nature portfolio

Peer Review File

scConfluence : single-cell diagonal integration with regularized Inverse Optimal Transport on weakly connected features

Reviewer #1 (Remarks to the Author):

See attachment.

Reviewer #1 Attachment on the following page

Title: scConfluence: single-cell diagonal integration with regularized Inverse Optimal Transport on weakly connected features

Authors: J Samaran, G Peyré and L Cantini

Summary Samaran et al. (2024) proposed scConfluence, a computational method for the diagonal integration of single-cell multiomics data, which is specifically designed for the case where no paired multimodal data is available. To link features from different modalities, scConfluence assumes that we can define, based on a priori biological knowledge, weakly connected features between each pair of modalities. That is, it assumes that a (at least partial) mapping between features from each pair of modalities exists, so that we can, for example, convert ATAC peaks to gene activity scores to match the gene expressions the RNA modality. The methodological novelty of scConfluence lies in its utilisation of regularised inverse optimal transport (rIOT) to align cells from different modalities in a common lower-dimensional space. The scConfluence framework is very flexible and can be applied to many different data types and use cases.

Major comments

1. The smFish case study appeared to be less convincing compared to the other case studies, mainly due to the close behaviours of all methods in the gene imputation task. Indeed, it's not clear from Fig 4d that scConfluence enjoys any significant advantage compared to competing methods. The authors claimed (lines 324–325) that according to the mSCC criterion scConfluence outperforms all competing methods. While scConfluence had a median mSCC value slightly higher than the second highest median mSCC value, their inter-quartile ranges overlapped too much to make scConfluence clearly outperforming. Fig 4e didn't fully help as the spatial patterns of some genes were conspicuously different from the ground truth. For example, the true Kcnip2 expression had a clearer layered structure, whereas was missing in the imputed version.

2. In the 3-omics (scRNA+scATAC+CyTOF) study, the clustering analyses (Fig 5g–i) was not maximally convincing. First, the authors highlighted (lines 375–376) the fact that the B cells in the RNA modality was split into three biologically meaningful subclusters. However, this was hardly surprising as 'B cell' is a very broad cell type

and it is usually sufficient to use the RNA modality alone to identify some B cell subtypes. That is to say, by subclustering the gene expression profiles of the B cells in the RNA modality, we should be able to identify some B cell subtypes, without the help of additional modalities. Furthermore, it seemed that the authors didn't really discuss cluster 8, which, on Fig 5g, appeared to correspond to memory CD8 T cells and natural killers (NKs). This was a bit concerning since there are some significantly transcriptional differences between these two cell types and usually they shouldn't be very difficult to distinguish. Was this result due to some confusion of these two cell types by scConfluence, or was it due to some mislabelling in the original dataset? More clarification is necessary to make this case study more solid.

3. The last case study on Patch-seq data was interesting and novel, but explanations of some key details, concerning the format of the morphology modality and the definition of weakly connected features, seemed to be lacking. It appeared that by morphology the authors referred to some imaging data. Then how to link gene expressions and image pixels would be the key. However the authors' explanations, centred around lines 448–450, was somewhat confusing. This hindered the reader's appreciation of the novelty of this case study.

Minor comments

1. Line 156 featured a phrase 'Unbalanced Optimal Transport', which seemed to be undefined. Does 'unbalanced' refer to the use of unbalanced Sinkhorn divergence? Lines 157-158 emphasised the advantage of using unbalanced OT. It seemed that the 'unbalanced' nature of the OT was really essential for not forcing all cells to align. More explanations?

2. In all figures, the colour scheme used to distinguish different methods was sometimes unhelpful, especially in the line charts (eg Fig 2b). This is because the colour for scConfluence, which needed maximal emphasis, was sometimes hard to distinguish from some of the competing methods. It may help to change the scConfluence colour to something that stands out more easily, eg red.

3. It would be more helpful for the potential users of scConfluence if the authors could comment on the selection of the lambda values in the loss function. This is

because the authors were using different lambda values in different case studies, so it appeared that parameter tuning really mattered. In addition, a key factor for the user to choose among different methods is computation time. Could the authors comment on scConfluence's computation time compared to competing methods?

References

Samaran, J., Peyré, G., & Cantini, L. (2024, February). Scconfluence : Single-cell diagonal integration wit

Reviewer #2 (Remarks to the Author):

The authors describe a framework to construct an integrated latent space across unpaired multi-modal single-cell datasets, informed by weak feature correspondences across profiling modalities. Auto-encoders are placed over each modality into a shared latent space, where integration is guided with an inverse OT-based approach. An unbalanced OT coupling is computed on the (weakly) corresponding features between each modality pairwise, and the latent representations are informed according to this plan. This is further regularized with an OT coupling between the latents themselves.
The paper is nicely written and thorough, and the approach seems robust and consistent. Futher more, the code is well documented, integrated in the scverse framework, and reproducibility notebooks are available online.
Major

1. How are the train/validation/test splits performed? Ideally training is done on train split, stopping is done on validation split, and figures, metrics, comparisons are made on the test split. Testing on the same data used for stopping is an info leak, as you can overfit to the validation set. This is particularly relevant when comparing to other methods. If three data splits are made and results reported in this way, it should be explicitly stated in the text, otherwise the comparisons to baseline methods are not fair.

2. A trivial solution to the integration problem is to perform cell typing on each modality and integrate by random assignment within cell types across modalities. While these methods typically do not require the cell type information, it is a standard part of single-cell analysis pipelines. A major advantage of these computational methods, especially those that construct latent spaces, is their ability to represent cells beyond discrete cell types.
What cell-type substructure, if any, is identified in the integrated latent space?
- Compare ranks of shared features between coupled cells
- Is the integration cycle-consistent? I.e. map rna -> atac -> rna, is the input and output similar vs a random in-cell type mapping?
It would be convincing enough to show this in the supplement fora few settings in the first experiment

2a. A version of FOSCTTM that only considers cells of the same cell type would be informative here. If we can assume that cell types are well separated, then the FOSCTTM is artificially inflated by the number of cell types. If you want to keep this between [0, 1] then you can e.g. normalize by the cells within the cell type plus the number of cells outside of the cell type that are located closer to the target cell.

3. It is not clear why a subset of metrics are reported for each experiment. For instance, FOSCTTM is not reported for the first experiment even though a ground truth pairing exists. Either provide all introduced metrics for each experiment in the supplement or explain why a certain metric is not applicable for the experiment

4. As a major contribution of the method seems its robustness to imbalanced cell types, showing the distribution of cell types across each modality for each experiment would be beneficial.

5a. This is a flexible and modular framework, consisting of many hyperparameters. While most practioneers will be familiar with e.g. choices in neural network architectures & optimizers, it might not be obvious how to adapt parameters for the loss and OT to new settings. Future users would benefit from more intuition on how (non-default OT) parameters and the lambdas are selected. Knowing what a training curve or initialization should look could also suffice.
5b. It would also be beneficial to understand how robust performance is to the partial OT m and if this parameter needs to be tuned if imbalance becomes more extreme.

Minor
Figure 1: Error bars are not visible, please plot your approach over baselines and decrease alpha. Consider changing the size, alpha and color of median indicators. Plotting as a grouped boxplot instead of a lineplot might also be easier to interpret.

Figure 4e: scatter plot/2D kde of ground truth vs imputed values would be useful here
669: change λ_p to $\lambda_{AE^{(p)}}$ to be consistent with figures, rest of paper
242: as per the methods section, FOSCTTM was modified from its original formulation and this should be noted in the main text

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

Reviewer #1

Summary

Samaran et al. (2024) proposed scConfluence, a computational method for the diagonal integration of single-cell multiomics data, which is specifically designed for the case where no paired multimodal data is available. To link features from different modalities, scConfluence assumes that we can define, based on a priori biological knowledge, weakly connected features between each pair of modalities. That is, it assumes that a (at least partial) mapping between features from each pair of modalities exists, so that we can, for example, convert ATAC peaks to gene activity scores to match the gene expressions of the RNA modality. The methodological novelty of scConfluence lies in its utilization of regularised inverse optimal transport (rIOT) to align cells from different modalities in a common lower-dimensional space. The scConfluence framework is very flexible and can be applied to many different data types and use cases.

Major comments

1. The smFish case study appeared to be less convincing compared to the other case studies, mainly due to the close behaviours of all methods in the gene imputation task. Indeed, it's not clear from Fig 4d that scConfluence enjoys any significant advantage compared to competing methods. The authors claimed (lines 324–325) that according to the mSCC criterion, scConfluence outperforms all competing methods. While scConfluence had a median mSCC value slightly higher than the second highest median mSCC value, their inter-quartile ranges overlapped too much to make scConfluence clearly outperform. Fig 4e didn't fully help as the spatial patterns of some genes were conspicuously different from the ground truth. For example, the true Kcnip2 expression had a clearer layered structure, which was missing in the imputed version.

Reply: We agree with the Reviewer that the results of scConfluence in gene imputation are not strikingly better than those of other state-of-the-art methods. At the same time, all methods seem to not have great performances in imputation, with low correlation with the ground truth (mSCC and aSCC close to 0.2), despite some of them being designed to achieve this task specifically.

We toned down our claim in this section highlighting that the performances of scConfluence are comparable with those of the state-of-the-art. Finally, we added some nuance to our description of the qualitative plots. Indeed, while the coarse pattern is identified in our imputations, finer patterns are not perfectly reconstructed as seen in the imputation of Kcnip2. Indeed we correctly find that Kcnip2 is expressed in pyramidal neurons (both in the hippocampus and upper layers) and in inhibitory neurons from the caudoputamen. However, we do not capture the layered structure in pyramidal neurons as present in the ground truth figure where Kcnip2 is specifically expressed in layers ⅔, 4, and 6. To have a comparison with a state-of-the-art method, we created Supplementary Figure 9 where we can see that GimVI (which was designed specifically and only for imputation) achieves lower correlations than scConfluence on the same three genes. Focusing for example Kcnip2 it fails to impute its expression in the caudoputamen.

We now updated section 4 of Results, modified Figure 4 adding the correlation values below the spatial visualization of the imputations, and added Supp Figure 9 to showcase the performances of GimVI on the same genes.

2. In the 3-omics (scRNA+scATAC+CyTOF) study, the clustering analyses (Fig 5g–i) was not maximally convincing. First, the authors highlighted (lines 375–376) the fact that the B cells in the RNA modality were split into three biologically meaningful subclusters. However, this was hardly surprising as 'B cell' is a very broad cell type and it is usually sufficient to use the RNA modality alone to identify some B cell subtypes. That is to say, by subclustering the gene expression profiles of the B cells in the RNA modality, we should be able to identify some B cell subtypes, without the help of additional modalities.

Reply: While we agree that "B cells" is a very broad cell type, the authors of the original study that produced and analyzed this data did not obtain any finer annotation (AJ Wilk, Nature Medicine, 2020).

To answer your question, we however clustered the B cells only using the scRNA and identified 3 subclusters Supp Figure 10A. As shown in Supp Figure 10 B these 3 subclusters largely overlap in terms of cells with the clusters we had identified in the integrative analysis (clusters 0, 1, and 2 in Figure 5) as naive B, memory B, and plasma cells. However, with the scRNA analysis alone it would not have been possible to annotate cluster 2 as plasma cells. Indeed, while two of the scRNA clusters expressed respectively markers of naive and memory B cells, the third scRNA cluster did not have a clear association with the markers of plasma cells (see Supp Figure 10 A, C). For example, the marker gene XBP1 was expressed sparsely over several clusters.

On the other hand, our multimodal integration enabled us to identify the 3 subpopulations, including plasma cells. Indeed, despite the absence of a clear overexpression of plasma markers, we were able, through statistical testing, to verify that there's a significant overlap between the features differentially expressed in the scRNA and scATAC cells in this subcluster. The finer level of annotations available for ATAC and CyTOF subclusters therefore allowed us to annotate this subpopulation (see Figure 5). The integration of the three omics was thus crucial in this case to identify meaningful subpopulations in the 'B cells' cell type.

Furthermore, it seemed that the authors did not discuss cluster 8, which, in Fig. 5g, appeared to correspond to memory CD8 T cells and natural killers (NKs). This was a bit concerning since there are some significant transcriptional differences between these two cell types and usually they shouldn't be very difficult to distinguish. Was this result due to some confusion of these two cell types by scConfluence, or was it due to some mislabelling in the original dataset? More clarification is necessary to make this case study more solid.

Reply: We thank the Reviewer for this interesting observation. To address this point, we subclustered the latent embeddings in cluster 8, composed of cells labeled in the original publication (AJ Wilk, Nature Medicine, 2020) as either memory CD8 T cells or NK cells. This analysis showed that those "NK cells" in cluster 8 formed a distinct subpopulation of cells expressing both markers of NK and CD8 T cells (CD3E and NCAM1). Therefore, this

subpopulation was incorrectly labeled as NK cells and represented instead NKT cells (Värynen J. P., Cancer Immunol. Res., 2022), which are a heterogeneous group of T cells that share properties of both T cells and Natural Killer cells. We now updated Figure 5 as well as the fifth section of the results to report this subpopulation and adjusted accordingly the cluster labels (cluster 8 is the NKT cells, cluster 9 is the former cluster 8 and all clusters above are shifted by one). In addition, we now provide a new Supplementary Figure (Supp. Figure 11) in which we can clearly see that while NK cells (cluster 7) are NCAM1+CD3− and CD8 T cells are NCAM1-CD3+ (new cluster 9), NKT cells (new cluster 8) are NCAM1+CD3+.

3. The last case study on Patch-seq data was interesting and novel, but explanations of some key details, concerning the format of the morphology modality and the definition of weakly connected features, seemed to be lacking. It appeared that by morphology the authors referred to some imaging data. Then how to link gene expressions and image pixels would be the key. However, the authors' explanations, centered around lines 448–450, were somewhat confusing. This hindered the reader's appreciation of the novelty of this case study.

Reply: we are sorry to the Reviewer if our explanation of the data used for the Patch-seq experiment was not sufficiently clear. The Patch-seq dataset we used was composed of two subsets of cells: in subset A only scRNA-seq profiles were available. In subset B, both images of neurons and their gene expression profiles were measured in the same cells. In the scConfluence integration, we used as X matrices the scRNA profiles for subset A and the images for subset B. We agree with the reviewer that it would be impossible to define connected features if we only relied on images in subset B. However, since scRNA-seq profiles were also available for cells of subset B, we used these data to create the Y matrices (connected features). Indeed, scRNA-seq from subset A and scRNA-seq from subset B contain the same features and therefore provide a very strong bridge for scConfluence integration.

We now improved the explanation in section 6 of Results.

Minor comments

1. Line 156 featured a phrase 'Unbalanced Optimal Transport', which seemed to be undefined. Does 'unbalanced' refer to the use of unbalanced Sinkhorn divergence?

Reply: Yes, but not only as the computation of the L_IOT term also involves Unbalanced Optimal Transport since it is based on a transport plan which is unbalanced as we now indicate in the main text (section 1 of Results). In addition, to clarify this part we also send to the Methods section for the definition of Unbalanced Optimal Transport and its application in the computation of those losses.

Lines 157-158 emphasized the advantage of using unbalanced OT. It seemed that the 'unbalanced' nature of the OT was really essential for not forcing all cells to align. More explanations?

Reply: In a balanced OT problem, there is a constraint on the marginals of the transport plan such that all the cells from the two distributions must be transported onto a suitable match in the other distribution. Unbalanced OT relaxes this constraint to allow the unbalanced plan to ignore cells for which there is no "suitable enough" match in the other distribution (as explained

in the Methods section). Unbalanced OT thus helps, for example, in situations where some populations of cells are present in one modality but not the other. We now clarified this point in section 1 of the results.

2. In all figures, the color scheme used to distinguish different methods was sometimes unhelpful, especially in the line charts (eg Fig 2b). This is because the colour for scConfluence, which needed maximal emphasis, was sometimes hard to distinguish from some of the competing methods. It may help to change the scConfluence colour to something that stands out more easily, eg red.

Reply: We improved the visibility of scConfluence in the benchmark figures by plotting its performances in fuchsia on top of other methods and by choosing the colors of other methods to increase their contrast with respect to scConfluence.

3. It would be more helpful for the potential users of scConfluence if the authors could comment on the selection of the lambda values in the loss function. This is because the authors were using different lambda values in different case studies, so it appeared that parameter tuning really mattered.

Reply: We added a paragraph on this in the Methods section of the article, under the title “Training hyperparameters” and included as well a link to the documentation website where these questions are discussed more extensively.

In addition, a key factor for the user to choose among different methods is computation time. Could the authors comment on scConfluence’s computation time compared to competing methods?

Reply: We now provide a new Supplementary Table (Supp. Table 6) reporting a comparison of the computation times of the different methods on the PBMC 10X dataset (where approximately 10k cells were profiled for both modalities). Overall, scConfluence running time is comparable with the state-of-the-art. It is indeed slower than the three CPU methods (Seurat, liger, and MultiMAP). Nonetheless, these methods operate on the full dataset in one go and therefore can’t be applied to very large datasets. On the other hand, scConfluence is faster than the other two neural network-based methods, all three being much more scalable due to their mini-batch approach.

We added this comment to the Methods section “Computational runtime”.

Reviewer #2

The authors describe a framework to construct an integrated latent space across unpaired multi-modal single-cell datasets, informed by weak feature correspondences across profiling modalities. Auto-encoders are placed over each modality into a shared latent space, where integration is guided with an inverse OT-based approach. An unbalanced OT coupling is computed on the (weakly) corresponding features between each modality pairwise, and the

latent representations are informed according to this plan. This is further regularized with an OT coupling between the latent distributions themselves.
The paper is nicely written and thorough, and the approach seems robust and consistent. Furthermore, the code is well documented, and integrated into the scverse framework, and reproducibility notebooks are available online.

Major

1. How are the train/validation/test splits performed? Ideally training is done on train split, stopping is done on validation split, and figures, metrics, comparisons are made on the test split. Testing on the same data used for stopping is an info leak, as you can overfit to the validation set. This is particularly relevant when comparing to other methods. If three data splits are made and results reported in this way, it should be explicitly stated in the text, otherwise the comparisons to baseline methods are not fair.

Reply: We apologize to the Reviewer if our explanation was not completely clear. As mentioned in the Methods' section "Training hyper parameters" (first 4 lines), we split each dataset randomly (80/20 %) into train and validation sets. We only use the validation set to stop the training of the model with Pytorch Lightning's early stopping callback. We then use all samples (both training and validation) after training to compute cell embeddings and evaluation metrics. This is the standard practice for representation learning on single-cell data and it has been adopted by the majority of the community (Cao Z.-J., Nat. Biotechnol, 2021; Cao K., Nat. Commun., 2022; Lopez R., Nat. Methods, 2018). While for classical machine learning tasks like regression and classification, it is necessary to evaluate on a held-out test dataset, this is not the case for unsupervised dimension reduction methods. Indeed, in our task, the goal is to encode the whole given dataset on which the model was trained. There is no notion of unseen samples and generalizability for this problem since the models we train are not meant to be then used on different datasets at inference time. There is no information leakage either since the ground truth information (i.e. cell type labels and pairing information) used to evaluate the methods are not used during training.

We now better detail these aspects in the Methods' section "Training hyper parameters".

2. A trivial solution to the integration problem is to perform cell typing on each modality and integrate by random assignment within cell types across modalities. While these methods typically do not require the cell type information, it is a standard part of single-cell analysis pipelines. A major advantage of these computational methods, especially those that construct latent spaces, is their ability to represent cells beyond discrete cell types.
What cell-type substructure, if any, is identified in the integrated latent space?

Reply: We agree with the Reviewer that a core advantage of embedding methods is their ability to represent cells beyond discrete cell types. Our work highlights the ability of scConfluence to go beyond discrete cell types and identify additional heterogeneity. For example, in Results section 6 we report the variation of the height of apical dendrites in intra telencephalic neurons. Such continuous morphological heterogeneity inside well-defined transcriptomic cell types indicates that the integrated latent space contains more information than discrete cell types. In addition, we investigated this aspect in our benchmark (Results section 2-3) by reporting the FOSCTTM score which doesn't rely on discrete annotations.

However, this same Reviewer proposed in one of his/her points below to add a “cell type FOSCTTM” score. This same score can help to further investigate the point here raised by the Reviewer. Indeed, the trivial solution mentioned above would result in a 0.5 cell type FOSCTTM. We can thus assert whether the benchmarked methods learn cell-type substructures by comparing their result with this baseline.

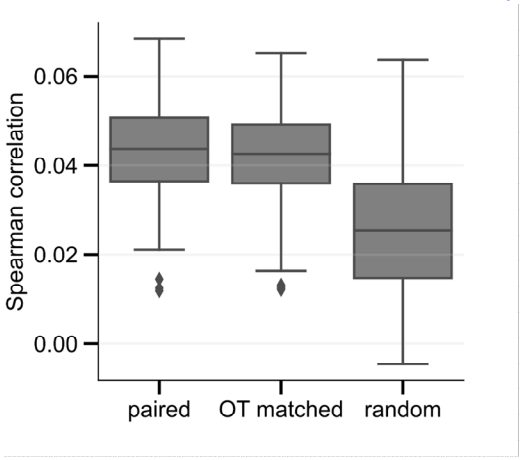
For all the benchmark datasets (Results section 3) we observed that scConfluence’s score was significantly below the 0.5 baseline (Supp Figure 6) indicating that the scConfluence embeddings capture cell-type substructures. We did not compute this same score on cell lines as no heterogeneity is expected among cells from the same cell line.

These results are now presented in section 3 of Results.

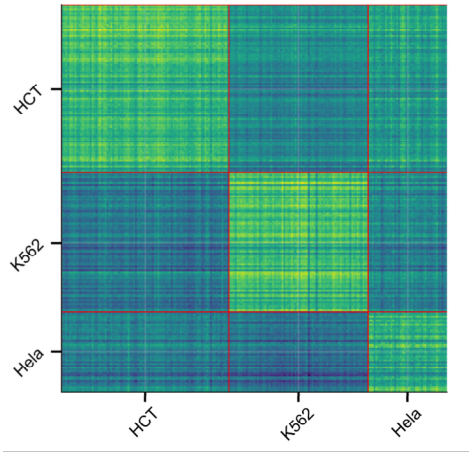
- Compare ranks of shared features between coupled cells

Reply: We did not understand if the Reviewer meant as “coupled cells” the ground truth pairs available from scCAT-seq (1) or the cells that are coupled by scConfluence (2). We thus performed both experiments on cell lines using the Spearman correlation between gene expression profiles and atac-derived gene activity profiles to compare the ranks of shared features across modalities.

In the Figure below, we plot the distribution of correlations for (1) in the first boxplot and the distribution of the correlations for (2) in the second boxplot. As a reference, we added a third boxplot containing correlations between random pairs of cells. The distribution (1) is higher than the correlations between random pairs of cells and very similar to the distribution (2).



This indicates that scConfluence aligns cells that are similar based on the shared features. In addition, we plotted the correlations between all pairs of cells and found that while the values in (1), corresponding to the diagonal of the heatmap, are not significantly higher than pairs taken from the same cell type, they are higher than pairs taken from different cell types.



We hope that this addresses the point of the Reviewer, if this is not the case, we would be happy to provide additional information. At the same time, we failed to see the critical point of the paper that made the Reviewer raise this question. We would be happy to include this analysis in the paper once this point would be more clear.

- Is the integration cycle-consistent? I.e. map rna -> atac -> rna, is the input and output similar vs a random in-cell type mapping?
It would be convincing enough to show this in the supplement for a few settings in the first experiment

Reply: We thank the Reviewer for this intriguing question. While AE-based methods allow in principle users to map cells back and forth across modalities, no state-of-the-art method for single-cell integration has investigated cycle consistency before. Indeed, the focus in single-cell diagonal integration is on the quality of the latent embeddings as well as the ability to do a one-way translation from one modality to the other. Cycle consistency is thus not a necessary condition for a diagonal integration method to work well.

To investigate the cycle-consistency of the integration, we compared the embeddings of the RNA cells to the embeddings obtained after decoding them through the ATAC decoder and mapping them back to the embedding space through the ATAC encoder. The results are presented in the Supp. Fig. 14 where the Reviewer can see that while “cycle reconstructed embeddings” are mapped close to cells from the same cell line, there is a discrepancy with respect to the original embeddings. This suggests that the results are worse than a random in-cell type mapping.

We now discuss the cycle-consistency of scConfluence in the Methods section “Outputs of scConfluence”.

2a. A version of FOSCTTM that only considers cells of the same cell type would be informative here. If we can assume that cell types are well separated, then the FOSCTTM is artificially inflated by the number of cell types. If you want to keep this between [0, 1] then you can e.g. normalize by the cells within the cell type plus the number of cells outside of the cell type that are located closer to the target cell.

Reply: This is a very relevant question. As discussed above, we now added a new score called “cell type FOSCTTTM” to evaluate this aspect. The results obtained with this score are presented in Supp Figure 6, in addition, the score is described in the Methods section “Evaluation metrics”.

3. It is not clear why a subset of metrics are reported for each experiment. For instance, FOSCTTM is not reported for the first experiment even though a ground truth pairing exists. Either provide all introduced metrics for each experiment in the supplement or explain why a certain metric is not applicable for the experiment

Reply: That’s true that cell line data are paired, but the different scenarios that we investigated in this experiment created partially unpaired situations (situations where a cell’s pair in the other modality was not present or where no cells from the same cell type were present in the other modality). For this reason, we originally didn’t report all metrics for the cell lines experiment. However, we now used exactly the same scores for both benchmarks. To deal with the partial unpairedness we only evaluated FOSCTTM score for paired cells and, for the transfer accuracy, we only classified cells whose cell type was present in the other modality. In addition, the graph connectivity scores for the benchmark experiment were moved from the Supplementary Materials to Figure 3. Furthermore, to be consistent across experiments, we also used the same scores in Figure 4 (apart from FOSCTTM since the mouse cortex data are fully unpaired).

We now updated Figures 2 and 3 to report all scores and the Results text accordingly.

4. As a major contribution of the method seems its robustness to imbalanced cell types, showing the distribution of cell types across each modality for each experiment would be beneficial.

Reply: We now added Supplementary Figures 15 to 18, reporting the proportion of cells per cell type in the unpaired datasets. We refer to these Figures in the Methods section “Data preprocessing”.

5a. This is a flexible and modular framework, consisting of many hyperparameters. While most practitioners will be familiar with e.g. choices in neural network architectures & optimizers, it might not be obvious how to adapt parameters for the loss and OT to new settings. Future users would benefit from more intuition on how (non-default OT) parameters and the lambdas are selected. Knowing what a training curve or initialization should look could also suffice.

Reply: We added a paragraph on this in the Methods section of the article, under the title “Training hyper parameters” and included as well a link to the documentation website where these questions are discussed more extensively.

5b. It would also be beneficial to understand how robust performance is to the partial OT m and if this parameter needs to be tuned if imbalance becomes more extreme.

Reply: We compared the results of our method for m = [0.2, 0.35, 0.5, 0.85] to show the impact of m on the performance as the cell line scenarios become more unbalanced, see Supp Figure 13. As expected, we observe that higher values of m give better results in more

balanced scenarios and lower values give better results in unbalanced scenarios. This means that when users know about the degree of unbalancedness of the dataset, they can adjust m to get better results. However, since this prior knowledge is generally not available, we see that setting m to 0.5 produces robust performances across all metrics.

We now refer to the new Supp Figure 13 and we provide the additional information on m in the Methods section “Optimal Transport solvers”.

Minor

Figure 1: Error bars are not visible, please plot your approach over baselines and decrease alpha. Consider changing the size, alpha and color of median indicators. Plotting as a grouped boxplot instead of a lineplot might also be easier to interpret.

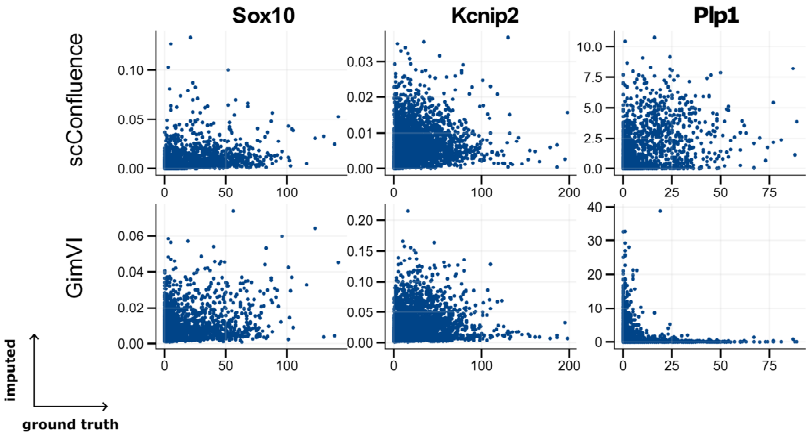
Reply: We improved the visibility of scConfluence in the benchmark figures by plotting its performances in fuchsia on top of other methods and by choosing the colors of other methods to increase their contrast with respect to scConfluence. We also increased the width of the error bars to make them more visible. Figures 2-4 are now updated accordingly.

Figure 4e: scatter plot/2D kde of ground truth vs imputed values would be useful here

Reply:

We found 2D spatial plots to be more informative to assess visually the quality of the imputations as they allow readers to compare the spatial patterns of expression between imputed and ground-truth counts. Since the exact values of the imputations matter much less than the rank of cells, we chose to report the spearman correlation for the quantitative evaluation and we now modified Figure 4 to report these correlation scores below the plotted spatial imputations. These correlations give a complete quantitative assessment of the relationship between imputed and ground-truth values.

However, to address the comment of the Reviewer, we provide below the scatter plots of imputed vs ground-truth counts. As you can see, they don’t carry additional information which is not already provided in Figure 4.



669: change λ_p to $\lambda_{AE^{(p)}}$ to be consistent with figures, rest of paper

Reply: We uniformized the notation by using λ_p everywhere (figures and text) to designate the weights of the reconstruction losses.

242: as per the methods section, FOSCTTM was modified from its original formulation and this should be noted in the main text

Reply: We now specify in Results section 2 that the FOSCTTM was modified from its original formulation and we send the reader to the methods where this is further detailed.

Reviewer #1 (Remarks to the Author):

I appreciate the authors' efforts on addressing my comments. All my major concerns have been satisfactorily addressed. In particular, in response to my major comment 2, the authors reevaluated the cell type labelling of data from Wilk et al. (2020) and identified some possible mislabelling issue, showcasing scConfluence's ability for more accurate cell typing. Their demonstration that multi-omics annotation is more robust than a single modality (scRNA-seq) in differentiating B cell subpopulations is also convincing.

Minor comments have also been appropriately addressed. Overall the current version of the manuscript is more readable.

Reviewer #1 (Remarks on code availability):

scConfluence comes with a detailed documentation. I tried their notebook on imputing smFISH mouse cortex data (https://scconfluence.readthedocs.io/en/latest/tutorials/RNA_FISH_tutorial.html). Below are my observations.

When I tried to import scconcluence I encountered an version clash with NumPy 2.X, I had to downgrade Numpy to 1.26.4 so that I could correctly import scconfluence

When running code block 11, I encountered error asking me to pip install --user scikit-misc

I ran the training step on a Mac without gpu, so it took longer the finish. It would be helpful if there is a progress bar (as in scVI model training).

Apart from these minor problems I could run all the code and got the same results as shown in the notebook.

Reviewer #2 (Remarks to the Author):

The authors have addressed my comments and concerns.

Reviewer #1 (Remarks on code availability):

scConfluence comes with a detailed documentation. I tried their notebook on imputing smFISH mouse cortex data (https://scconfluence.readthedocs.io/en/latest/tutorials/RNA_FISH_tutorial.html). Below are my observations.

When I tried to import scconcluence I encountered an version clash with NumPy 2.X, I had to downgrade Numpy to 1.26.4 so that I could correctly import scconfluence. When running code block 11, I encountered error asking me to pip install --user scikit-misc

Answer: To fix those two points, we added Numpy <= 1.26.4 and scikit-misc to the requirements of scConfluence and updated both the pip package and the github repository.

I ran the training step on a Mac without gpu, so it took longer the finish. It would be helpful if there is a progress bar (as in scVI model training).Apart from these minor problems I could run all the code and got the same results as shown in the notebook.

Answer: As scConfluence's number of training iterations isn't exactly determined by the user which only gives a maximum number of epochs which is never attained because of early stopping, it is impossible to know a priori when the training will end which is why we decided not to put a progress bar as the indicated predicted training length would be largely overestimated and could mislead the users.