nature portfolio

Peer Review File

Segmentation aware probabilistic phenotyping of single-cell spatial protein expression data

Corresponding Author: Dr Kieran Campbell

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)
This article introduces STARLING, a novel probabilistic clustering method designed to address segmentation errors in spatial protein expression data. The challenge of inaccurate cell annotation due to flawed cell segmentation is a significant technical hurdle in data analysis. The authors propose a probabilistic clustering model that explicitly accounts for segmentation errors, showcasing both innovation and practicality. The authors propose a probabilistic clustering model that explicitly accommodates segmentation errors, showcasing both innovation and practicality. The overarching concern addressed in the article is both necessary and well-reasoned. However, there are still some areas of concern:
Major concerns:
1. The evaluation of the approach is indirect and lacks ground truth. Including simulated data, such as mixed data generated from single-cell RNA-seq or single-cell proteomics, could enhance the accuracy assessment of identifying cell segmentation errors and cell phenotypes after applying the STARLING method.
2. A comprehensive review of similar approaches and benchmarking against those approaches is warranted.
3. It would be beneficial to provide a discussion and guidance on the selection of hyperparameters, such as λ in the loss function and the number of clusters.
4. The method can only identify cell overlaps involving two cells expressing different marker proteins, making it infeasible for cells of the same type. Similarly, STARLING does not address cell overlaps involving more than two cells.
5. The underlying principle of the approach suggests its potential application to other data types, such as spatial transcriptomes. Demonstrating this extension would further enhance the method's versatility and applicability.
Minor concerns:
1. In Figure 2A, although the co-expression conditions are explained in the methods, it would be beneficial to include direct explanations in the figure legend to improve readability. In addition, the colors are difficult to distinguish.
2. In Figure 3A, the differences in cell segmentation error probabilities appear to be minimal between different groups. Please provide information on the statistical test used and the corresponding p-values to support this observation.
3. In Figure S1, while it can be seen that there are more co-expression phenotypes for CD20 and CD3 in the third panel, the first two panels do not seem to show a significant number. How can it be proven that these are caused by cell segmentation errors? It would be helpful to use STARLING and present the results to assess if there is an improvement in the identification.
4. In Figure S4, the author mentioned that CD3 and CD20 exhibit lower expression levels in B and T cell clusters in the Baseline method heatmaps. However, this is not clearly visible in the figure. It would be informative to directly compare the expression differences of CD3 and CD20 specifically in the B and T cell clusters among different methods. Additionally, comparing the number of B and T cells identified in each method would be valuable.
5. There's a minor error in line 211 of the article. I think that the correct figure references should be "for both PhenoGraph (Fig. 4F) and STARLING (Fig. 4G)," not the original "for both PhenoGraph (Fig. 4G) and STARLING (Fig. 4H)."

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Main comments
Introduction
Define the scope more clearly
The authors start by introducing IMC as one multiplexed image technology (Line 32) but do not follow up on any of the other multiplexed image technologies. The rest of the introduction seems to indicate that this method is primarily developed for IMC as all the benchmarking data is generated using IMC. The authors should specify clearly if their method is specifically developed for Imaging Mass Cytometry (IMC) or if it can or could be applied to similar multiplex image technologies such as image-based Spatial Transcriptomics (ST).

Define the objective more clearly
The authors should specify more clearly in the introduction what the purpose and objectives of their method are: The authors state on Line 71, that STARLING is a novel clustering method. However, in the result section (lines 139 to 142) it is noted that STARLING is initialized with some clustering and then computes a de-noised clustering along the per-cell segmentation error probabilities (also shown in Figure 1). From reading the introduction it is unclear if their method presents a novel clustering algorithm or a post-processing algorithm for denoising clusters and identification of segmentation errors. The definition of the method and the fact it is initialized with a clustering strongly suggests the latter: a post-processing method for clustering algorithms. I recommend making the definition and purpose of STARLING more precise in the introduction.
Additionally, I suggest adding one sentence explaining the method just after introducing it on Line 73. Something like: “STARLING takes as input a clustering and the potentially erroneous cell segmentation masks and applies a probabilistic optimization algorithm to output a corrected clustering”. This will clarify the objective and help readers to understand the method and manuscript better.

Results
STARLING Input data - Lines 99-112
It is unclear why the high-dimensional image is required for STARLING other than for the generation of the clusters. From an algorithmic point of view, it would be more “clean” to take as input a pre-computed clustering and the imperfect cell segmentation masks. This goes in a similar direction to a previous comment, that the authors should specify more clearly what STARLING does: Is it a clustering algorithm or a post-processing method that is initialized with a clustering such as KMeans or PhenoGraph?
Decoupling STERLING from the initialized clustering algorithms would provide a more intuitive algorithmic framework and might allow the application of STERLING to other multiplex image technologies.
The authors should consider making this change or clearly state why this design decision was taken.

neighboring and non-neighboring cells
The plausibility score is evaluated by comparing cells with "no neighbors" to cells in densely packed regions. However, the definition of neighboring and non-neighboring cells is not clearly defined.
Authors should clarify their definition of neighboring cells: What is the quantification of two adjacent cells? Similarly, Figure 2B compares cells with no neighbors to cells with neighbors. This term might be misleading or confusing for readers. It would be more appropriate to compare cells in regions of low vs. high cell density.
Authors might consider a more appropriate definition of cells in densely packed tissue regions vs cells in low-density regions for example the mean/max/min distance to its neighboring cells.
In any case, the authors should specify the definition in the main manuscript and not in the Methods section to help readers understand the benchmark results better.

Plausibility score
The authors introduce the notion of “plausibility scores” to benchmark the differences in clusterings. Plausibility scores are calculated based on an a priori known set of proteins that either exhibit mutually exclusive expression or conditional co-expression. Plausibility is obtained by calculating the proportions of proteins that are co-expressed.
Although this score seems reasonable to estimate the segmentation results, it is unclear how the absolute scores can be interpreted. There is no range of “good and bad” scores defined. Figure 2C shows improvements in plausibility scores but it is not possible to estimate the significance of these improvements.
It would be more appropriate to use statistical tests to determine the plausibility of segmentation. For example: Calculate the p-value of two mutually exclusive proteins where the null hypothesis is that these are co-expressed. Such statistics would directly identify clusters that do not correspond to actual cell types and give a better understanding of the plausibility scores.

Discussion
Limitations and related work
Although this manuscript presents a novel approach to handling segmentation errors in single-cell analysis the overall issue is well-known in similar fields such as ST and has been addressed in other publications. Multiple probabilistic framework for

cell segmentation of ST data exists which take into account the gene/protein expressions in addition to nuclei and membrane stains. Such algorithms, such as [1, 2], tackle the same problem of segmentation errors from a different angle by incorporating additional information. It was previously shown that incorporating the gene/protein expression levels into cell segmentation can improve the segmentation result significantly by removing common segmentation errors such as those discussed in this manuscript. STARLING, on the other hand, simply identifies segmentation errors and improves the downstream clustering but does not solve the issue of segmentation errors.
Authors should address these approaches and briefly discuss differences.

Future work
Adapting STARLING for other multiplex imaging technologies such as image-based ST (e.g. MERFISH) would greatly increase the impact of this work. Authors should briefly discuss the limitations of their work regarding the application to other technologies and potential steps needed to extend STERLING to such platforms.

Minor comments
Line 133: Authors claim that "... plausibility score is significantly higher for cells with no neighbours than those with neighbours…" - Authors present statistical proof of this claim (test for statistical significance) or adjust the wording (e.g. use "considerably higher").
Figure 2: Authors should consider changing subfigure C to provide a better comparison of the individual plausibility scores. Consider using grid lines. Similarly, for sub-figure B, it is hard to compare the results from the different clustering algorithms. Consider plotting them side by side.

Line 142: The hyperparameter λ was not previously introduced and is only explained in the method section. Consider introducing this important hyperparameter earlier when the STERLING algorithm is first explained (first paragraph in Result section, Lines 99-112).

Line 142: "Note that given the multiple ways a segmentation error can occur beyond doublets we do not expect the segmentation error probability to be zero in cells where the nuclear segmentation matches the whole-cell segmentation." - It is not entirely clear what is meant by this statement. Please define "doublets".

Figure 3A: Line 159 states: "The results in Fig. 3A shows a clear increase in segmentation
error probability for cells whose nuclear segmentation conflicts with the whole-cell Segmentation." - Subfigure 3A is too small and too crowded to visually observe this increase. Please consider making this plot bigger or more easy to read. For example, consider adding grid lines or removing some preprocessing plots (e.g. CF=5 and CF=10) to give the figure more space. Alternatively, since B and C are well visible, consider making A full width and place B and C below.

Line 230-233: Applying STARLING resulted in 25 clusters of which only 9 could be associated with actual cell types. Compared to the previous experiment the fraction of associated cell clusters is low. Please elaborate on why only 9 clusters could be associated.
Line 236/237: "Notably, interpretation of cell clusters was improved compared to baseline methods (S. Fig. 4)" - Please elaborate on how the cell clustering was improved and what the baseline is. Should this not also refer to Figure 5 (as the rest of the paragraph)?

Line 300-304: Consider removing this statement about previous work. It has limited relevance for the discussion. If an a priori list of cell types is known and the assay contains marker genes/proteins for each of them, identifying segmentation errors becomes somewhat trivial. Furthermore, there is no need for clustering anymore as the cell types can be called based on the marker protein alone. However, this is typically not the case in most experiments.

Line 457- 459: Please explain the importance and impact of these hyperparameters. How were the parameters chosen and how are they impacting the results?

References
[1] Petukhov, V., Xu, R.J., Soldatov, R.A. et al. Cell segmentation in imaging-based spatial transcriptomics. Nat Biotechnol 40, 345–354 (2022). https://doi.org/10.1038/s41587-021-01044-w
[2] Qian, X., Harris, K.D., Hauling, T. et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nat Methods 17, 101–106 (2020). https://doi.org/10.1038/s41592-019-0631-4

(Remarks on code availability)
The authors provide a well-structured code repository including a README, installation instructions and tutorials. The code seems to be well-written and clean. The authors created various tests and a CI pipeline to maintain the code.
Installation instructions clearly define the necessary steps and an additional Docker file is provided.

Reviewer #4

(Remarks to the Author)

The authors propose STARLING, a probabilistic machine learning model to address errors from cell segmentation in highly multiplexed imaging data, focusing on errors that lead to the expression of implausible protein combinations in cell clusters. The authors further propose a plausibility score to quantify how biologically realistic clusters are, and metrics to quantify the ability to detect cell segmentation errors. The applicability of STARLING is further demonstrated on IMC data of controlled cell lines and human tonsil ROIs.

Overall, in my opinion, this paper addresses an important problem and presents interesting approaches to quantify this issue. However, I had difficulty appreciating the benefits of STARLING over existing methods, and its stability across datasets, settings, cell types and morphologies. I have the following comments:

1. The manuscript can be strengthened by including comparisons of outputs from STARLING and existing/state-of-the-art doublet detection methods. Otherwise, the specific advantages of STARLING over approaches that seek to address similar issues is unclear.
2. How comprehensive does the list of protein pairs need to be and how does the number of pairs affect the performance of STARLING / plausibility score?
3. Figure 2B – different cell types may be more or less likely found in densely packed regions (i.e., with more neighbours). How does the plausibility score look like broken down across the different cell types?
4. Line 124 / Fig 2A / Supplementary Table 2 – can the authors provide results for a more comprehensive set of thresholds to better understand the effects. How does the threshold value affect the plausibility score across different datasets? Can the authors also confirm if the threshold applies to all protein pairs (and mutually exclusive/conditional).
5. The method seems sensitive to the settings used, with the optimal settings being inconsistent for different datasets/tissue types (Fig 2C). This can impede generalisability of the method, and the user may need to spend a long time searching for the best settings.
6. Fig 3A – Can the authors provide an explanation as to why probabilities can be below 0 or above 1? Furthermore, as additional validation, it would be useful to include plots of the probabilities for nuclei only. The expectation is that nuclei are relatively pure are without contamination.
7. I suggest the inclusion of a figure showing examples of cells with more than one nucleus and the predicted segmentation error probability associated with the cells, for better understanding.
8. Regarding the metric based the on ratio of convex area to total area, how robust is this metric for cells of different shapes and cell types? I'm wondering because some cell types may have typical shapes (e.g., round / oval / long).
9. Additionally, how robust are the proposed metrics to the segmentation method used? Currently only Mesmer is used. What is the sensitivity of the metrics methods to different characteristics of the segmentation method (e.g., the method could produce more complex cell shapes, e.g. with protrusions).
10. Although the a priori known set of protein pairs contains some commonly known pairs, I suggest adding some references to clarify their source.
11. I think the utility of STARLING is not fully clear in Figs 5/6. What are the results like when STARLING is not used?

Minor comments:

1. Inconsistent use of style for names, e.g. Mesmer and MESMER
2. Fig 2A – hard to tell apart the colours in the pairs. Also consider using proteins C and D for the conditional example.
3. Usually referred to as the METABRIC dataset rather than Metabrics.
4. Repetition between lines 108-112 with 101-102
5. Figs 4F and FG are unclear and visually busy, obscuring their message. I suggest they be redrawn.

(Remarks on code availability)
1. I followed the installation steps using virtualenvwrapper, but was unable to proceed past installing virtualenvwrapper. The next step (mkvirtualenv starling) yielded the error: mkvirtualenv: command not found. I was also unable to install using poetry. It stalled at - Installing pytorch-lightning (2.1.0): Pending... - Installing scanpy (1.9.5) for at least 90 minutes. I was using Ubuntu 18.04.1 and Python 3.6.8.
2. The Colab notebook ran without issue. Please check typos in the notebook, e.g. "Showing initial expression centriods", "Setting trainning log"
3. Some documentation about the contents of the input data file (the .h5ad file) would help with usability.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
All my questions has been addressed and I have no further comments on the manuscript.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed the concerns raised in the review.
In particular, the authors have included an example on Akoya PhenoCycler demonstrating STARLING's ability to generalize to other multiplexed imaging technologies which increases the impact of the manuscript considerably.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors have addressed the comments satisfactorily. The manuscript has improved. The additional results help to show the robustness of STARLING and is informative for potential users.

(Remarks on code availability)
I appreciate that the authors created a Python package - I was able to run and install it.

Reply to reviewers – Lee et al. 2024

We are very grateful to the reviewers for the time taken to review our manuscript and their comments and suggestions, to which we reply below. Our replies are in **blue** and updates to the manuscript are inline in **red**.

Reviewer 1

Major concerns:

1. The evaluation of the approach is indirect and lacks ground truth. Including simulated data, such as mixed data generated from single-cell RNA-seq or single-cell proteomics, could enhance the accuracy assessment of identifying cell segmentation errors and cell phenotypes after applying the STARLING method.

We thank the reviewer for raising this important point. Deciding the correct benchmarking strategy was something we (and the field) has struggled with. The fundamental issue in tackling this problem is that we lack a robust single-cell ground truth. Even expert trained pathologists struggle to manually annotate cells that are robustly segmented (see e.g. **Fig. 4C**). Our initial reluctance to include simulations as a benchmarking strategy stems from the fact that it is hard to realistically simulate multiplexed imaging data, particularly since there is no true single-cell multiplexed imaging data as a starting point. Therefore, we focussed on metrics that recapitulate the ultimate end goal of the analysis, which is to infer biologically meaningful clusters, and therefore introduced and validated the plausibility score. We hope that the formalization of this is on its own useful for the field.

We would like to emphasize that we did generate IMC data of controlled cell pellets of different cell lines. Given each cell line uniquely over-expresses a single marker, each resulting cluster should uniquely over-express a single marker, giving us a way to quantitatively and directly benchmark different approaches. However, the reviewer is correct we did not benchmark this against the full range of clustering approaches in the paper. We now include this in the revised manuscript and find that for all clustering methods and thresholds used to assign clusters to lineages, STARLING outperforms existing baselines:

Supplementary Table 2. Threshold used for discriminate cell types

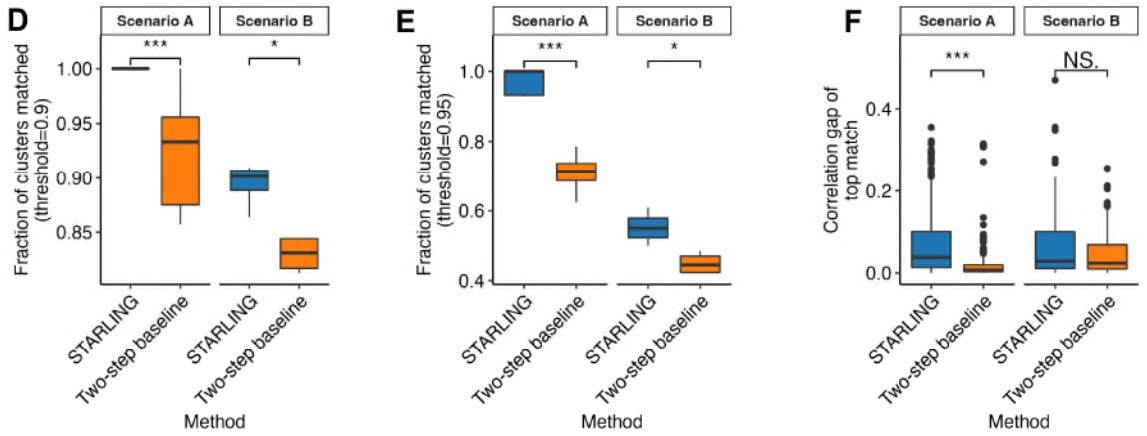
Clustering Method	Identifiability Threshold	Number of unidentifiable initial clusters	Number of unidentifiable STARLING clusters
Phenograph	0.1	4	0
Phenograph	0.15	4	1
Phenograph	0.2	5	1
FlowSom	0.1	4	2

FlowSom	0.15	6	4
FlowSom	0.2	6	5
K-means	0.1	2	1
K-means	0.15	3	1
K-means	0.2	5	2

However, we appreciate the reviewer's suggestion that *in silico* mixing of suspension mass cytometry followed by deconvolution using STARLING along with a baseline adds a valuable aspect to the proposed benchmarking. To do this, we have performed the experiment as requested by the reviewer, with the approach now outlined in the revised manuscript:

Finally, we assessed the ability of STARLING to recapitulate single-cell clusters in the presence of segmentation errors by *in silico* mixing of pure single-cells from suspension mass cytometry. We retrieved single-cell data from a recent paper[3] and simulated segmentation errors by randomly mixing two cells with 80% frequency or three cells with 20% frequency. For this we created two scenarios: scenario *A* where single cells are mixed with equal proportion, and scenario *B* where cells are mixed in an imbalanced setting (70:30 for two cells, 50:25:25 for three). We next clustered the single-cell suspension data, then ran STARLING on the *in silico* mixed data along with the two-step procedure where a random forest classifier is first trained to identify cells with likely segmentation errors, which are then removed before clustering, with the same clustering algorithm used throughout to minimize sources of variability (**Methods**). We subsequently matched the inferred clusters from the mixed data to the original from the single-cell data via correlating the cluster centroids. Overall, STARLING matched significantly more clusters to the single-cell data using different thresholds on the correlation (**Fig. 3D-E**). In addition, the gap between the highest and second highest correlations between the single-cell and mixed data was higher on average for STARLING than the two step approach (**Fig. 3F**), implying clusters could be more uniquely assigned. Together, these results suggest STARLING can infer clusters closer to those obtained from true single-cell data compared to existing approaches.

With updated Fig. 3D-F:



We believe this now adds strong quantitative evidence on top of the existing ground truth data in the manuscript that STARLING can re-infer single-cell populations in the presence of segmentation errors even in the absence of *a priori* knowledge about what cell populations should be present.

2. A comprehensive review of similar approaches and benchmarking against those approaches is warranted.

Our motivation for developing STARLING was that no equivalent method existed—the closest methods are those used to infer clusters from multiplexed imaging, against which we benchmarked three common methods (PhenoGraph, FlowSOM, K-Means). To our knowledge, STARLING is the first (and remains only) computational method for quantifying cell phenotypes from highly multiplexed imaging while accounting for cell segmentation errors. This was outlined in the introduction of the original paper with

To-date, no methods have been proposed to learn denoised cellular phenotypes while accounting for cell segmentation errors in highly multiplexed imaging.

Despite this, we attempted to collate a strong set of baseline methods for this task for us to compare against (and hopefully improve upon). This involved developing a separate two-step approach whereby a supervised learning model (random forest) was trained to identify likely segmentation errors after which these are removed and the cells re-clustered using each of the above clustering methods (though note that this has the distinct disadvantage that all cells no longer have probabilities of cluster association as per STARLING). If the reviewer knows of existing methods that specifically learn cell clusters while accounting for segmentation errors, we would of course be happy to compare. Alternatively, if they believe the manuscript would be improved by referencing in depth more of the above discussion, we would be happy to do so.

3. It would be beneficial to provide a discussion and guidance on the selection of hyperparameters, such as λ in the loss function and the number of clusters.

We apologize this was omitted from the original manuscript—while we sought to show robustness to the choice of hyperparameters, we did not provide useful guidance for setting them which is of course important for the end-user. Given our plausibility score benchmarking consistently shows highest values for λ=0.1, we now recommend this value. We have added the following to the manuscript:

While we show the robustness of the results to choice of λ, we recommend λ=0.1 as it returns the strongest metrics in benchmarking (below) and use this consistently thereafter.

Regarding the second point, the correct number of clusters in any single-cell analysis is both a contentious and variable decision that we believe is best left to the user to decide in line with existing publications. As such, we have added the following to the discussion:

As most existing single-cell clustering algorithms allow the user to set the number of clusters that best aligns with prior biological knowledge—either directly or indirectly via resolution parameters in approaches such as PhenoGraph— the user's choice of initialization guides this

In light of this, several figures in the remainder of the paper have changed as we have sought to consistently use λ=0.1 hereafter.

4. The method can only identify cell overlaps involving two cells expressing different marker proteins, making it infeasible for cells of the same type. Similarly, STARLING does not address cell overlaps involving more than two cells.

Regarding the first point, it is correct that if two (or more) cells expressing identical marker combinations are mis-segmented as a single cell, then STARLING will not find them. However, this is not considered a weakness for two reasons. First, if the two cells come from identical cell types and the segmentation algorithm applied cannot tell them apart, then it is unlikely any method could discern them as there is simply no information in the image to do so with. Second, if our aim is to identify the cell clusters in an image, then having two identical cells co-segmented as one won't disadvantage STARLING (or any other clustering algorithm) as it still represents a "clean" read-out of a single cell cluster. This was outlined in the original submission via

The consequence is that when the error involves cells from different cell types — previously described as heterotypic doublets[24] — clusters are formed due to the composition of different cells within a segment that are not "real" cell types in the source biological material.

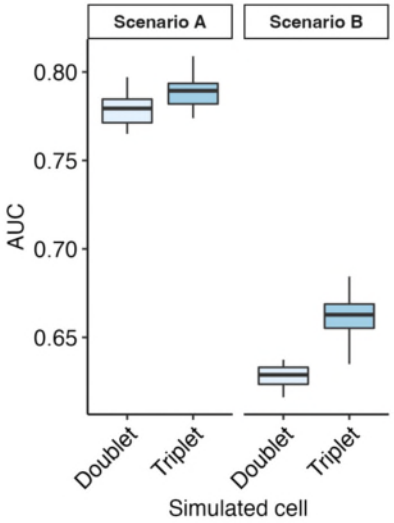
We now make it explicit that STARLING is designed to denoise only segmentation errors from multiple cell types:

*To help address this, we developed STARLING (**S**egmen**T**ation **A**waRe c**L**uster**ING**), a novel probabilistic clustering method to identify true cell phenotypes in the presence of segmentation errors involving cells from multiple different phenotypes.*

Regarding the second point, STARLING *can* denoise cell expression when the segmentation involves more than two cells, but we agree with the reviewer this was both poorly explained and poorly demonstrated in the original manuscript. To correct this, we have added the following explanation in the results (it was previously alluded to in discussion):

While the segmentation errors are modelled as pairwise combinations of underlying cell phenotypes, by averaging over all possible pairings STARLING can model true cluster identities in the presence of multiple individual cell phenotypes contributing to a segmentation error.

Second, inspired by the reviewer's suggestion above of *in silico* mixing different cells from suspension mass cytometry, rather than creating segmentation errors by mixing only two cells ("doublet"), we also mixed three ("triplet") cells in varying proportions and benchmarked the results. These are included in the above simulations and results, showing that STARLING can indeed reverse out cell phenotypes closer to "ground truth" when there are segmentation errors comprising >3 cells. In addition, we extracted the segmentation error probabilities from the STARLING model, and contrasted the area under the ROC curve of STARLING discriminating two cell errors vs three cell errors:



Perhaps counter-intuitively, STARLING performs better at discriminating triplets vs. doublets. To understand why, we appeal again to the above reasoning in that to compute the likelihood of a segmentation error, STARLING averages over all possible pairings (not any given one). We have added the above as a supplementary figure in the manuscript, and added the following to the results:

Interestingly, STARLING exhibited a higher ability to detect mis-segmented cells formed of three cell phenotypes compared to two (**S. Fig.**), in line with the intuition that averaging over all cell pairs when computing segmentation errors enables detection of multi-phenotype errors.

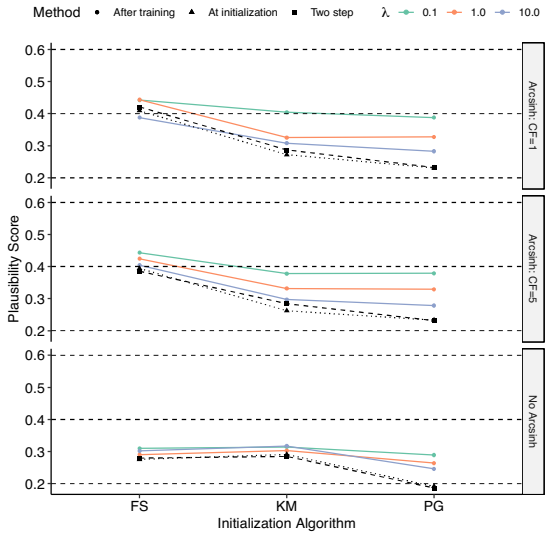
5. The underlying principle of the approach suggests its potential application to other data types, such as spatial transcriptomes. Demonstrating this extension would further enhance the method's versatility and applicability.

We thank the reviewer for highlighting this point. To demonstrate this, we know apply STARLING to an example Akoya PhenoCycler (previously CODEX) dataset. We have added the following text to the manuscript:

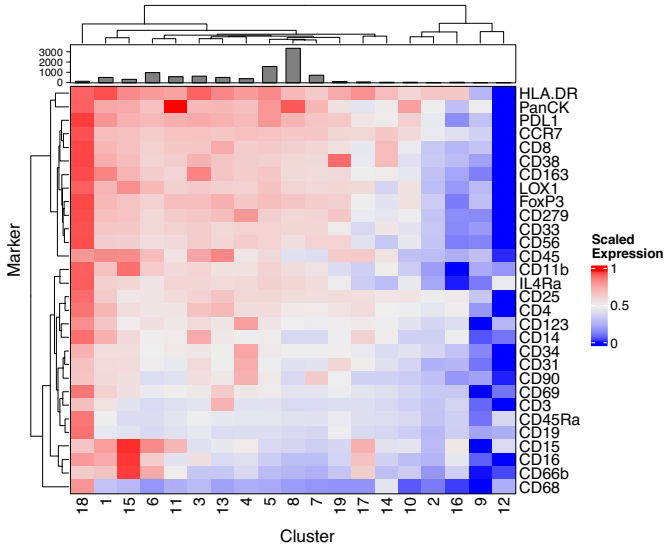
To demonstrate the applicability of STARLING to multiplexed imaging technologies beyond IMC, we retrieved Akoya PhenoCycler (previously Codex) data of the tumour immune microenvironment of head and neck squamous cell carcinoma[3]. We applied STARLING using multiple hyperparameter combinations and initial clusterings as before for an equivalent set of markers to calculate the plausibility score (**S. Table**). Contrasting the plausibility scores from the initial to STARLING clusters shows an overall increase in all but one scenario (**S. Fig**). Indeed, comparing cellular phenotypes returned using FlowSOM to initialize demonstrates several improvements in interpretation compared to the initialization (**S. Fig**). For example, at initialization, a cluster is present (19) with relatively high CD31 (endothelial marker) and CD38 (hematopoietic marker[4]), which is no longer present after fitting with STARLING, which instead exhibits a single cluster overexpressing markers known to be expressed in endothelial cells such as CD31, CD34, and CD90. Despite this, STARLING still assigns a small number of cells to a cluster expressing almost every marker (18), highlighting the fundamental difficulty in completely resolving all cell phenotypes from such data. Together, these results suggest STARLING will have strong utility for highly multiplexed imaging technologies beyond IMC.

With the relevant new supplementary figures referenced:

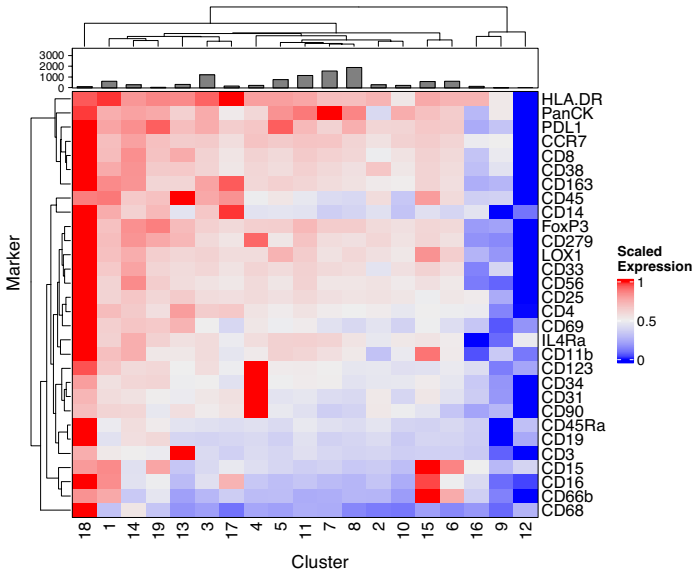
Plausibility scores for Codex:



Initial clustering example:



Clustering after STARLING:



Minor concerns:

1. In Figure 2A, although the co-expression conditions are explained in the methods, it would be beneficial to include direct explanations in the figure legend to improve readability. In addition, the colors are difficult to distinguish.

We have updated Fig. 2A as requested using a different colour palette to better distinguish the centroids.

2. In Figure 3A, the differences in cell segmentation error probabilities appear to be minimal between different groups. Please provide information on the statistical test used and the corresponding p-values to support this observation.

We have remade Fig. 3A as requested to be clearer and include p-values.

3. In Figure S1, while it can be seen that there are more co-expression phenotypes for CD20 and CD3 in the third panel, the first two panels do not seem to show a significant number. How can it be proven that these are caused by cell segmentation errors? It would be helpful to use STARLING and present the results to assess if there is an improvement in the identification.

We thank the reviewer for highlighting this. A limitation of STARLING is it cannot provide denoised cell expression profiles for individual cells, only cluster level quantification, so we are unable to investigate this. We have added this to the discussion:

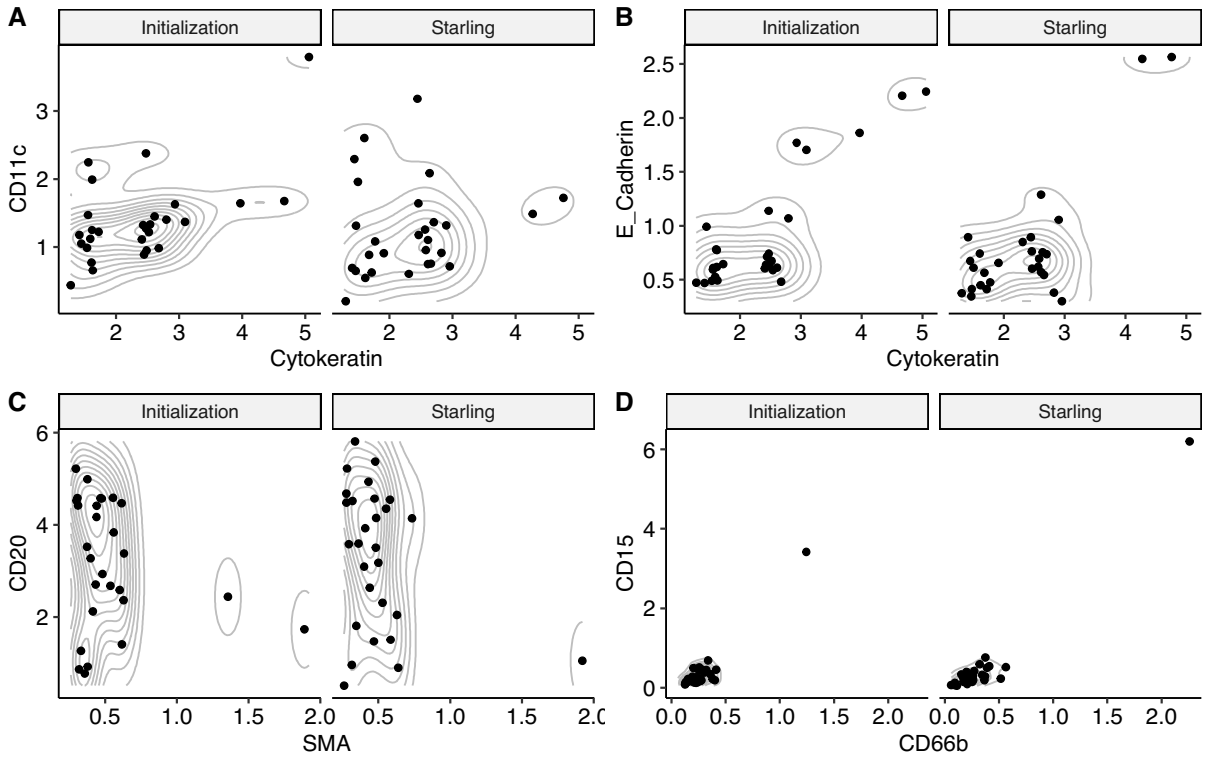
Finally, STARLING cannot provide denoised expression profiles for individual cells that would be highly valuable for quantifying whether an observed expression pattern is the result of segmentation errors or real biological variation. Future work developing this will be important for refined interpretation of highly multiplexed imaging data.

4. In Figure S4, the author mentioned that CD3 and CD20 exhibit lower expression levels in B and T cell clusters in the Baseline method heatmaps. However, this is not clearly visible in the figure. It would be informative to directly compare the expression differences of CD3 and CD20 specifically in the B and T cell clusters among different methods. Additionally, comparing the number of B and T cells identified in each method would be valuable.

We agree this was lacking in the original submission. The reason we focused on CD3/CD20 was it was how we first became aware of this problem, motivating the investigation and creation of the method. However, in the Tonsil example at hand, while STARLING does find fewer clusters that co-express these markers, it is not the best example to show where cell type interpretation has been improved (i.e. there are still clusters that co-express these markers). Instead, in response to points raised by other reviewers, we have focussed on where STARLING did in fact improve the interpretation of the clusters in the tonsil data. We have added the following to the main text:

Notably, the interpretation of multiple cell clusters was improved compared to the initialization clustering (**S. Fig.**). For example, at initialization a cluster was present with high co-expression of CD11c (expressed on various leukocytes) and Cytokeratin (expressed on epithelial cells), which was no longer present after applying STARLING (**S. Fig. A**). Similarly, STARLING identified two clear populations of epithelial cells denoted by high co-expression of E-Cadherin and Cytokeratin, while at initialization multiple other clusters were observed with high E-Cadherin expression, but moderate level of Cytokeratin expression similar to that observed on E-Cadherin-low clusters (**S. Fig. B**). In addition, at initialization a cluster expressing moderate CD20 (a B cell marker) and Smooth Muscle Actin (SMA, expressed on certain endothelial cells) was present, while after applying STARLING the only cluster remaining with high SMA expression had low CD20 expression (**S. Fig. C**). Finally, at initialization, the CD66b+CD15+ population demarcating potential granulocytes contained only moderate expression of each, while STARLING adjusted this cluster to contain almost double the expression (**S. Fig. D**). Together, these results demonstrate several ways in which cellular interpretation was improved compared to the default initialization.

with the relevant supplementary figure:



5. There's a minor error in line 211 of the article. I think that the correct figure references should be "for both PhenoGraph (Fig. 4F) and STARLING (Fig. 4G)," not the original "for both PhenoGraph (Fig. 4G) and STARLING (Fig. 4H)."

Thank you, this has been corrected.

Reviewer 2

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for their time spent reviewing our work.

Reviewer 3

Main comments
Introduction
Define the scope more clearly
The authors start by introducing IMC as one multiplexed image technology (Line 32) but do not follow up on any of the other multiplexed image technologies. The rest of the introduction seems to indicate that this method is primarily developed for IMC as all the benchmarking data is generated using IMC. The authors should specify clearly if their method is specifically developed for Imaging Mass Cytometry (IMC) or if it can or could be applied to similar multiplex image technologies such as image-based Spatial Transcriptomics (ST).

We thank the reviewer for this comment. While the method has been primarily tested on IMC given our primary research focus, in response to all reviewers we now include a further example on Akoya PhenoCycler (previously Codex). To make this explicit in the introduction, we now include the following:

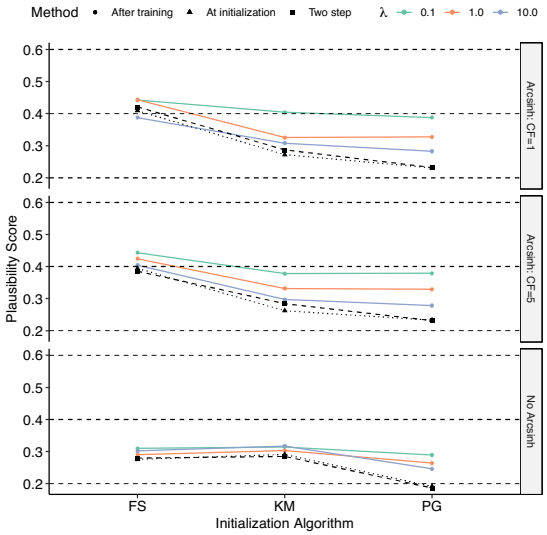
While we primarily focus on IMC, we also apply STARLING to Akoya PhenoCycler (previously Codex) data, and show that in this setting it also improves the plausibility scores of the resulting clustering.

To show this, we have included an analysis using Codex data. We have added the following text to the manuscript:

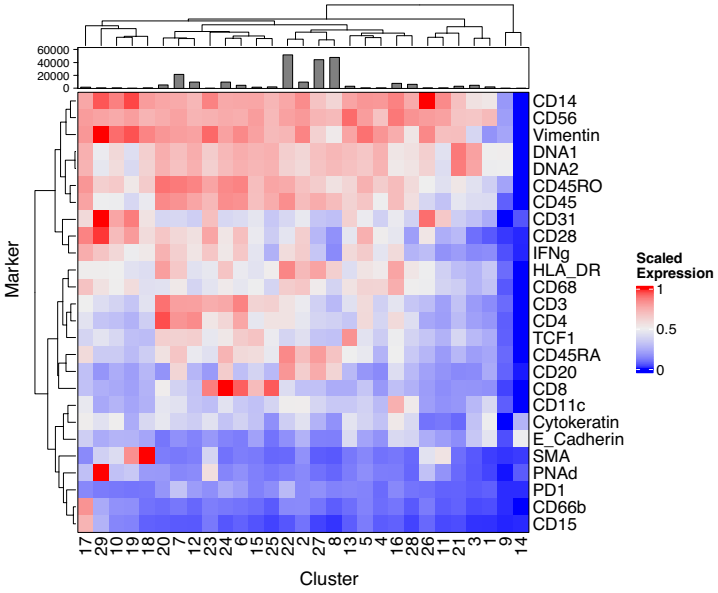
To demonstrate the applicability of STARLING to multiplexed imaging technologies beyond IMC, we retrieved Akoya PhenoCycler (previously Codex) data of the tumour immune microenvironment of head and neck squamous cell carcinoma[3]. We applied STARLING using multiple hyperparameter combinations and initial clusterings as before for an equivalent set of markers to calculate the plausibility score (**S. Table**). Contrasting the plausibility scores from the initial to STARLING clusters shows an overall increase in all but one scenario (**S. Fig**). Indeed, comparing cellular phenotypes returned using FlowSOM to initialize demonstrates several improvements in interpretation compared to the initialization (**S. Fig**). For example, at initialization, a cluster is present (19) with relatively high CD31 (endothelial marker) and CD38 (hematopoietic marker[4]), which is no longer present after fitting with STARLING, which instead exhibits a single cluster overexpressing markers known to be expressed in endothelial cells such as CD31, CD34, and CD90. Despite this, STARLING still assigns a small number of cells to a cluster expressing almost every marker (18), highlighting the fundamental difficulty in completely resolving all cell phenotypes from such data. Together, these results suggest STARLING will have strong utility for highly multiplexed imaging technologies beyond IMC.

With the relevant new supplementary figures referenced:

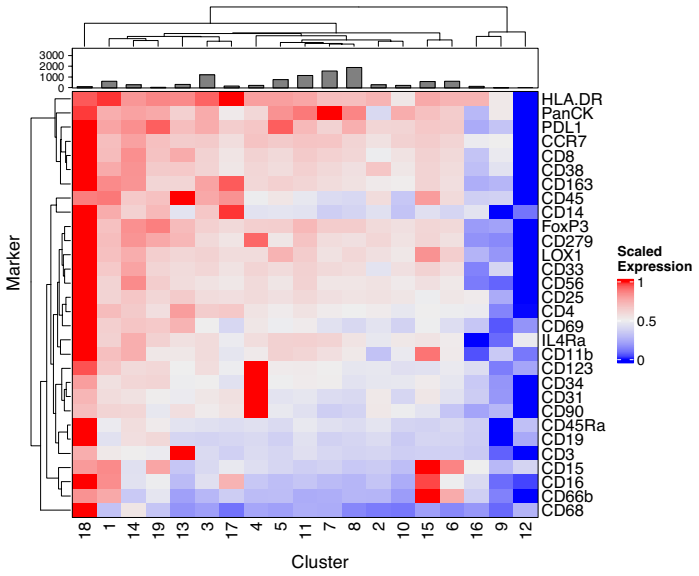
Plausibility scores for Codex:



Initial clustering example:



Clustering after STARLING:



We do not apply STARLING to spatial transcriptomic data, as the widespread availability of single-cell transcriptomic references in this area provide more information that is not currently exploited by STARLING. This is referenced in the reply to the reviewer's point *Limitations and related work.*

Define the objective more clearly

The authors should specify more clearly in the introduction what the purpose and objectives of their method are: The authors state on Line 71, that STARLING is a novel clustering method. However, in the result section (lines 139 to 142) it is noted that STARLING is initialized with some clustering and then computes a de-noised clustering along the per-cell segmentation error probabilities (also shown in Figure 1). From reading the introduction it is unclear if their method presents a novel clustering algorithm or a post-processing algorithm for denoising clusters and identification of segmentation errors. The definition of the method and the fact it is initialized with a clustering strongly suggests the latter: a post-processing method for clustering algorithms. I recommend making the definition and purpose of STARLING more precise in the introduction. Additionally, I suggest adding one sentence explaining the method just after introducing it on Line 73. Something like: "STARLING takes as input a clustering and the potentially erroneous cell segmentation masks and applies a probabilistic optimization algorithm to output a corrected clustering". This will clarify the objective and help readers to understand the method and manuscript better.

We thank the reviewer for this insight and agree the paper needs clarity regarding this point. With regards to whether STARLING represents a novel clustering algorithm or not, we would argue yes, since many clustering algorithms require an initialization, then optimize some loss function to arrive at a final clustering. For example, Gaussian Mixture models in the sklearn python package are initialized using a different clustering algorithm (kmeans) (discussion at https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_init.html), and then go on to optimize

an objective function (the negative log likelihood). In the same manner, STARLING takes an initial clustering, optimizes a negative log likelihood, and returns a clustering, so in as much as Gaussian mixture models represent a clustering algorithm, so does STARLING.

With regards to why we initialize to a clustering solution rather than randomly (as does e.g. kmeans), we reasoned that the clusterings returned by most existing methods are clearly not just noise and contain real biological variability. Therefore, if we really want to arrive at the best cell phenotypes, we should start there and improve, rather than start from scratch. This also enables an apples-to-apples comparison of what STARLING is improving, rather than comparing to a random baseline.

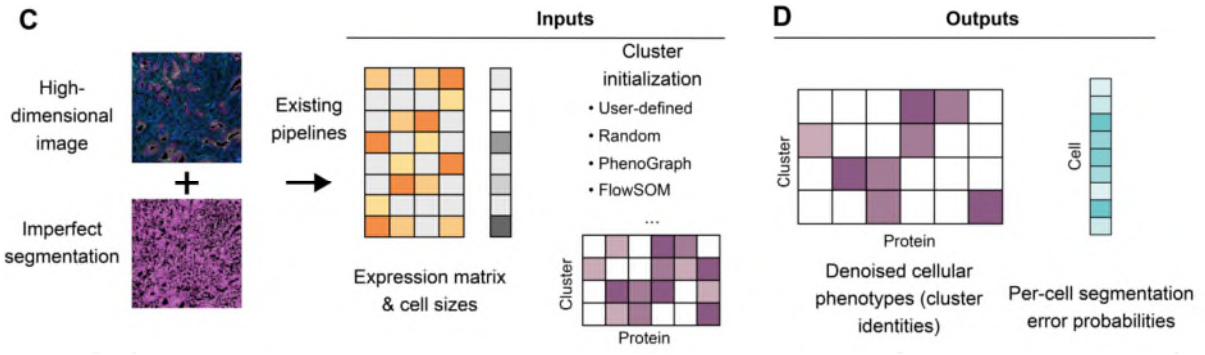
In addition, we were ultimately striving for honesty here—given there are multiple ways to initialize a clustering algorithm (STARLING), we wanted to evaluate all of them (and specifically the improvement over the initialization) rather than just selecting a single one and calling it a day.

However, we agree none of this was made clear in the original manuscript. Therefore, we have updated it as follows:

In addition, as with many existing cluster algorithms, STARLING takes as input a cluster initialization that may be either user-defined or generated using existing clustering algorithms[1,2]. The reason for this choice was twofold: (i) existing clustering algorithms applied to multiplexed imaging clearly learn cellular phenotypes that are not just noise, so these should form an important starting point to improve upon, and (ii) since the initialization choice reflects an important source of variability in algorithm performance, an honest benchmarking approach will examine multiple different initializations. As such, most subsequent benchmarking compares metrics computed from the STARLING clusters compared to the initialization, as this reflects the value-add rather than compared to a random clustering.

Results
STARLING Input data - Lines 99-112
It is unclear why the high-dimensional image is required for STARLING other than for the generation of the clusters. From an algorithmic point of view, it would be more “clean” to take as input a pre-computed clustering and the imperfect cell segmentation masks. This goes in a similar direction to a previous comment, that the authors should specify more clearly what STARLING does: Is it a clustering algorithm or a post-processing method that is initialized with a clustering such as KMeans or PhenoGraph? Decoupling STERLING from the initialized clustering algorithms would provide a more intuitive algorithmic framework and might allow the application of STERLING to other multiplex image technologies.
The authors should consider making this change or clearly state why this design decision was taken.

We apologize this was initially unclear and hope the above discussion and modifications to the text help address this. In addition, we have updated **Fig. 1C** to make it explicit the inputs are the summarized expression along with a cluster initialization, and specifically that a user-defined cluster initialization may be use, thereby increasing the generalization of the approach:



neighboring and non-neighboring cells
The plausibility score is evaluated by comparing cells with "no neighbors" to cells in densely packed regions. However, the definition of neighboring and non-neighboring cells is not clearly defined.
Authors should clarify their definition of neighboring cells: What is the quantification of two adjacent cells? Similarly, Figure 2B compares cells with no neighbors to cells with neighbors. This term might be misleading or confusing for readers. It would be more appropriate to compare cells in regions of low vs. high cell density.
Authors might consider a more appropriate definition of cells in densely packed tissue regions vs cells in low-density regions for example the mean/max/min distance to its neighboring cells.
In any case, the authors should specify the definition in the main manuscript and not in the Methods section to help readers understand the benchmark results better.

We apologize the manuscript was unclear with this respect. We defined two cells as neighbouring if there is no gap between their respective segmentation masks, i.e. the first cell has at least one pixel in its segmentation mask that is directly adjacent to the second cell. We have updated this in the main results of the manuscript:

We define two cells as neighbouring if the first cell has at least one pixel in its segmentation mask that is directly adjacent in the image to any pixel in the segmentation mask of the second cell, and not neighbouring otherwise.

We agree an extension to low- and high-density regions would be interesting, but given the additional set of choices needing defined (how big is a region, how do we define low vs. high, how do we deal with variable density within the region), we believe the more simple approach of defining cells as having neighbours or not is equally as informative.

Plausibility score
The authors introduce the notion of “plausibility scores” to benchmark the differences in

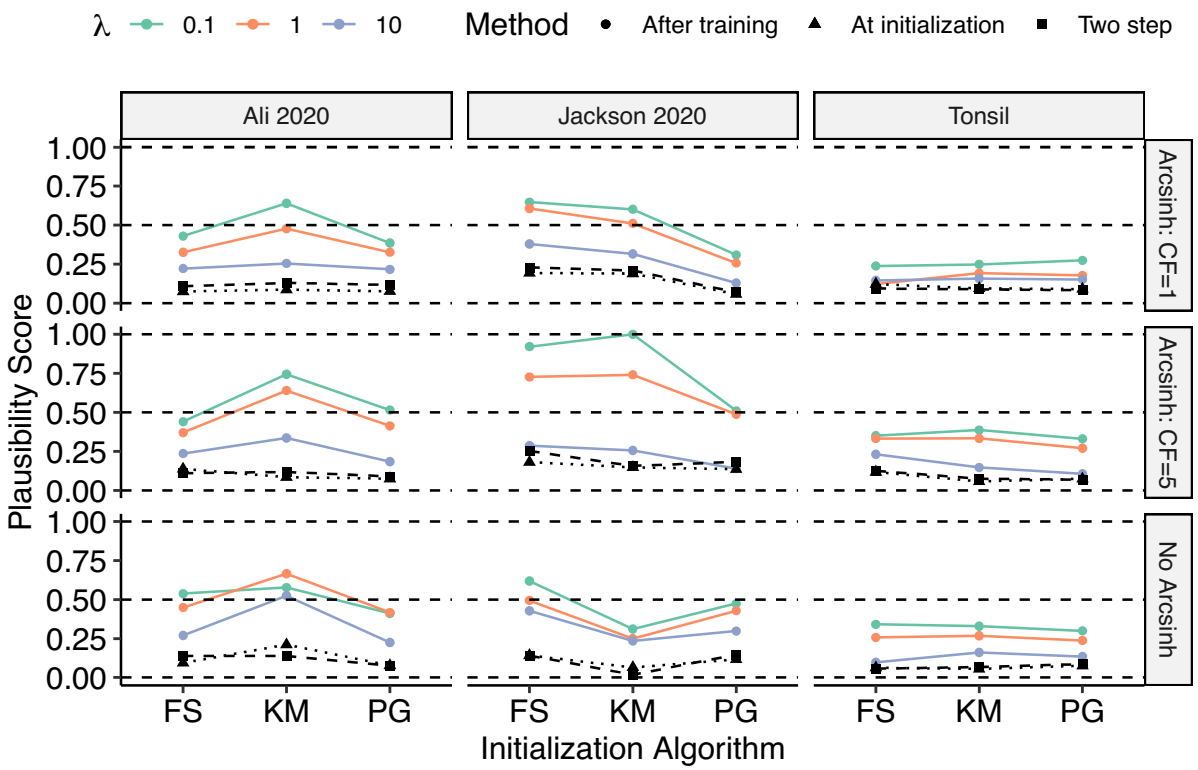
clusterings. Plausibility scores are calculated based on an a priori known set of proteins that either exhibit mutually exclusive expression or conditional co-expression. Plausibility is obtained by calculating the proportions of proteins that are co-expressed. Although this score seems reasonable to estimate the segmentation results, it is unclear how the absolute scores can be interpreted. There is no range of “good and bad” scores defined. Figure 2C shows improvements in plausibility scores but it is not possible to estimate the significance of these improvements.
It would be more appropriate to use statistical tests to determine the plausibility of segmentation. For example: Calculate the p-value of two mutually exclusive proteins where the null hypothesis is that these are co-expressed. Such statistics would directly identify clusters that do not correspond to actual cell types and give a better understanding of the plausibility scores.

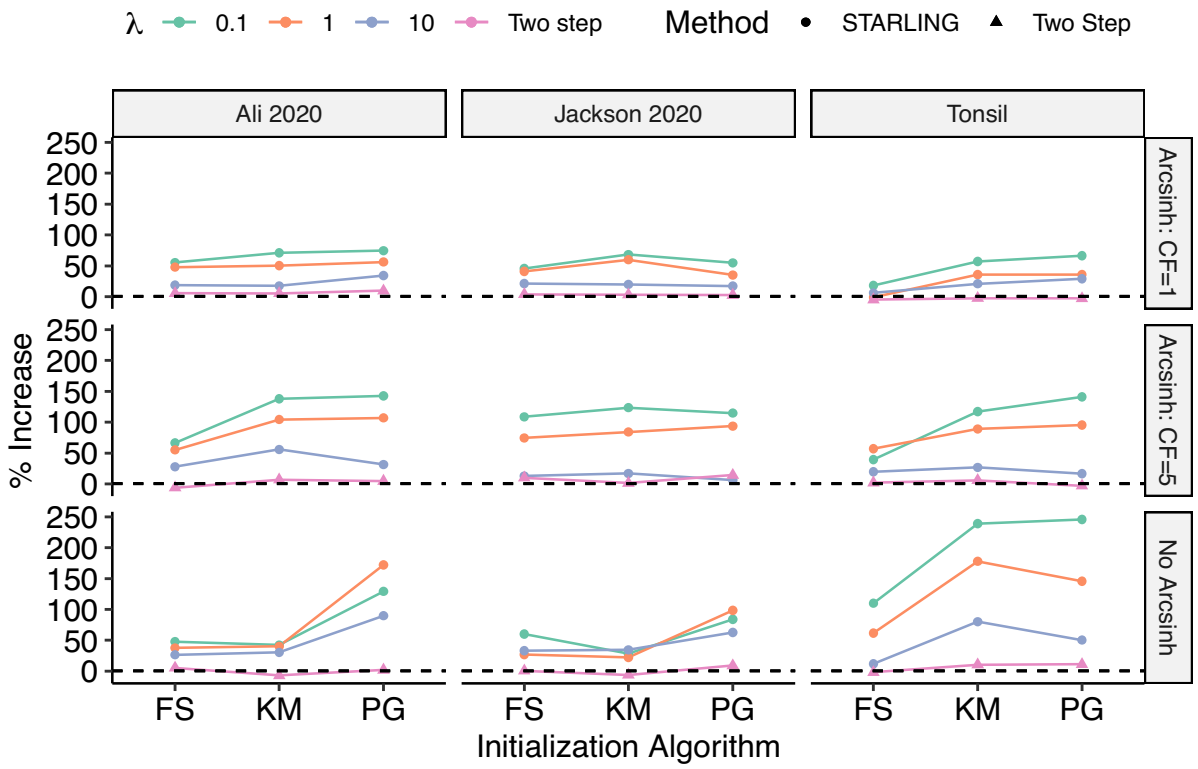
We agree with the reviewer that having some method of discerning absolute improvement is desirable. The issue with assigning a p-value and why we steered away from it in the original paper is two-fold: (i) the pairs are not statistically independent (since a given protein can appear in more than one pair, and we have e.g. different isoforms of the same protein) making it invalid for most two-sample tests, and (ii) the number of pairs to include is a user preference (we selected a set we considered reasonable, but in reality could have included many more) meaning we could get an arbitrarily low p-value by including more pairs which is scientifically unsatisfactory.

However, we appreciate the importance of trying to contextualize what a “good” vs “poor” plausibility score is, we have exploited the points in the last reply regarding neighbouring and no-neighbouring cells, and for each cohort/method, rescaled the plausibility score such that the minimum (0) represents the average score of cells with neighbours, and the maximum (1) represents the average with no neighbours. Therefore, this rescaled plausibility score can be interpreted as the improvement to the clustering from all cells being densely packed (0) to cells being approximately measured in suspension (1). We can also quantify this as the % improvement We have added the following to the manuscript addressing this:

As the plausibility score represents a relative improvement in cell phenotype quality, we sought to anchor it on an absolute interpretable scale. To do so, for each cohort and method, we compute the plausibility scores on cells with no neighbours and with neighbours as above and rescaled the score for each method such that the average score for cells with no neighbours was 1 and the average score for cells with neighbours was 0. Therefore, the increase in this normalized plausibility score can be interpreted as the improvement an algorithm could make relative to measuring all cells individually rather than densely packed regions. The results show STARLING can in many cases raise this normalized plausibility score closer to 1 than 0, frequently improving the initial plausibility score by >50-100% (**S. Fig.**). Together, these results suggest STARLING has the effect of presenting single-cell phenotypes closer to those that would be quantified if all cells were clearly separated and segmentable across the tissue.

Referencing the following supplementary figures:


λ 0.1 1 10
Method After training At initialization Two step
Ali 2020
Jackson 2020
Tonsil
Arcsinh: CF=1
Arcsinh: CF=5
No Arcsinh
Plausibility Score
1.00
0.75
0.50
0.25
0.00
FS KM PG
Initialization Algorithm


λ 0.1 1 10 Two step
Method STARLING Two Step
Ali 2020
Jackson 2020
Tonsil
Arcsinh: CF=1
Arcsinh: CF=5
No Arcsinh
% Increase
250
200
150
100
50
0
FS KM PG
Initialization Algorithm

Discussion
Limitations and related work
Although this manuscript presents a novel approach to handling segmentation errors in single-cell analysis the overall issue is well-known in similar fields such as ST and has been addressed in other publications. Multiple probabilistic framework for cell segmentation of ST data exists which take into account the gene/protein expressions in addition to nuclei and membrane stains. Such algorithms, such as [1, 2], tackle the same problem of segmentation errors from a different angle by incorporating additional information. It was previously shown that incorporating the gene/protein expression levels into cell segmentation can improve the segmentation result significantly by removing common segmentation errors such as those discussed in this manuscript. STARLING, on the other hand, simply identifies segmentation errors and improves the downstream clustering but does not solve the issue of segmentation errors.
Authors should address these approaches and briefly discuss differences.

We thank the reviewer for highlighting this. We have added to the discussion citing the selected papers and others:

Specifically, highly multiplexed imaging assays that target protein expression lack equivalent single-cell data for a matched tissue that could be used as a reference for deconvolution, meaning STARLING was designed under the assumption no *a priori* knowledge about the pure cell types in the tissue was available. In contrast, in spatial transcriptomics methods tackling similar problems can exploit the widespread availability of scRNA-seq of all tissues and cell types as a reference for deconvolution or segmentation[36–39].

Future work
Adapting STARLING for other multiplex imaging technologies such as image-based ST (e.g. MERFISH) would greatly increase the impact of this work. Authors should briefly discuss the limitations of their work regarding the application to other technologies and potential steps needed to extend STERLING to such platforms.

In response to this along with reviewer 1 point 5, we now include an example on Akoya PhenoCycler (previously CODEX) that we hope demonstrates the applicability of STARLING to other highly multiplexed imaging technologies. We cover this in response to the point *Define the scope more clearly* above.

Minor comments
Line 133: Authors claim that “... plausibility score is significantly higher for cells with no neighbours than those with neighbours…” - Authors present statistical proof of this claim (test for statistical significance) or adjust the wording (e.g. use “considerably higher”).

We have now added significance testing to the revised Figure 2A.

Figure 2: Authors should consider changing subfigure C to provide a better comparison

of the individual plausibility scores. Consider using grid lines. Similarly, for sub-figure B, it is hard to compare the results from the different clustering algorithms. Consider plotting them side by side.

We have added gridlines to Fig 2C.

Line 142: The hyperparameter λ was not previously introduced and is only explained in the method section. Consider introducing this important hyperparameter earlier when the STERLING algorithm is first explained (first paragraph in Result section, Lines 99-112).

We have moved the description of λ to when the method is introduced as well as guidance on what value it should take in response to reviewer #1:

STARLING includes a hyperparameter denoted λ that controls the trade-off between optimizing the model likelihood and ability to detect synthetic segmentation errors as well as different data pre-processing strategies. While we show the robustness of the results to choice of λ, we recommend λ=0.1 as it returns the strongest metrics in benchmarking (below) and use this consistently thereafter.

Line 142: “Note that given the multiple ways a segmentation error can occur beyond doublets we do not expect the segmentation error probability to be zero in cells where the nuclear segmentation matches the whole-cell segmentation.” - It is not entirely clear what is meant by this statement. Please define “doublets”.

We thank the reviewer for catching this. We have updated this example to read:

Note that given the multiple ways a segmentation error can occur beyond this example where we find whole segmented cells that likely have more than one nucleus, we do not expect the segmentation error probability to be zero in cells where the nuclear segmentation matches the whole-cell segmentation.

Figure 3A: Line 159 states: “The results in Fig. 3A shows a clear increase in segmentation
error probability for cells whose nuclear segmentation conflicts with the whole-cell Segmentation.” - Subfigure 3A is too small and too crowded to visually observe this increase. Please consider making this plot bigger or more easy to read. For example, consider adding grid lines or removing some preprocessing plots (e.g. CF=5 and CF=10) to give the figure more space. Alternatively, since B and C are well visible, consider making A full width and place B and C below.

We apologize the initial figure was unclear. We have made multiple aesthetic changes including (i) replacing violin plots with boxplot in 3A, (ii) text size larger throughout, (iii) figure is now close to full page (with additional plots).

Line 230-233: Applying STARLING resulted in 25 clusters of which only 9 could be associated with actual cell types. Compared to the previous experiment the fraction of associated cell clusters is low. Please elaborate on why only 9 clusters could be associated.

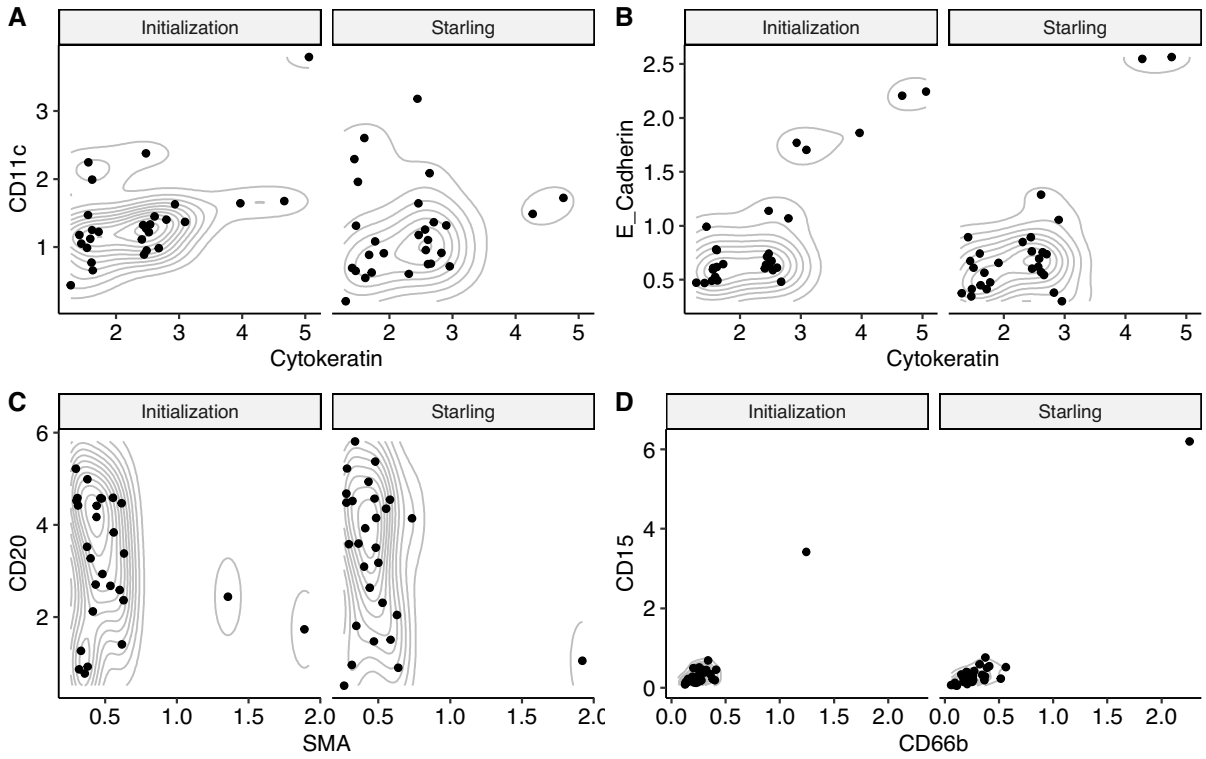
We apologize for the confusion. It was not that only 9/25 clusters could be assigned, it was that the 25 clusters were interpreted as representing sub-clusters of 9 major cell lineages. In fact, only 1 cluster couldn't be interpreted, so the sentence could have read "Of the 25 clusters inferred, 24 were confidently mapped as representing sub-clusters of 9 major cell types. Note that these numbers are no longer exact as we have rerun this entire analysis in an attempt to standardize the hyperparameter settings of STARLING.

Line 236/237: "Notably, interpretation of cell clusters was improved compared to baseline methods (S. Fig. 4)" - Please elaborate on how the cell clustering was improved and what the baseline is. Should this not also refer to Figure 5 (as the rest of the paragraph)?

We apologize for the detail lacking here. We have expanded on this to highlight specific details of how the cell type performance was improved:

Notably, the interpretation of multiple cell clusters was improved compared to the initialization clustering (**S. Fig.**). For example, at initialization a cluster was present with high co-expression of CD11c (expressed on various leukocytes) and Cytokeratin (expressed on epithelial cells), which was no longer present after applying STARLING (**S. Fig. A**). Similarly, STARLING identified two clear populations of epithelial cells denoted by high co-expression of E-Cadherin and Cytokeratin, while at initialization multiple other clusters were observed with high E-Cadherin expression, but moderate level of Cytokeratin expression similar to that observed on E-Cadherin-low clusters (**S. Fig. B**). In addition, at initialization a cluster expressing moderate CD20 (a B cell marker) and Smooth Muscle Actin (SMA, expressed on certain endothelial cells) was present, while after applying STARLING the only cluster remaining with high SMA expression had low CD20 expression (**S. Fig. C**). Finally, at initialization, the CD66b+CD15+ population demarcating potential granulocytes contained only moderate expression of each, while STARLING adjusted this cluster to contain almost double the expression (**S. Fig. D**). Together, these results demonstrate several ways in which cellular interpretation was improved compared to the default initialization.

with the relevant supplementary figure:



Regarding figure referencing, Fig 5c was the post-STARLING fit, while previous S. Fig 4 (now S. Fig. 16) was the initialization heatmap. We hope this has now been clarified by replacing “baseline” with “initialization”.

Line 300-304: Consider removing this statement about previous work. It has limited relevance for the discussion. If an a priori list of cell types is known and the assay contains marker genes/proteins for each of them, identifying segmentation errors becomes somewhat trivial. Furthermore, there is no need for clustering anymore as the cell types can be called based on the marker protein alone. However, this is typically not the case in most experiments.

We have removed this as requested.

Line 457- 459: Please explain the importance and impact of these hyperparameters. How were the parameters chosen and how are they impacting the results?

We have added clarifying sentences on this, specifically:

These control the user’s *a priori* expectation of what the segmentation error rate in a given dataset is expected to be, and may be user defined in a custom analysis should the researcher anticipate a higher or lower error rate.

and

in line with the expectation that *a priori* no cluster should be more prevalent than the others given the cluster interpretations are not known in advance

References

[1] Petukhov, V., Xu, R.J., Soldatov, R.A. et al. Cell segmentation in imaging-based spatial transcriptomics. Nat Biotechnol 40, 345–354 (2022). https://doi.org/10.1038/s41587-021-01044-w

[2] Qian, X., Harris, K.D., Hauling, T. et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nat Methods 17, 101–106 (2020). https://doi.org/10.1038/s41592-019-0631-4

Reviewer #3 (Remarks on code availability):

The authors provide a well-structured code repository including a README, installation instructions and tutorials. The code seems to be well-written and clean. The authors created various tests and a CI pipeline to maintain the code.
Installation instructions clearly define the necessary steps and an additional Docker file is provided.

Thank you.

Reviewer 4

The authors propose STARLING, a probabilistic machine learning model to address errors from cell segmentation in highly multiplexed imaging data, focusing on errors that lead to the expression of implausible protein combinations in cell clusters. The authors further propose a plausibility score to quantify how biologically realistic clusters are, and metrics to quantify the ability to detect cell segmentation errors. The applicability of STARLING is further demonstrated on IMC data of controlled cell lines and human tonsil ROIs.

Overall, in my opinion, this paper addresses an important problem and presents interesting approaches to quantify this issue. However, I had difficulty appreciating the benefits of STARLING over existing methods, and its stability across datasets, settings, cell types and morphologies. I have the following comments:

1. The manuscript can be strengthened by including comparisons of outputs from STARLING and existing/state-of-the-art doublet detection methods. Otherwise, the specific advantages of STARLING over approaches that seek to address similar issues is unclear.

We apologize for the confusion here. We would like to emphasize that the primary purpose of STARLING is not to identify doublets, but to infer cell types in the presence of segmentation errors in the absence of prior knowledge of which cell types we expect in the data. To our knowledge, no method exists to do this, which was our primary motivation to develop STARLING. We apologize if this was not clear in the original manuscript, it was implied in

To-date, no methods have been proposed to learn denoised cellular phenotypes while accounting for cell segmentation errors in highly multiplexed imaging.

in the introduction, though if the reviewer thinks it should be worded differently or moved we would be happy to incorporate any changes.

That said, we are also unaware of any doublet detection/cell QC algorithms developed for highly multiplexed imaging despite their utility in scRNA-seq analysis. This motivated us to develop the “two-step” baseline where first single segmented cells are computationally mixed and a random forest classifier is trained to predict single cell vs doublets, similar to several approaches taken with scRNA-seq. Then, the classifier is applied to all the original single cells, with those likely classified as doublets removed prior to clustering. We note that this has the distinct dis-advantage that only a subset of cells are then clustered, compared to STARLING where all cells receive cluster assignments regardless of cell segmentation error probability. This was not clearly motivated in the original paper, so we have added to the discussion:

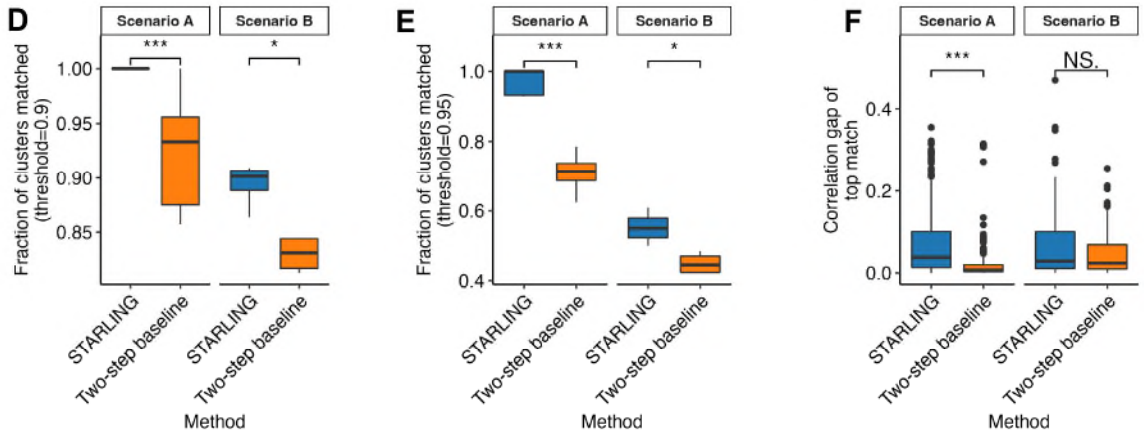
Finally, we note that “two-step” solutions where a set of cells with likely segmentation errors are removed prior to clustering has the disadvantage that not all cells will result

as belonging to a cluster, in contrast to STARLING where every cell has a probability of cluster assignment even in the presence of segmentation errors.

We would like to emphasize that the two-step solution was benchmarked against in the original manuscript, and we now include an extensive simulation benchmarking in response to reviewer 1, where we *in silico* mix cells from suspension mass cytometry and show STARLING is better at recovering the true single-cell populations compared to the two-step doublet finding approach:

Finally, we assessed the ability of STARLING to recapitulate single-cell clusters in the presence of segmentation errors by *in silico* mixing of pure single-cells from suspension mass cytometry. We retrieved single-cell data from a recent paper[3] and simulated segmentation errors by randomly mixing two cells with 80% frequency or three cells with 20% frequency. For this we created two scenarios: scenario *A* where single cells are mixed with equal proportion, and scenario *B* where cells are mixed in an imbalanced setting (70:30 for two cells, 50:25:25 for three). We next clustered the single-cell suspension data, then ran STARLING on the *in silico* mixed data along with the two-step procedure where a random forest classifier is first trained to identify cells with likely segmentation errors, which are then removed before clustering, with the same clustering algorithm used throughout to minimize sources of variability (**Methods**). We subsequently matched the inferred clusters from the mixed data to the original from the single-cell data via correlating the cluster centroids. Overall, STARLING matched significantly more clusters to the single-cell data using different thresholds on the correlation (**Fig. 3D-E**). In addition, the gap between the highest and second highest correlations between the single-cell and mixed data was higher on average for STARLING than the two step approach (**Fig. 3F**), implying clusters could be more uniquely assigned. Together, these results suggest STARLING can infer clusters closer to those obtained from true single-cell data compared to existing approaches.

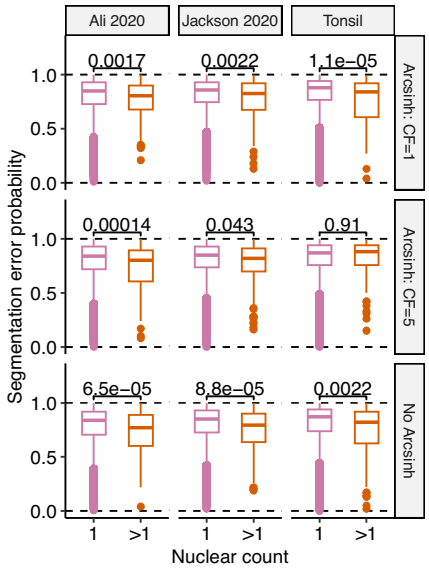
Updated Fig. 3D-F:



We also now include a comparison to the random forest baseline for Fig. 3A to contrast the ability to identify doublets for a simple baseline. We have added the following to the text:

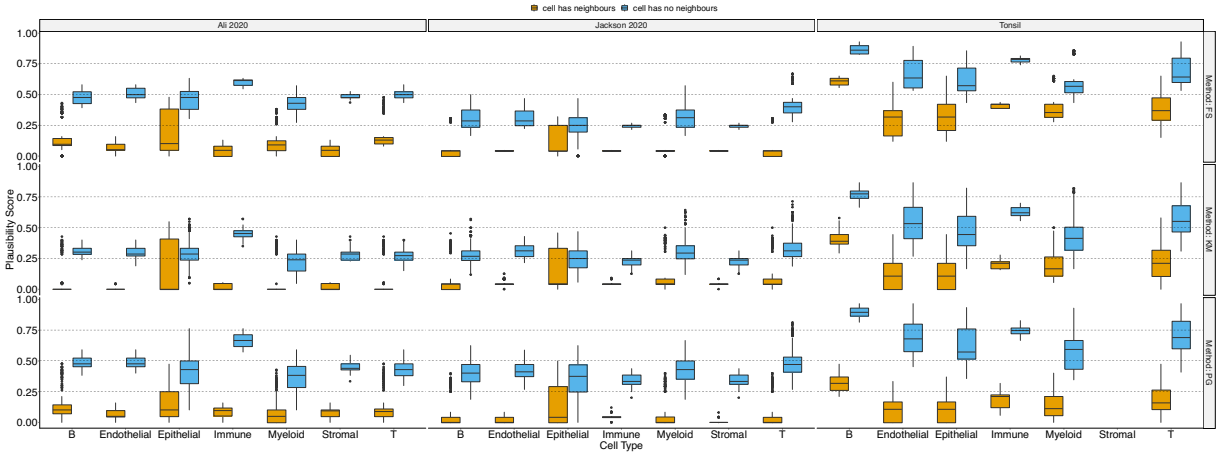
Notably, the increase in segmentation error probability in cells segmented with more than one nucleus was not present using predictions derived from the random forest baseline (**S. Fig**).

With the corresponding supplementary figure:



2. How comprehensive does the list of protein pairs need to be and how does the number of pairs affect the performance of STARLING / plausibility score?

We have recreated Fig. 2C—that demonstrates the plausibility score is overall higher in cells with no neighbours than others—but randomly subsampling to half the number of marker pairs for each computation. It can be seen the overall results are highly consistent:

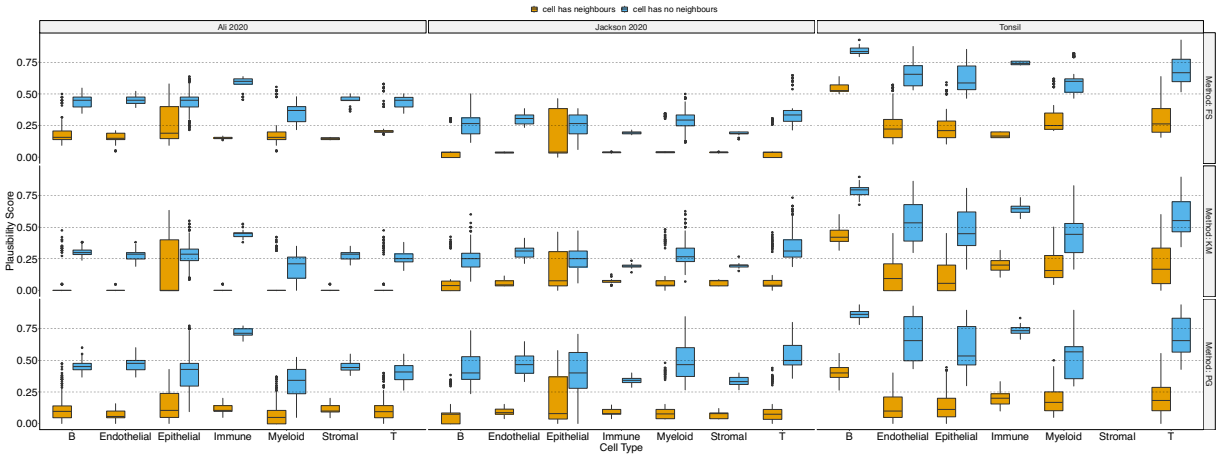


We have added this as a supplementary figure, referenced in the main text via:

a result that is consistent when … the overall pairs of markers used in the computation are randomly subsampled by 50% (**S. Fig.**).

3. Figure 2B – different cell types may be more or less likely found in densely packed regions (i.e., with more neighbours). How does the plausibility score look like broken down across the different cell types?

We thank the reviewer for highlighting this important point. We have assigned each pair of proteins in the plausibility score to a particular cell type (e.g., CD45-CD3 → T cell, see **Supplementary Table 2** for complete list). If we then break Fig. 2B down over cell type, we see the plausibility scores remain well separated between neighbouring cells and non neighbouring cells in all cell types:



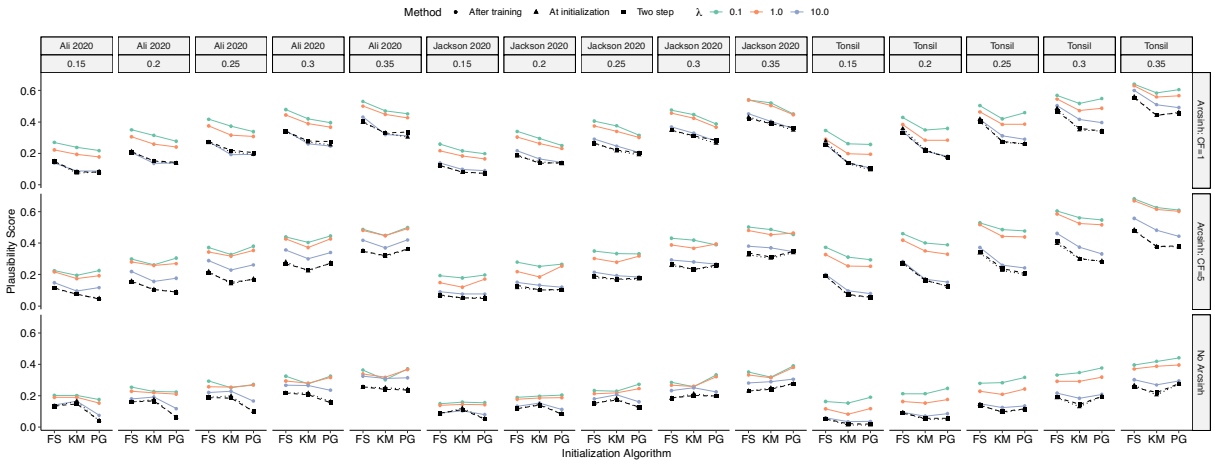
We have highlighted these results in the updated manuscript:

Across all datasets and clustering algorithms, the plausibility score is significantly higher for cells with no neighbours than those with neighbours (**Fig. 2B**), a result that is

consistent when considered individually for marker pairs belonging to individual cell types (**S. Fig. 2** and **S. Table 2**).

4. Line 124 / Fig 2A / Supplementary Table 2 – can the authors provide results for a more comprehensive set of thresholds to better understand the effects. How does the threshold value affect the plausibility score across different datasets? Can the authors also confirm if the threshold applies to all protein pairs (and mutually exclusive/conditional).

We can confirm the threshold applies equally to all pairs and mutually exclusive/conditional. To address the reviewer's point, we have repeated Fig 2C across five different thresholds (0.15, 0.2, 0.25, 0.3, 0.35) and find the results highly consistent for all thresholds (with the average PS increasing as expected as the thresholds are relaxed):



We have added this as a supplementary figure and now reference it in the manuscript:

These results are highly consistent when the threshold used to compute the plausibility score is varied over a range of values (**S. Fig.**).

5. The method seems sensitive to the settings used, with the optimal settings being inconsistent for different datasets/tissue types (Fig 2C). This can impede generalisability of the method, and the user may need to spend a long time searching for the best settings.

We appreciate this point was communicated poorly in the original manuscript. We highlighted that STARLING always outperformed the baseline regardless of hyperparameter, which we felt was important to demonstrate (i.e. the method doesn't only work under some hyperparameter settings). However, the reviewer raises an important point (also raised by reviewer #1), which is that this doesn't actually help the end-user in their analysis. Accordingly, we have distilled the results of the benchmarking on the plausibility score to recommend

STARLING includes a hyperparameter denoted λ that controls the trade-off between optimizing the model likelihood and ability to detect synthetic segmentation errors as well as different data pre-processing strategies. While we show the robustness of the results to choice of λ, we recommend λ=0.1 as it returns the strongest metrics in benchmarking (below) and use this consistently thereafter.

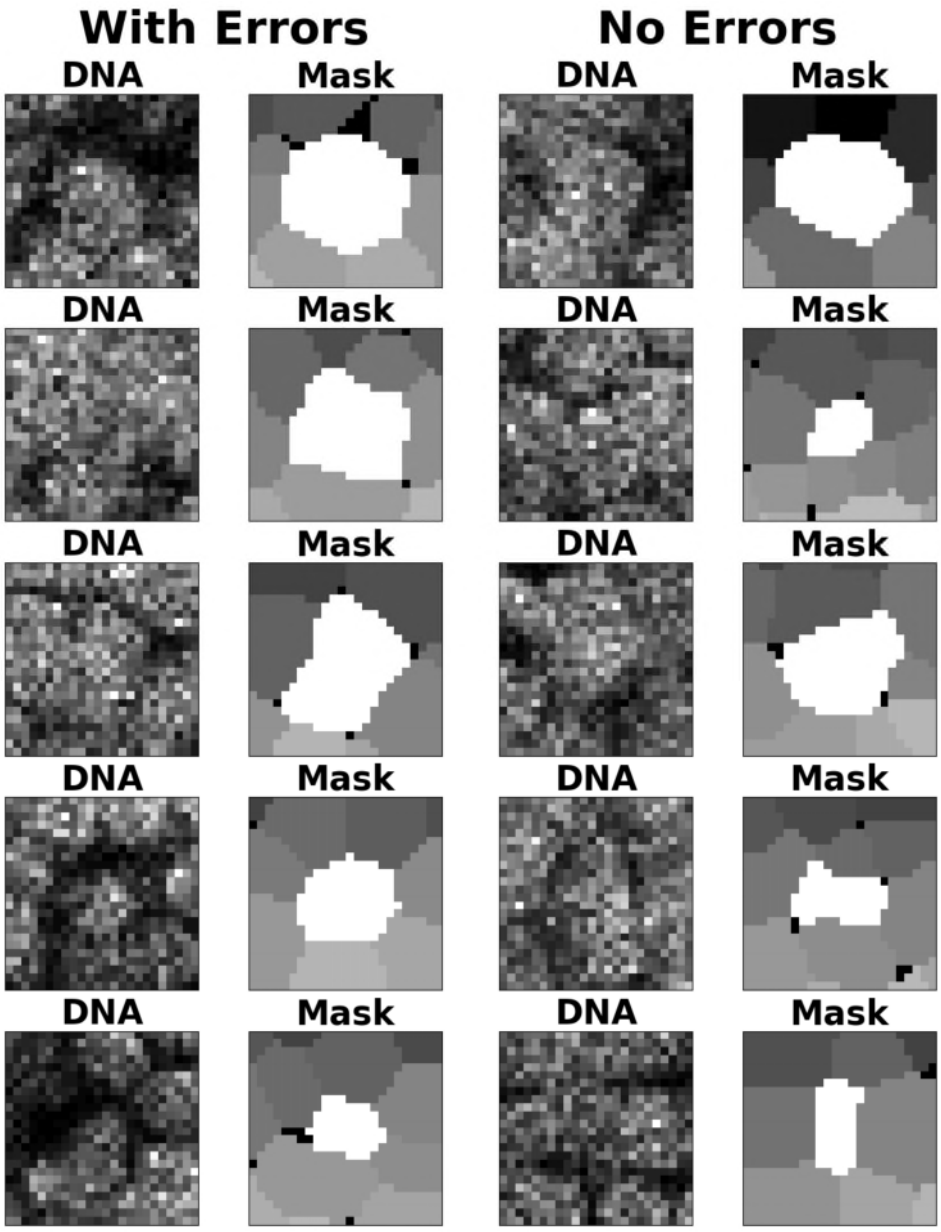
6. Fig 3A – Can the authors provide an explanation as to why probabilities can be below 0 or above 1? Furthermore, as additional validation, it would be useful to include plots of the probabilities for nuclei only. The expectation is that nuclei are relatively pure are without contamination.

We apologize for the previous issue with the figure. The probabilities themselves were never below 0 or above 1, but ggplot's violin plotting function effectively estimated a smoothed density outside that range. We have replaced 3A with a boxplot that now explicitly shows the probabilities are bounded in [0,1].

Regarding the next point, we would ask for clarification on what exactly should be plotted. For the nuclear segmentation, each segmented cell would contain only one nucleus, so we could not recreate 3A with nuclei only, but if we are misunderstanding the request we would be happy to correct this.

7. I suggest the inclusion of a figure showing examples of cells with more than one nucleus and the predicted segmentation error probability associated with the cells, for better understanding.

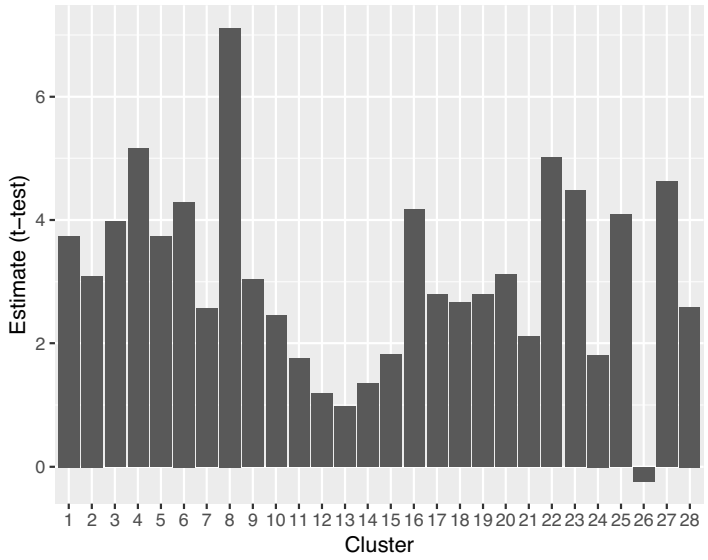
This is an excellent suggestion. The difficulty comes from the (relatively) poor resolution and high noise of IMC, which thwarted our initial efforts to build a supervised model based on human annotation of likely segmentation errors. To demonstrate this, from our cell pellet data we have selected 5 cells predicted to have segmentation errors by STARLING, and 5 without, and plotted both the DNA (nuclear) channel and the MESMER segmentation mask:



There are some cells (e.g. 3rd with error) that conceivably look like they have a segmentation error, but others (e.g. 4th) that don't (that may be due to 3D tissue sectioning artefacts or otherwise). We are happy to include this figure with an explanation if the reviewer would like it in, though it's not as informative as perhaps expected and motivated our focus on cell expression-based approaches.

8. Regarding the metric based the on ratio of convex area to total area, how robust is this metric for cells of different shapes and cell types? I'm wondering because some cell types may have typical shapes (e.g., round / oval / long).

To answer this, we took the final STARLING fit for the clustering of the Tonsil data, and computed the ratio on a per-cluster basis:



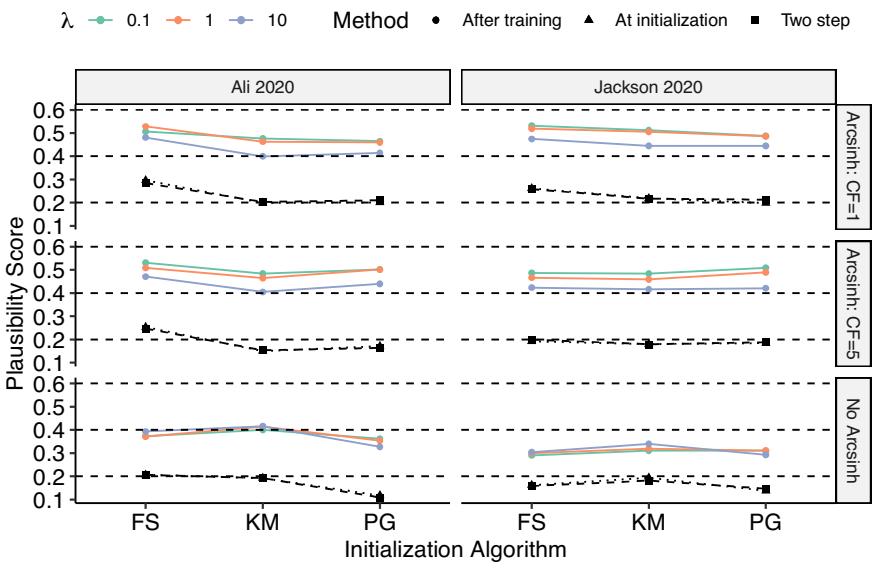
It can be seen the ratio is relatively robust in the sense it is positive for all but one of the cell types, which happens to be a myeloid cell type, aligning exactly with the reviewer's intuition given the expected cellular projections. We have added the following to the main text addressing this:

However, certain cell types are expected to have non-convex morphology, particularly those with large cellular projections. To assess this, we re-computed this ratio individually for each finalized cluster for the Tonsil dataset (see below). This showed a positive effect size for all but one cluster (**S. Fig**) that was interpreted as a myeloid cell type, consistent with the intuition that this metric may not be suitable for every cell type.

9. Additionally, how robust are the proposed metrics to the segmentation method used? Currently only Mesmer is used. What is the sensitivity of the metrics methods to different characteristics of the segmentation method (e.g., the method could produce more complex cell shapes, e.g. with protrusions).

We thank the reviewer for raising this important point. We previously sought to standardize the pipeline to control for as much variation as possible between datasets. We have now repeated the benchmarking on the Jackson 2020 cohort using their original segmentation (a combination of Cellprofiler+Ilastik+Watershed) rather than MESMER. The results show the conclusions regarding STARLING improving the plausibility score are consistent across segmentation methods and we have integrated this into the paper:

These results are highly consistent when…the original cell segmentations presented in the paper were used instead of re-segmentation using Mesmer (**S. Fig.**).



Regarding complex cell shapes and protrusions—it's a significant limitation of the technology (specifically the resolution) that it is currently difficult to segment cells that are not approximately circular. For example, protrusions from macrophages or elongated fibroblasts are almost always missed by all segmentation algorithms. We have added a brief discussion regarding this:

In addition, fundamental resolution limitations of current-generation multiplexed imaging technologies make it difficult to segment cell protrusions, so the effect of these on our benchmarking results remains unknown.

10. Although the a priori known set of protein pairs contains some commonly known pairs, I suggest adding some references to clarify their source.
We have added references covering the main proteins used in calculation of the plausibility score:

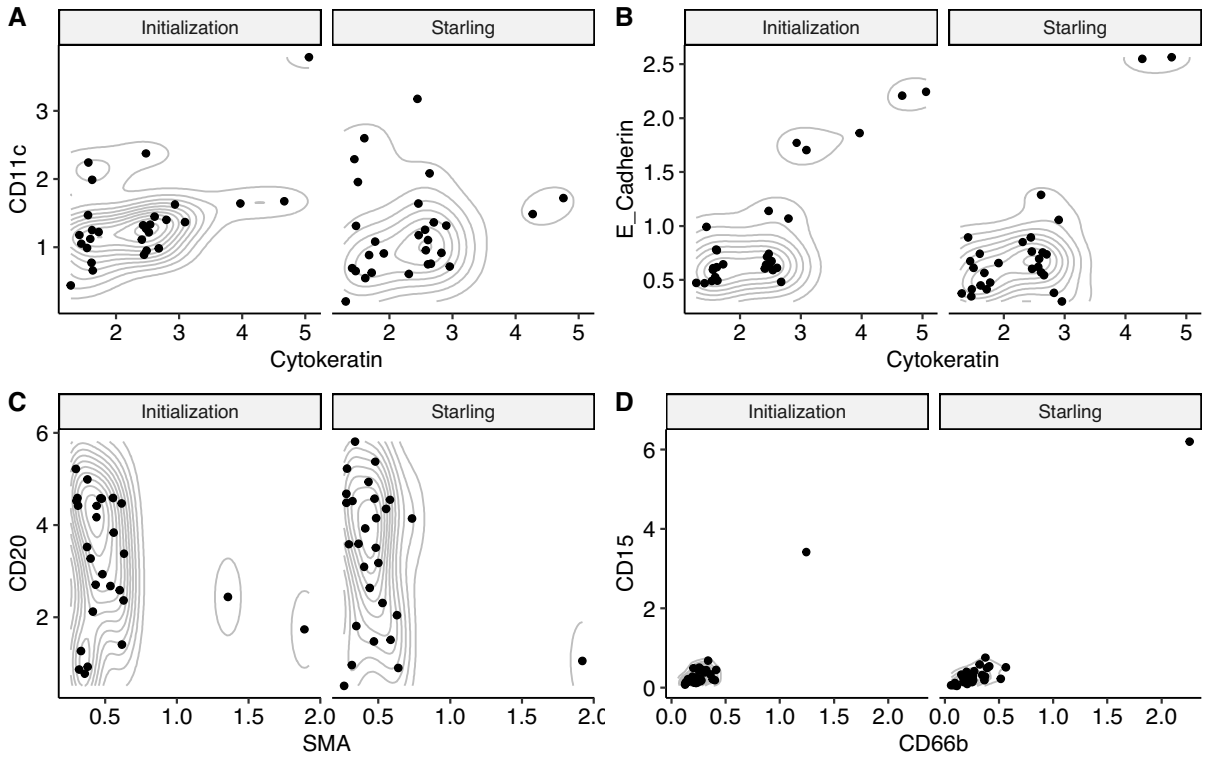
For example, the protein CD3 is highly expressed in T lymphocytes while the protein CD20[3] is highly expressed in B lymphocytes, meaning the pair should not be co-expressed. Similar over-expression patterns come from proteins like CD68 overexpressed in myeloid cell types[4], EpCAM overexpressed in epithelial cells[5], and CD31 overexpressed in endothelial cells[6].

11. I think the utility of STARLING is not fully clear in Figs 5/6. What are the results like when STARLING is not used?

In response to this and comments from the other reviewers we have included specific examples of how the STARLING fit improved cell type interpretation compared to the initialization. Specifically, we have added to the manuscript:

Notably, the interpretation of multiple cell clusters was improved compared to the initialization clustering (**S. Fig.**). For example, at initialization a cluster was present with high co-expression of CD11c (expressed on various leukocytes) and Cytokeratin (expressed on epithelial cells), which was no longer present after applying STARLING (**S. Fig. A**). Similarly, STARLING identified two clear populations of epithelial cells denoted by high co-expression of E-Cadherin and Cytokeratin, while at initialization multiple other clusters were observed with high E-Cadherin expression, but moderate level of Cytokeratin expression similar to that observed on E-Cadherin-low clusters (**S. Fig. B**). In addition, at initialization a cluster expressing moderate CD20 (a B cell marker) and Smooth Muscle Actin (SMA, expressed on certain endothelial cells) was present, while after applying STARLING the only cluster remaining with high SMA expression had low CD20 expression (**S. Fig. C**). Finally, at initialization, the CD66b+CD15+ population demarcating potential granulocytes contained only moderate expression of each, while STARLING adjusted this cluster to contain almost double the expression (**S. Fig. D**). Together, these results demonstrate several ways in which cellular interpretation was improved compared to the default initialization.

with the relevant supplementary figure:



Minor comments:

1. Inconsistent use of style for names, e.g. Mesmer and MESMER

This has been corrected.

2. Fig 2A – hard to tell apart the colours in the pairs. Also consider using proteins C and D for the conditional example.

Thank you for this suggestion. This has been updated in Fig 2A.

3. Usually referred to as the METABRIC dataset rather than Metabrics.

This has been corrected.

4. Repetition between lines 108-112 with 101-102

This section has been substantially re-written. While the repetition is still there, it's now broken into a paragraph that explains how the model works, and one that explains the precise input/output to help users precisely understand both the contributions of the model and what would be required to run it.

5. Figs 4F and FG are unclear and visually busy, obscuring their message. I suggest they be redrawn.

We apologize that this is visually busy and agree with this, but have struggled to think of a way to present the data clearly while retaining all relevant information. If the reviewer has a specific suggestion we would be happy to incorporate it.

Reviewer #4 (Remarks on code availability):

1. I followed the installation steps using virtualenvwrapper, but was unable to proceed past installing virtualenvwrapper. The next step (mkvirtualenv starling) yielded the error: mkvirtualenv: command not found. I was also unable to install using poetry. It stalled at - Installing pytorch-lightning (2.1.0): Pending... - Installing scanpy (1.9.5) for at least 90 minutes. I was using Ubuntu 18.04.1 and Python 3.6.8.
Since our original submission, STARLING has been submitted to pypi so should be available via "pip install biostarling". If the reviewer still has an issue, we would ask them to (anonymously) submit an issue on GitHub.

2. The Colab notebook ran without issue. Please check typos in the notebook, e.g. "Showing initial expression centriods", "Setting trainning log"

Thank you, these have been updated.

3. Some documentation about the contents of the input data file (the .h5ad file) would help with usability.

Thank you for the suggestion, we have added the following to the notebook:

```
The input anndata object should contain a cell-by-protein matrix of
segmented single-cell expression profiles in the `.X` position.
Optionally, cell size information can also be provided as a column of
the `.obs` DataFrame. In this case `model_cell_size` should be set to
TRUE and the column specified in the `cell_size_col_name`argument.
```