotherapy, Breast Cancer Research and Treatment **123**, 691 (2010).

[10] V. Popovici, W. Chen, B. D. Gallas, C. Hatzis, W. Shi, F. W. Samuelson, Y. Nikolsky, M. Tsyganova, A. Ishkin, T. Nikolskaya, *et al.*, Effect of training-sample size and classification difficulty on the accuracy of genomic predictors, Breast Cancer Research **12**, 1 (2010).

[11] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T.-M. Chu, F. M. Goodsaid, L. Pusztai, *et al.*, The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models, Nature biotechnology **28**, 827 (2010).

[12] A. Tabchy, V. Valero, T. Vidaurre, A. Lluch, H. Gomez, M. Martin, Y. Qi, L. J. Barajas-Figueroa, E. Souchon, C. Coutant, *et al.*, Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer30-gene t/fac response predictor in breast cancer, Clinical Cancer Research **16**, 5351 (2010).

[13] T. Iwamoto, G. Bianchini, D. Booser, Y. Qi, C. Coutant, C. Ya-Hui Shiang, L. Santarpia, J. Matsuoka, G. N. Hortobagyi, W. F. Symmans, *et al.*, Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer, Journal of the National Cancer Institute **103**, 264 (2011).

[14] S. Glück, J. S. Ross, M. Royce, E. F. McKenna, C. M. Perou, E. Avisar, and L. Wu, Tp53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine±trastuzumab, Breast Cancer Research and Treatment **132**, 781 (2012).

[15] K. Shen, Y. Qi, N. Song, C. Tian, S. D. Rice, M. J. Gabrin, S. L. Brower, W. F. Symmans, J. A. O'Shaughnessy, and F. A. Holmes, Cell line derived multi-gene predictor of pathologic response to neoadjuvant chemotherapy in breast cancer: a validation study on us oncology 02-103 clinical trial, BMC Medical Genomics **5**, 1 (2012).

[16] J. A. Bauer, A. B. Chakravarthy, J. M. Rosenbluth, D. Mi, E. H. Seeley, N. De Matos Granja-Ingram, M. G. Olivares, M. C. Kelley, I. A. Mayer, I. M. Meszoely, *et al.*, Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiationmarkers of taxane sensitivity in breast cancer, Clinical Cancer Research **16**, 681 (2010).

[17] C. Hatzis, L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, *et al.*, A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer, JAMA **305**, 1873 (2011).

[18] M. J. M. Magbanua, D. M. Wolf, C. Yau, S. E. Davis, J. Crothers, A. Au, C. M. Haqq, C. Livasy, H. S. Rugo, L. Esserman, *et al.*, Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response, Breast Cancer Research **17**, 1 (2015).

[19] T. Miyake, T. Nakayama, Y. Naoi, N. Yamamoto, Y. Otani, S. J. Kim, K. Shimazu, A. Shimomura, N. Maruyama, Y. Tamaki, *et al.*, Gstp 1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in er-negative breast cancer, Cancer Science **103**, 913 (2012).

[20] J. C. Liu, V. Voisin, G. D. Bader, T. Deng, L. Pusztai, W. F. Symmans, F. J. Esteva, S. E. Egan, and E. Zacksenhaus, Seventeen-gene signature from enriched her2/neu mammary tumor-initiating cells predicts clinical outcome for human her2+: Erα- breast cancer, Proceedings of the National Academy of Sciences **109**, 5832 (2012).

[21] A. Prat, G. Bianchini, M. Thomas, A. Belousov, M. C. Cheang, A. Koehler, P. Gómez, V. Semiglazov, W. Eiermann, S. Tjulandin, *et al.*, based pam50 subtype predictor identifies higher responses and improved survival outcomes in her2-

positive breast cancer in the noah study, Clinical Cancer Research **20**, 511 (2014).

[22] E. Danenberg, H. Bardwell, V. R. Zanotelli, E. Provenzano, S.-F. Chin, O. M. Rueda, A. Green, E. Rakha, S. Aparicio, I. O. Ellis, *et al.*, Breast tumor microenvironment structures are associated with genomic features and clinical outcome, Nature Genetics **54**, 660 (2022).

Response to Reviewers

We thank the reviewer for their feedback. Below, we address each comment. Below, we address their comments (changes made to the main text/SI are in blue for clarity).

COMMENTS FROM REVIEWER #5

I have substantial concerns about this manuscript. Its overall scope is limited, the conceptual advances feel incremental, and the rationale for several analyses is not sufficiently robust. The idea that MHC+ macrophages drive T-cell exhaustion (T_{ex}) has already been mechanistically explored in at least two solid immunology studies (PMID: 39724910, Immunity, 2025; PMID: 35623342, Cancer Cell, 2022), which significantly reduces the novelty of the main analytical conclusions here. Overall, compared with other works of a similar type currently published in Nature Communications, the competitiveness and reader appeal of this study are inadequate.

Reply: We thank the reviewer for their evaluation of our revised manuscript and the opportunity to clarify the conceptual and methodological advances of our study:

In this paper we present an explainable machine learning (XML)–guided, integrative multi-omics framework to breast cancer, unifying bulk, spatial, and single-cell analyses from more than 5,000 human samples. This framework is not only interpretable and mechanistically grounded, but also robustly validated across modalities, providing a level of internal consistency rarely achieved in studies of this scale. Cross-validation of findings between deconvolution and IMC-based estimates, as well as between IMC and single-cell analyses, ensures that the observed associations are biologically reproducible rather than dataset-specific. This systematic integration distinguishes our study from previous reports that rely on single modalities or model systems. Indeed, we acknowledge the importance of the cited works (one of which was published during our revision process), which bring evidence from mouse melanoma and human glioblastoma models supporting the findings in our XML models, despite the substantial difference from human breast cancer in terms of tissue architecture, immune composition and clinical context. It is well recognized that the tumor microenvironment is highly context-dependent, and findings from other cancer types cannot be directly extrapolated to breast tumors [1, 2]. Our work specifically addresses this critical gap by capturing all signal from real world biological specimens from patients.

In brief, the important novel aspects of our work, which have been acknowledged by previous reviewers during earlier rounds of evaluation, include:

- (i) Discovery of a striking dichotomy in ER-positive/Luminal breast cancers, where macrophages correlate positively with pCR, but negatively with relapse free survival.
- (ii) HLA-ABC^{hi} macrophages associated with T_{Reg} and T_{Ex} in breast tumors.
- (iii) Quantitative insight into how spatial colocaliation contributes to immunosupressive niches shaping (adverse) outcomes.

Together, these findings establish a cross-modally validated and explainable link between macrophage phenotypes and clinical outcomes in breast cancer, representing a conceptual and methodological advance that goes well beyond prior observations.

Q1. *In Fig. S9, the prognostic association of macrophages is significant in only one cohort, and is even reversed in TCGA. This suggests the phenomenon may not be broadly robust or generalizable. ...*

Reply: Thank you for raising these important points regarding the robustness of our findings. We agree that cohort-specific variability exist given differences in patient demographics, treatment regimens and data modalities. However, we maintain that the negative prognostic association of macrophages is broadly robust and consistent across multiple, distinct analytical approaches and datasets. For clarity, we reiterate our results for consistence association of macrophage fractions and clinical outcomes:

- Survival Scores: the results displayed in Figs 2 and 3 clearly show that macrophages are consistently associated with lower RFS probability in both TCGA and MBRC across LumA, LumB, ER+ and GP.

- Single Variant analysis (survival contour plots): These results consistently show that macrophages are associated with lower RFS probability in IMC data of MBRC, deconvoluted data of MBRC and TCGA across LumA, LumB, ER+ and GP.

- Kaplan–Meier survival plots: In these plots, samples are divided into two groups and then their survival probabilities compared: we observed this stratification lead to significant clinical outcomes in IMC data, deconvolution data of MBRC but not TCGA. We re-iterate that the lack of such observation does not undermine the role of macrophages in the survival probability of the TCGA dataset, as we consistently recovered negative Survival Scores in LumA, LumB, ER+, and GP in this dataset.

These results confirm that the observed associations are not dependent on a single analytical method. The consistent direction and reproducibility of the macrophage–survival relationship across these independent layers further confirms the robustness of our findings and the reliability of the XML-based approach in capturing clinically relevant patterns.

To clarify this consistency, and to further illustrate the robustness of the observed associations, we have now included additional contour plots (see new panels in Fig. S9 and Fig. S10). Here are all results that show that macrophages always are associated with lower survival probability and can stratify patients to distinct groups:

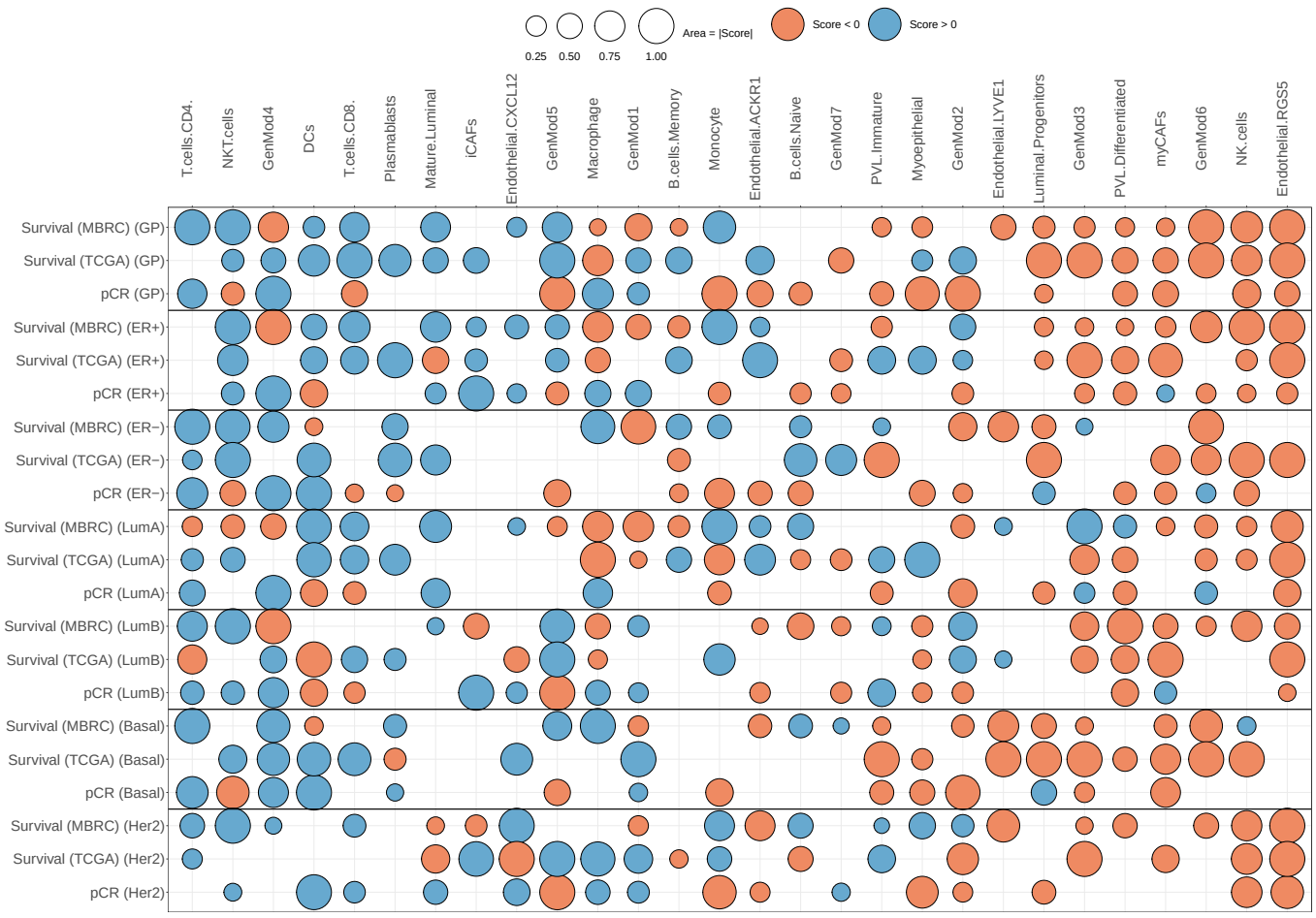
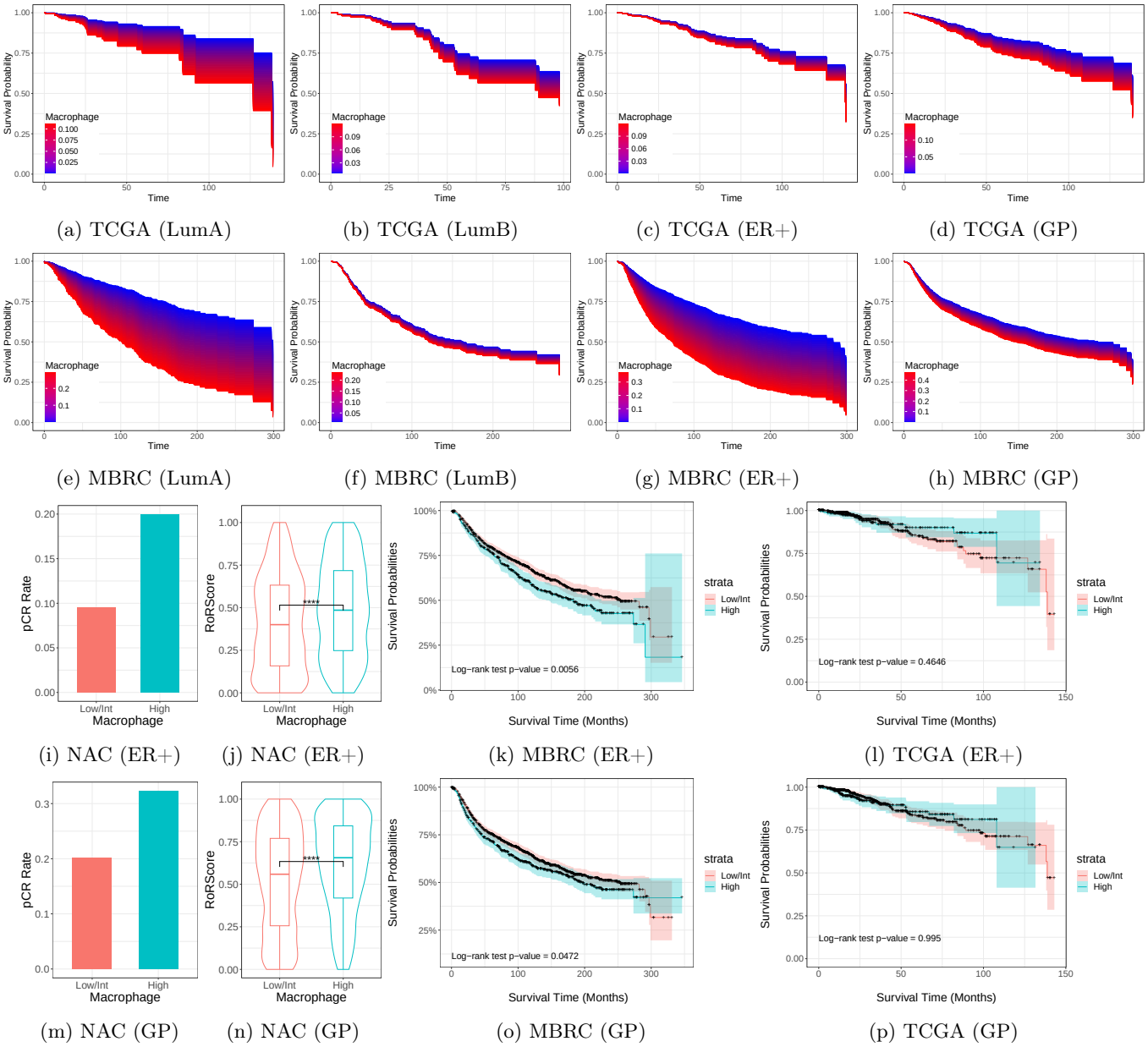


Figure 1: Survival Scores for subtypes in MBRC and TCGA datasets, alongside pCR scores.



(a) TCGA (LumA) (b) TCGA (LumB) (c) TCGA (ER+) (d) TCGA (GP)

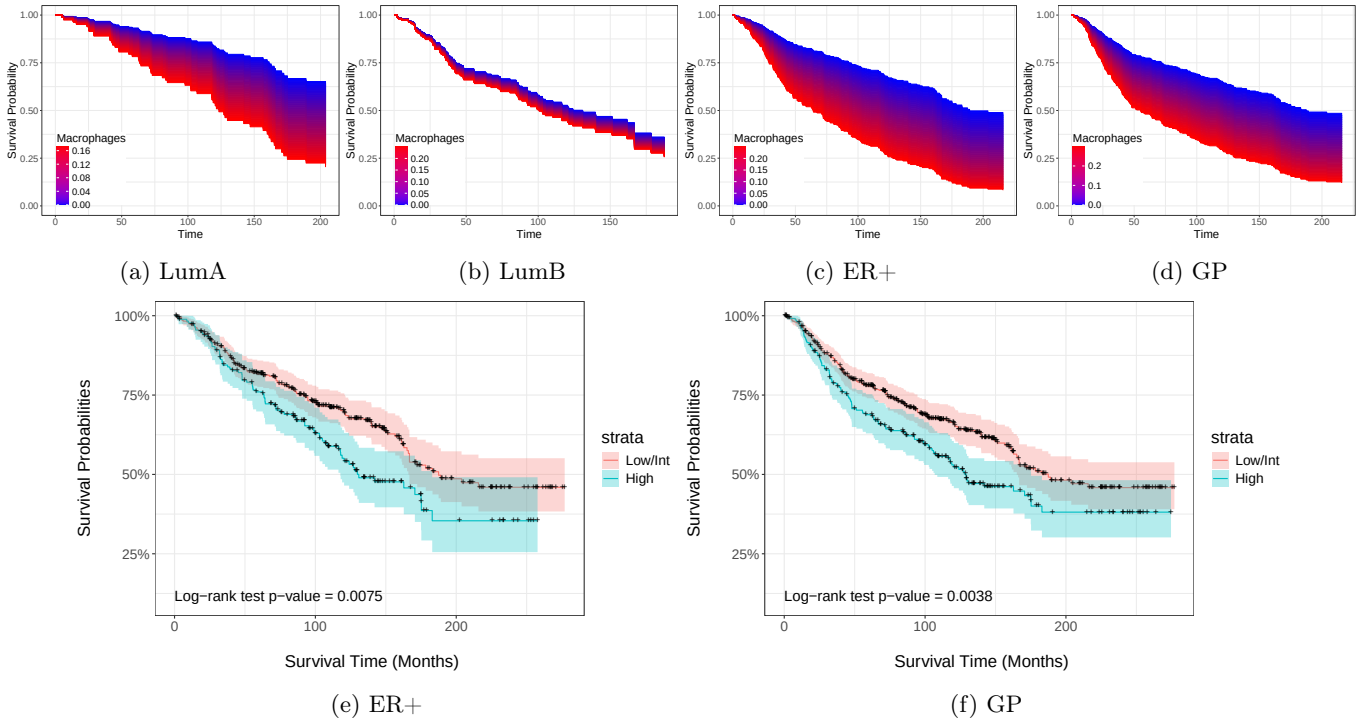
(e) MBRC (LumA) (f) MBRC (LumB) (g) MBRC (ER+) (h) MBRC (GP)

(i) NAC (ER+) (j) NAC (ER+) (k) MBRC (ER+) (l) TCGA (ER+)

(m) NAC (GP) (n) NAC (GP) (o) MBRC (GP) (p) TCGA (GP)

Figure 2: **Macrophage fraction inferred from bulk deconvolution stratifies breast cancer samples into groups with distinct clinical outcomes.** (a–h) Association between macrophage fraction and survival probability in TCGA and MBRC datasets across LumA, LumB, ER+, and GP subtypes. (i–l) Comparison of pCR rates and RoR scores and survival probabilities between macrophage-high and macrophage-low groups in NAC, MBRC, and TCGA cohorts. Higher macrophage fractions correspond to worse overall survival and, in the NAC cohort, higher pCR rates and RoR scores.

IMC Data:



(a) LumA (b) LumB (c) ER+ (d) GP

(e) ER+ (f) GP

Figure 3: **Macrophage fraction estimated from IMC data stratifies breast cancer samples into groups with distinct clinical outcomes.** (a–d) Association between IMC-derived macrophage fraction and survival probability across LumA, LumB, ER+, and GP subtypes. (e–f) Kaplan–Meier survival curves for ER+ and GP tumors, showing that macrophage-high samples exhibit significantly lower survival probabilities.

We have also revised the title of our paper to Explainable machine learning–guided integrated multi-omics analysis reveals macrophage-driven immune suppression in breast cancer to more clearly reflect the central message of the study.

Q1. *.... Moreover, macrophages are highly heterogeneous, so their functions should not be analyzed as a single lumped population. In Fig. 4, however, the authors stratify by overall macrophage abundance rather than by subset abundance, calling into question the true clinical and biological meaning of the analyses presented in Fig. 4.*

Reply: We fully agree that macrophages are functionally and transcriptionally heterogeneous, and we explicitly addressed this in our analyses. As shown in Fig. 4b, we first included macrophage subsets (CXCL10, EGR1, LAM1-FABP5, and LAM2-APOE) in the deconvolution to test whether the observed clinical associations could be explained by known transcriptional subtypes. However, this analysis did not account for the divergent behavior observed at the bulk level, suggesting that other biological mechanisms—beyond classical subset definitions underlie the variation in clinical outcome.

To investigate this, we next hypothesized that the level of macrophage infiltration itself could define distinct tumor microenvironments (TMEs). Indeed, we found that LumA tumors with high macrophage fractions showed both higher pCR and higher recurrence risk (Fig. 4c–e), demonstrating functional differences between macrophage-rich and macrophage-poor TMEs. We then asked whether such differences could be explained by macrophage composition. To address this, we stratified macrophages not by pre-defined molecular markers but by their spatial and clinical associations. This revealed that macrophages could be further divided by distance-based criteria (Figs. 4g–j), and that tumors with higher overall macrophage fractions also contained a greater proportion of macrophages associated with poor survival (Fig. 4k).

Thus, rather than treating macrophages as a single "lumped" population, Fig. 4 presents a stepwise exploration from bulk abundance to molecular and spatial stratification, aiming to identify macrophage subsets that are clinically and biologically relevant. This framework directly motivates the analyses in Fig. 5, where we define macrophage subtypes based on their functional and transcriptional states.

Q2. *In Fig. 3, why are CD8 T cells associated with non-pCR? As the canonical cytotoxic effector cells in anti-tumor*

immunity, how is this result explained? Does this finding call into question the validity of the pCR score?

Reply: We thank the reviewer for this important point. We agree that, mechanistically, CD8 T cells are key cytotoxic effectors, and one might expect their abundance to correlate positively with pCR. However, in breast cancer this association has been shown to be context-dependent and not uniformly positive across all subtypes. A previous meta-analysis reported that CD8 T-cell infiltration does not consistently predict improved response to neoadjuvant chemotherapy [3]. Similarly, [4] showed that while CD8 T-cell infiltration correlates with higher overall survival in colorectal cancer, it is associated with lower survival in renal cancer. Other studies have also observed increased CD8 T-cell presence in tumors exhibiting immune exhaustion or therapy resistance [5, 6].

Our findings are consistent with these observations. In our results, CD8 T cells are positively associated with survival (Fig. 3), consistent with their known role in long-term tumor control [7]. However, their inverse association with pCR in the overall population likely reflects the presence of exhausted or dysfunctional CD8 T cells that are unable to mount effective cytotoxic responses at the time of treatment. Notably, this trend is subtype-dependent: in the Her2-enriched subset, CD8 T cells show a positive association with pCR, consistent with higher immune activation and responsiveness in this context.

These results indicate that the apparent non-pCR association of CD8 T cells reflects functional differences in CD8 T-cell states within the tumor microenvironment—such as exhaustion or immunosuppression—rather than any limitation of the pCR score itself.

Q3. *The authors still do not clearly show which single-cell–defined macrophage subset corresponds to the MHC+ macrophages observed in the IMC data, nor how those subsets relate to the groups in Fig. 4. At present, the analyses of IMC, single-cell, and bulk RNA-seq data remain poorly integrated.*

Reply: We agree that the link between IMC and scRNA-seq must be explicit. In our study, we deliberately stratified scRNA-seq macrophages by HLA-ABC levels (HLA-ABC^{hi} vs. HLA-ABC^{lo}) to mirror the IMC marker (HLA-ABC used to define macrophages, thereby enabling a direct, cross-modal comparison. As shown in Fig. 5f, previously reported scRNA-seq subsets (CXCL10, EGR1, LAM1-FABP5, LAM2-APOE) are evenly distributed across HLA-ABC strata, indicating that HLA-ABC^{hi} corresponds to a functional state not captured by any single transcriptomic subset. This HLA-ABC axis unifies modalities: in IMC, HLA-ABC^{hi} macrophages colocalize with and are correlated with T_{Reg}/T_{Ex} (Fig. 5b–e); in scRNA-seq, they show interferon/immunoregulatory programs and correlate with T_{Reg}/T_{Ex} abundance (Fig. 5g–j); and in bulk cohorts, higher macrophage content aligns with poorer survival (Fig. 4c–e), consistent IMC data (Fig. 4f and Fig. S10).

Main text was changed as follows to reflect this logic:

... In scRNA-seq data, we have three genes for HLA class I: HLA-A, HLA-B, and HLA-C. We define a single value representing HLA-ABC as follows: HLA-ABC = log2(HLA-A + 1) + log2(HLA-B + 1) + log2(HLA-C + 1). To enable a direct comparison with IMC, and validate our findings, we *a priori* stratified scRNA-seq macrophages by HLA-ABC (HLA-ABC^{hi} vs. HLA-ABC^{lo}), the transcriptomic analogue of the IMC-defined phenotypes. Thus, scRNA-seq sample macrophages (10,700 cells across 57 samples) were divided into two groups using this single value (HLA-ABC^{lo} and HLA-ABC^{hi}). ...

Q4. *Please clarify the precise definition of Tex used here. Is it closer to Tex_prog, Tex_int, or Tex_term?*

We borrow annotation from Wu et al [8], where, exhausted T cells (Tex) correspond to a PD-1^{+}LAG3^{+}TIGIT^{+} CD8^{+} subset with preserved CD27/CD70-mediated cytotoxic capacity—functionally that can aligne with an intermediate or progenitor exhausted phenotype (Tex_prog/int), not terminal exhaustion (Tex_term).

Q5. *Please verify whether the group labels in Fig. 4g are incorrect.*
They are correct.
Q6. *In Fig. S5, the point colors and line colors should not be the same; as currently drawn they are hard to distinguish.*
Thank you, figure was updated for better visuals:

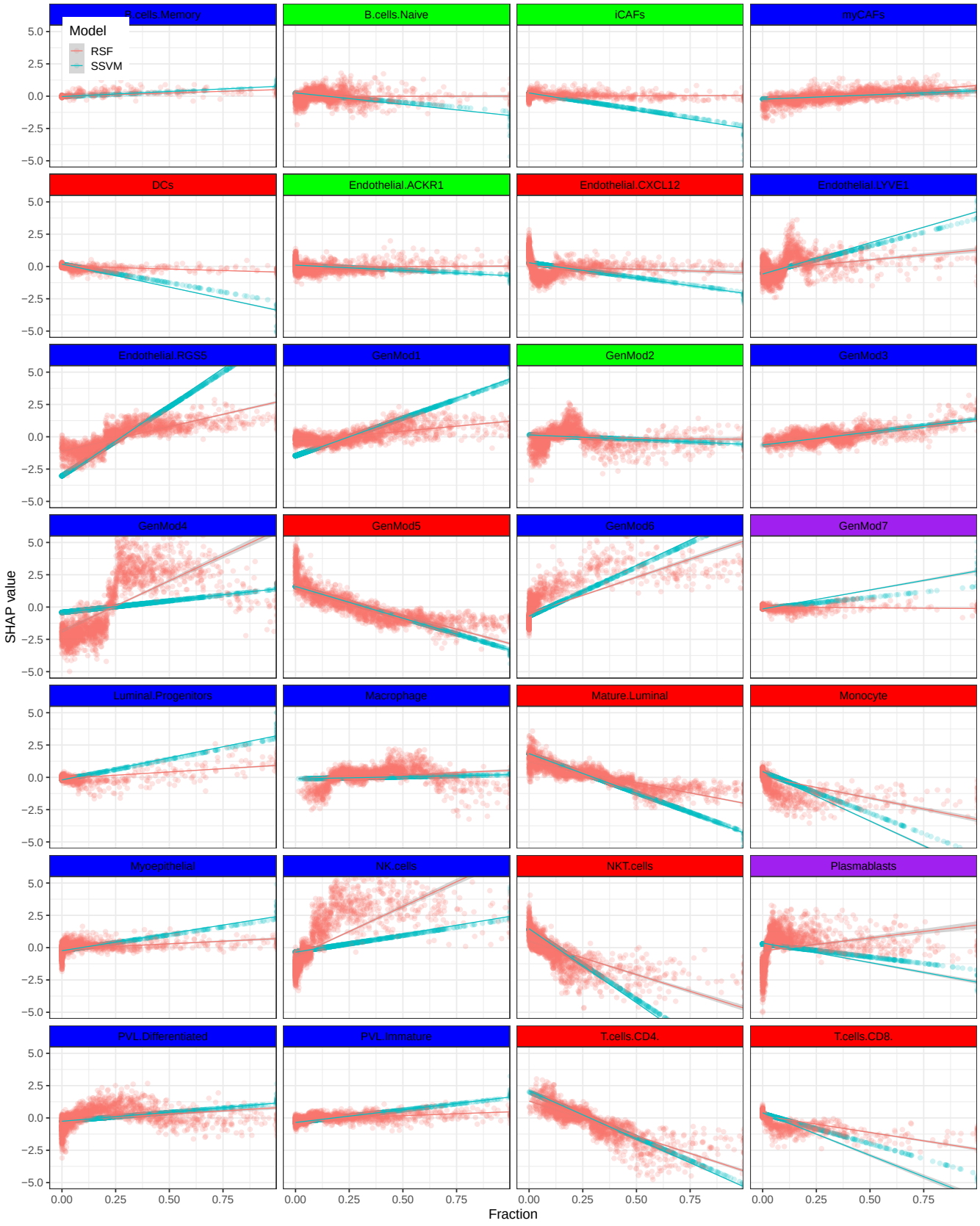


Figure 4: SHAP values for SSVM and RSF vs normalized cell type fractions, following the same color scheme introduced in Fig. 1. Some cell types are filtered out (green and purple) and for the remaining cells we calculate Survival scores.

[1] K. Aliazis, A. Christofides, R. Shah, Y. Y. Yeo, S. Jiang, A. Charest, and V. A. Boussiotis, The tumor microenvironment's role in the response to immune checkpoint blockade, Nature Cancer , 1 (2025).

[2] G. F. Beeghly, A. A. Shimpi, R. N. Riter, and C. Fischbach, Measuring and modelling tumour heterogeneity across scales, Nature Reviews Bioengineering **1**, 712 (2023).

[3] R. D. Bense, C. Sotiriou, M. J. Piccart-Gebhart, J. B. Haanen, M. A. van Vugt, E. G. de Vries, C. P. Schröder, and R. S. Fehrmann, Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer, JNCI: Journal of the National Cancer Institute **109** (2017).

[4] N. C. Blessin, W. Li, T. Mandelkow, H. L. Jansen, C. Yang, J. B. Raedler, R. Simon, F. Büscheck, D. Dum, A. M. Luebke, *et al.*, Prognostic role of proliferating cd8+ cytotoxic tcells in human cancers, Cellular Oncology **44**, 793 (2021).

[5] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, *et al.*, Supervised risk predictor of breast cancer based on intrinsic subtypes, Journal of Clinical Oncology **27**, 1160 (2009).

[6] J. Wagner, M. A. Rapsomaniki, S. Chevrier, T. Anzeneder, C. Langwieder, A. Dykgers, M. Rees, A. Ramaswamy, S. Muenst, S. D. Soysal, *et al.*, A single-cell atlas of the tumor and immune ecosystem of human breast cancer, Cell **177**, 1330 (2019).

[7] H. Ali, E. Provenzano, S.-J. Dawson, F. Blows, B. Liu, M. Shah, H. Earl, C. Poole, L. Hiller, J. Dunn, *et al.*, Association between cd8+ t-cell infiltration and breast cancer survival in 12 439 patients, Annals of Oncology **25**, 1536 (2014).

[8] S. Z. Wu, G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy, N. Bartonicek, *et al.*, A single-cell and spatially resolved atlas of human breast cancers, Nature Genetics **53**, 1334 (2021).

Response to Reviewers

We thank the reviewer for their feedback. Below, we address each comment. Below, we address their comments (changes made to the main text/SI are in blue for clarity).

COMMENTS FROM REVIEWER #5 (REMARKS TO THE AUTHOR)

The authors have addressed most of the concerns, but several issues remain to be discussed:

1. In Fig. 5g, the MHC-lo group shows upregulation of MHC genes. Please verify whether the group labels may have been incorrectly assigned. In addition, if the labels are indeed incorrect, the MHC-hi group shows upregulation of CXCL9. In PMID: 37535729 (2023, Science), a high CXCL9/SPP1 ratio was associated with better prognosis and superior immunotherapy pCR across multiple cancer types. Please present the CXCL9/SPP1 ratio and their individual expression levels in the MHC-lo/high groups, as well as the correlations between MHC gene expression and CXCL9, SPP1 individual expression, and the CXCL9/SPP1 ratio in macrophages. Please discuss how the above analytical results may align with or contradict the conclusions of the article PMID: 37535729 and the conclusions of the present study.

Reply: We thank the reviewer for pointing out this. As the reviewer suggested, the labels in the plot are reversed. They are now corrected. Regarding the additional analysis, analysis was done (please see SI fig S25) and a new paragraph was added to the discussion (in blue).

DISCUSSION

...

Another marker in macrophages is the CXCL9/SPP1 ratio, which has been linked to improved overall survival and higher pCR rates in multiple cancers [1]. While HLA-ABC levels display a positive correlation with CXCL9 expression and the CXCL9/SPP1 ratio (SI Fig. S25), our results show that HLA-ABC^{hi} macrophages preferentially colocalize with TReg and TEx populations and mark an immunosuppressive rather than an immunostimulatory niche. This indicates that the same molecular program—CXCL9-high antigen-presenting macrophages—can play fundamentally different roles depending on tumor context. In breast cancer, chronic antigen presentation coupled with spatially organized T-cell dysfunction appears to dominate, contrasting with the effector-recruiting role of CXCL9^{+} macrophages described in [1]. However, a more detailed analysis is required to explore this relationship more fully.

SI

HLA-ABC and Other Genes in Macrophages

...

To further contextualize the HLA-ABC–associated macrophage state, we next examined genes linked to macrophage functional polarization and immune modulation. Recent work has highlighted the balance between CXCL9 and SPP1 expression as a key axis distinguishing immunologically active versus tissue-remodeling macrophage programs [1]. We assessed the relationship between HLA-ABC expression and CXCL9, SPP1, and their ratio as shown in Fig. 1.

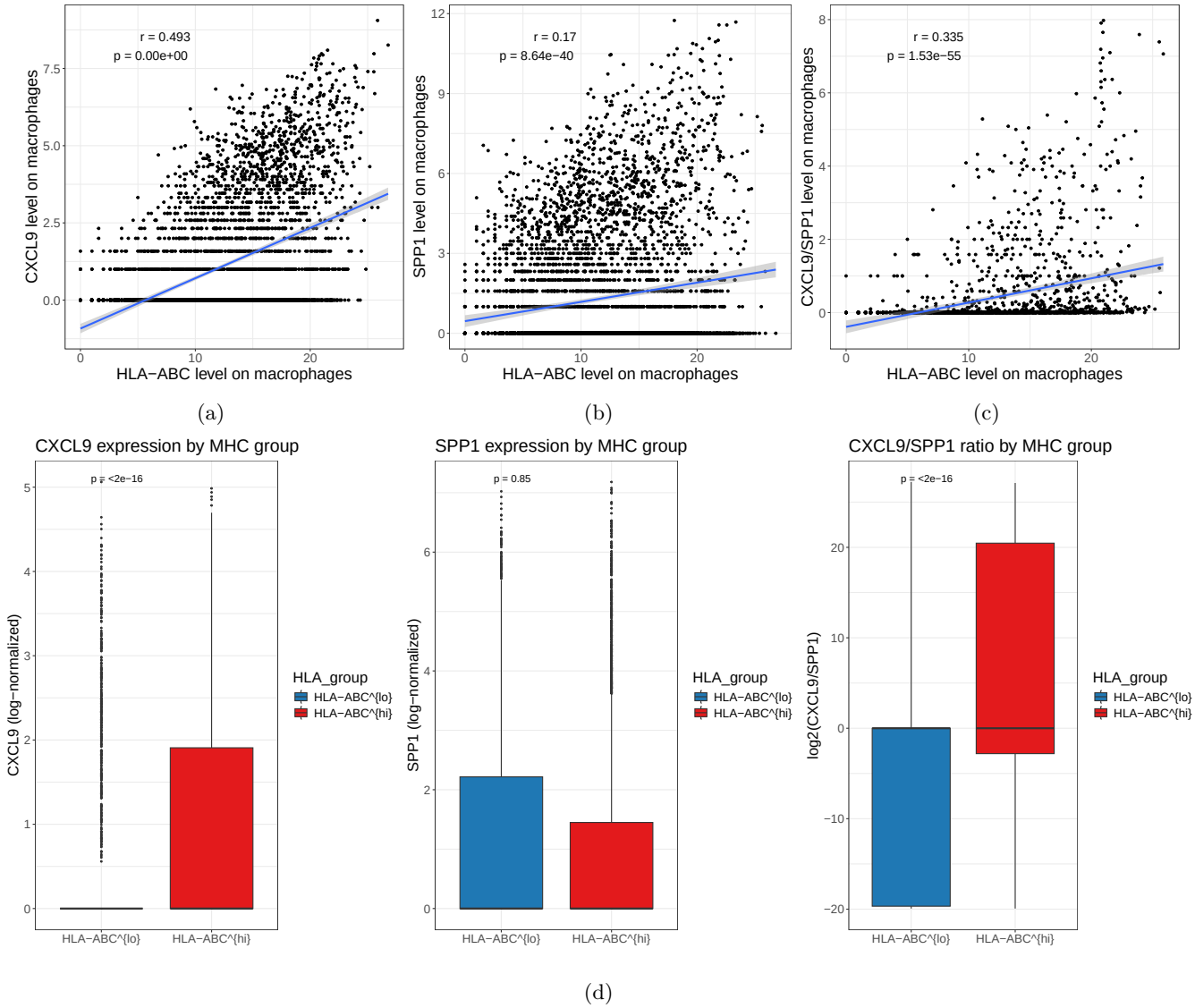


Figure 1: Correlation of HLA-ABC vs CXCL9 (a), SPP1 (b) and CXCL9/SPP1 (c) on macrophages in scRNA-seq. c) CXCL9, SPP1 and CXCL9/SPP1 values across macrophage subsets. Box plots show the median and 25th–75th percentiles in d. Source data are provided as a Source Data file.

2. Please replace Fig. 5b and Fig. 5c with versions in which Treg and Tex are labeled separately.

Reply:

We provide the corresponding updated plots in Fig. S18(a–c) (see below), which mirror Fig. 5(b,c). The results remain consistent with our original conclusions and further support the generalizability of our findings. Because this re-annotation is intended as an additional robustness analysis rather than part of the primary analysis workflow, we include it in the SI.

SI

Further stratification of T_{Reg}&T_{Ex}

In Danenberg *et al.* [2], T_{Reg} and T_{Ex} were not analyzed separately. Here, we have now re-examined the data by explicitly separating the two populations. We stratified the combined T-cell compartment by FOXP3 and PD-1 expressions: among T_{Reg}&T_{Ex} cells, top 33% with FOXP3 labeled as T_{Reg}. Among the remaining 67%, cells with

PD-1 > median were treated as T_{Ex}. This annotation led to FOXP3^{hi} as T_{Reg}, and FOXP3^{lo} PD-1^{hi} as T_{Ex} (FOXP3^{lo} PD-1^{lo} remain unlabeled). This is an effort to annotate cells with high confidence without complicating the analysis and should be seen as support for our main results. An example of such re-annotation is shown in 2(a). Similar to Fig. 5(b), both T_{Ex} and T_{Reg} are located closer to macrophages near MHC I & II^{hi} epithelial cells, consistent with Fig. 5(c). We then re-computed, the correlations between the frequencies of HLA-ABC^{hi}/HLA-ABC^{lo} macrophages and each T-cell subset separately in all subsets of tumors. As Fig. 2 (d) (and table S1) shows, HLA-ABC^{hi} macrophages are exclusively correlated with both T_{Ex} and T_{Reg}.

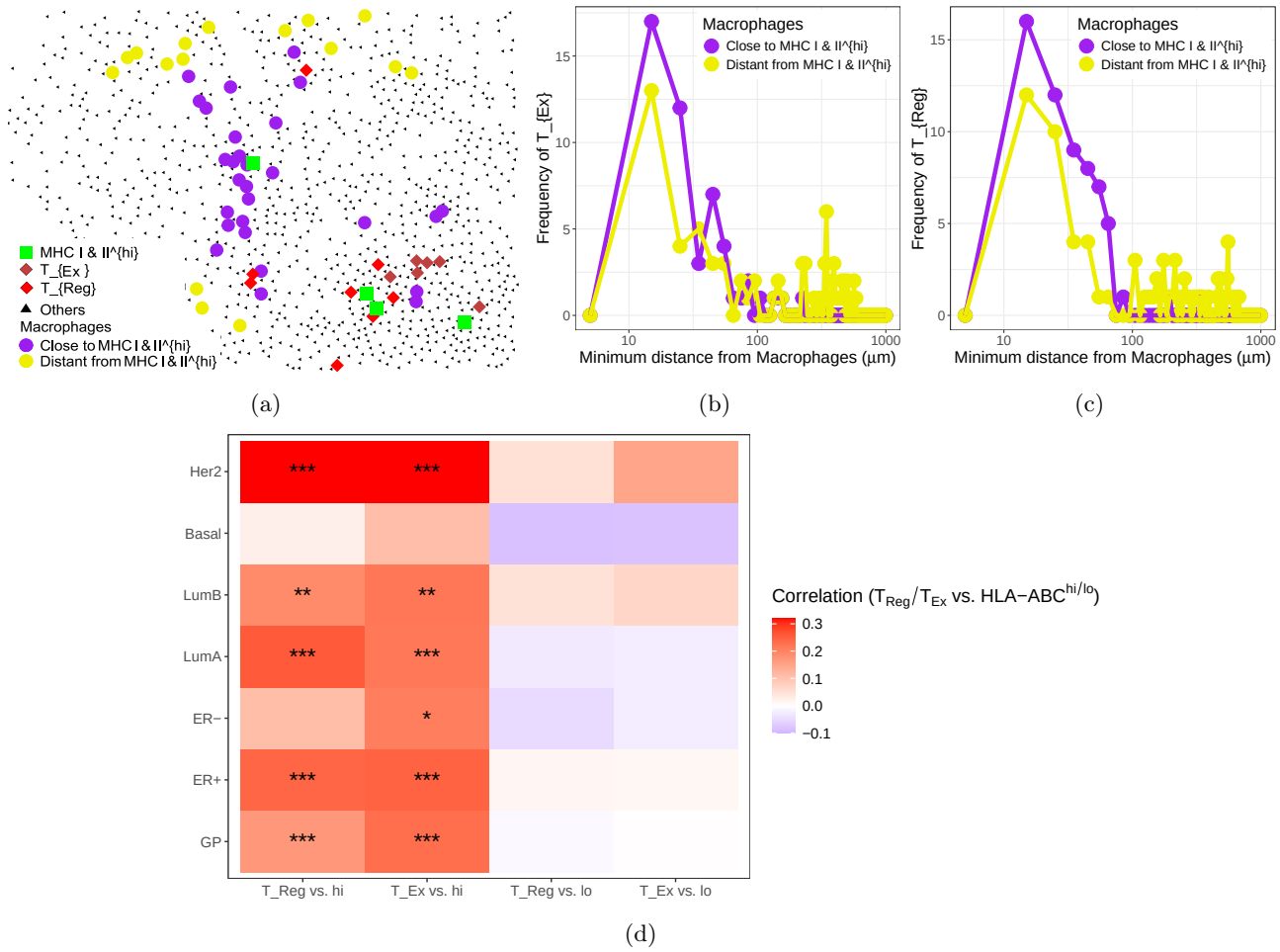


Figure 2: Spatial distribution of T_{Ex} and T_{Reg} cells (a) and their ferquencies in neighborhoods of two groups of macrophages (b-c). d) Correlations between the frequencies of T_{Ex} or T_{Reg} and HLA-ABC$^{hi/lo}$ macrophages across GP and tumor subtypes. See Table S1 for p-values calculated using two-sided Pearson correlation tests. Source data are provided as a Source Data file.

It should be noted that because cell-state annotation is inherently complex and context-dependent—and FOXP3/PD-1 can be transiently expressed—this FOXP3/PD-1-based separation is operational rather than definitive. We applied it to a limited set of cells to obtain an initial view of how these T-cell subsets associate with macrophage phenotypes.

Table S1: Significant correlations between HLA.ABC expression and T-cell populations across conditions.

Condition	r	p	n	Pair	stars	r_{lab}
GP	0.1695	8.67×10^{-6}	681	T_Reg vs. hi	***	0.17
GP	0.2303	1.20×10^{-9}	681	T_Ex vs. hi	***	0.23
ER+	0.2420	1.57×10^{-8}	532	T_Reg vs. hi	***	0.24
ER+	0.2456	9.44×10^{-9}	532	T_Ex vs. hi	***	0.25
ER-	0.2083	0.0108	149	T_Ex vs. hi	*	0.21
LumA	0.2571	8.90×10^{-5}	227	T_Reg vs. hi	***	0.26
LumA	0.2177	9.62×10^{-4}	227	T_Ex vs. hi	***	0.22
LumB	0.1894	0.0090	189	T_Reg vs. hi	**	0.19
LumB	0.2187	0.0025	189	T_Ex vs. hi	**	0.22
Her2	0.7094	5.14×10^{-13}	77	T_Reg vs. hi	***	0.71
Her2	0.7964	4.79×10^{-18}	77	T_Ex vs. hi	***	0.80

COMMENTS FROM REVIEWER #6 (REMARKS TO THE AUTHOR): ARBITRATING REVIEWER; EXPERT IN TUMOUR IMMUNOLOGY, MACROPHAGES, AND CANCER GENOMICS

The authors have addressed the comment of Reviewer 5 in an appropriate way, and have provided justified interpretation of their results. Figures 5 was corrected, as requested. The explanation for the link between the CXCL9/SPP1 ratio and HLA-ABC levels is fully satisfactory. The correlations can be cancer -specific , cohort-specific and model-specific. There is current gap of knowledge in our understanding what is the driving mechanism in macrophage-mediated immunosuppression, and what is just a biomarker manifestation, however, not reflecting the key regulatory pro-or anti-tumoral functions. The data provided by the submission are valuable, even if the results are controversial to some data published by others. Evidence provided by various research groups analysing of various cancer patient cohorts is need to discriminate between "drivee or passanger", and to generate clinically useful knowledge. In this context, the results of the submission are relevant.

Reply:

We sincerely thank the reviewer for the thoughtful evaluation of our manuscript and for recognizing the value of our work. We greatly appreciate the acknowledgment that the revisions have adequately addressed the concerns raised by Reviewer 5, as well as the positive assessment of the corrected Figure 5 and the clarification regarding the CXCL9/SPP1–HLA-ABC relationship.

We fully agree with the reviewer's important point regarding the current gap in understanding whether macrophage-associated immunosuppressive states function as true mechanistic drivers of tumor progression or represent downstream biomarker manifestations of broader tumor-intrinsic or microenvironmental programs. We believe this distinction is central to advancing the field and to translating macrophage-related findings into clinically meaningful strategies. In response to this valuable perspective, we have expanded the Discussion by adding the following paragraph:

An important unresolved question in the field is whether macrophage states associated with immunosuppression represent active drivers of tumor progression or, alternatively, phenotypic readouts of upstream tumor-intrinsic or microenvironmental programs. Our study is not designed to establish causality, and the observed associations between HLA-ABC$^{\text{hi}}$ macrophages, exhausted/regulatory T cells, and adverse survival may reflect driver, passenger, or amplifier roles. Nevertheless, the reproducibility of these associations across independent cohorts, molecular modalities (bulk RNA-seq, IMC, and scRNA-seq), and spatial definitions argues that these macrophage states capture a biologically meaningful dimension of the tumor microenvironment rather than stochastic variation. Importantly, even if such macrophage phenotypes act primarily as biomarkers rather than primary regulators, their strong and consistent association with clinical outcome and immunosuppressive niches provides clinically actionable stratification and defines specific interaction axes (e.g., SIGLEC9, ALCAM–CD6, CSF1–CSF1R) that can be functionally tested in future perturbation studies.

[1] R. Bill, P. Wirapati, M. Messemaker, W. Roh, B. Zitti, F. Duval, M. Kiss, J. C. Park, T. M. Saal, J. Hoelzl, *et al.*, Cxcl9: Spp1 macrophage polarity identifies a network of cellular programs that control human cancers, Science **381**, 515 (2023).

[2] E. Danenberg, H. Bardwell, V. R. Zanotelli, E. Provenzano, S.-F. Chin, O. M. Rueda, A. Green, E. Rakha, S. Aparicio,

I. O. Ellis, *et al.*, Breast tumor microenvironment structures are associated with genomic features and clinical outcome, Nature Genetics **54**, 660 (2022).