Cell Systems

Transparent Peer Review Record

Cell-Type Modeling in Mouse Brain Spatial Transcriptomics Data Elucidates Spatial Variation of Cellular Colocalization and Intercellular Communication

Francisco Jose Grisanti Canozo, Zhen Zuo, James F. Martin, Md. Abul Hassan Samee

Summary

Initial Submission: Received March 31, 2021
Preprint: https://doi.org/10.1101/2020.09.09.290064

Scientific editor: Ernesto Andrianantoandro, Ph.D.

First round of review: Number of reviewers: Two
Two confidential, Zero signed
Revision invited June 09, 2021
Major changes anticipated
Revision received August 06, 2021

Second round of review: Number of reviewers: One
One original, Zero new
One confidential, Zero signed
Accepted September 10, 2021

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Samee,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript.

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns regarding:

1. Proper contextualization with respect to previously published work.
2. Benchmarking to provide fair comparison to competing techniques.

Additionally, to move forward at Cell Systems, there needs to be a clearer demonstration of utility of the approach (generalizability to other datasets, biological insights not possible with competing techniques). I'd also like to be explicitly clear about an almost philosophical stance that we take at Cell Systems…

We believe that understanding how approaches fail is fundamentally interesting: it provides critical insight into understanding how they work. We also believe that all approaches do fail and that it's unreasonable, even misleading, to expect otherwise. Accordingly, when papers are transparent and forthright about the limitations and crucial contingencies of their approaches, we consider that to be a great strength, not a weakness. Please keep this in mind when addressing the reviewer comments.

As you address these concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by phone, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: Introduction

● The authors state that:
"However, current ST tools profile the transcriptional expression of only about half as many genes as scRNA-seq (1,000-10,000 compared to 20,000) (Vieth et al. 2019; Stuart and Satija 2019), an issue that can make it problematic to identify cell-types in ST datasets."

A statement which we would argue is partially incorrect, since several of the in situ capture-based methods like Slide-seq v{1,2} and Visium (as well as the first generation Spatial Transcriptomics platform, the predecessor to Visium) capture the near full-transcriptome (in theory all poly-adenylated transcripts). As the authors accurately point out, these in situ capture-based methods indeed operate at a pseudo-bulk level, but that does not change the fact that they profile the same population of genes as most scRNA-seq methods.

● In the context of in situ capture-based methods the authors state that these:
"[..] collect "pseudo-bulk" transcriptome spatially-resolved groups of cells and it becomes challenging to investigate the above questions using these datasets."

While true that answering questions with respect to cell type distribution across the tissue is not as straightforward as with technologies providing single-cell or sub-cellular resolution, there are several tools that perform so called "spatial-decomposition" using single cell data as a reference, for example:
○ stereoscope : https://www.nature.com/articles/s42003-020-01247-y
○ RCTD : https://www.nature.com/articles/s41587-021-00830-w
○ Tangram : https://www.biorxiv.org/content/10.1101/2020.08.29.272831v1
○ Cell2location :https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1
○ SPOTlight : https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab043/6129341

To give an appropriate background to the context, we would be fit to at least acknowledge these methods and that computational solutions (of course with certain drawbacks) exist.

● Speaking of the seqFISH+ data's consistency, the authors report that:
"as a field of view (FOV) (Fig. 1C). The data were highly concordant between sample (Pearson correlation of 0.95) [..]"

To us it's not fully clear how this Pearson correlation was computed? Is it an average correlation value for the pairwise correlation between all six sections? If so, it would be appropriate to report the standard deviation for the 15 pairs. Also, was the data from each section collapsed to represent a form of "bulk"

sample? Could the authors please elaborate a bit on this.

● The authors state that the seqFISH+ dataset profiles 33% of marker genes characterizing the different MOB cell-types, Looking at Supplementary Note 1 and Supplementary Table 1, it seems as if this number is calculated from a single data set (Tepe et al., 2018). While we fully agree that this high lights that marker genes are often lacking in some types of ST data, this is also highly dependent on the reference single cell data set. For example, would the number (33%) be the same had the single cell data from the site mousebrain.org been used instead? We would recommend the authors to revise their statement to be less strong (now saying that 33% marker genes in MOB are missing), to rather say that 33% of the genes in the particular data set was missing.

● In the last part of the introduction the authors write:
"To our knowledge, this is the first such systematic attempt to delineate the principles of brain architecture and intercellular communication by harnessing the unique features of the ST and scRNA-seq technologies."

We would like to make the authors aware of the following publications:

○ "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography" (https://www.nature.com/articles/s42003-020-01247-y). See Figure 2B for an example of how single cell and spatial transcriptomics data was used to chart the brain structure.
○ "Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram" (https://www.biorxiv.org/content/10.1101/2020.08.29.272831v1.full.pdf). see Figures 1-5 for examples of how ow single cell and spatial transcriptomics are used to delineate the brain architecture.
○ "Robust decomposition of cell type mixtures in spatial transcriptomics" (https://www.nature.com/articles/s41587-021-00830-w. See Figure 4-6 for examples of how ow single cell and spatial transcriptomics are integrated to survey the brain structure and cell composition,
○ "Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis" (https://www.sciencedirect.com/science/article/pii/S2211124719311325?via%3Dihub), see Figure 4 for an example of how cell-cell interactions in the mouse brain are surveyed in the spatial data using
Perhaps the authors should be slightly less strong in their claims of novelty.

Results

● The authors cite Kelley 1960 for the backpropagation algorithm, while true that Kelley presented some of the fundamental concepts used in the algorithm, we believe that the formulation and application of the backpropagation algorithm in neural networks most often is attributed to Rumelhart, Hinton and Williams (see: https://www.nature.com/articles/323533a0). Also, since the authors use Tensorflow for their implementation, it might be apt to cite the automatic differentiation scheme that the suite employs.
● The authors state that they used a ten-fold cross-validation scheme to make sure their model is accurate and generalizable, something we fully support and commend the authors for. However, in their description of the cross-validation setup (Supplementary Note 1):

"In order to test the generalizability of STANN on unseen data, we performed a 10-fold cross validation. On each fold, the data is split into 90% training and 10% testing."

To us, this does not sound like a k-fold cross validation setup. The idea with the k-fold cross validation is to first split up the data into k-folds, and then train the model i on the (k-1) partitions that remain when the i:th fold is held out (to be used as test data). If, as the authors describe, random resampling is done at every iteration - there's a risk that similar test/train combinations of train/test partitions occur at each round.[1]

[1]: https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

● In the main text the authors report the following accuracy numbers for training and test data:
"The optimal STANN model showed an average accuracy of 99.55 ± 0.05 % on training data and 95.24% ± 0.31% on separately held-out test data."

While in Supplementary Note 1, they state that:
"On each fold, the data is split into 90% training and 10% testing. Over the 10 folds, the model showed an average accuracy of 99.413 ± 0.059% on the training data and 95.150% ± 0.325% on the separately held-out test data"

While the numbers are similar, maybe the authors want to adjust these pairs of values to make sure that they agree.

● Request: we would like to see how well STANN generalizes to prediction between different single cell datasets. To elaborate, while similar in their superficial form the expression data collected from seqFISH+ and scRNA-seq both host different biases. Thus the data sets still represent different modalities, see [1] for a discussion about platform effects. Hence, we believe that the task of predicting cell type identity in a "held out" partition of the same single cell data set as the model is trained on is significantly easier than to predict the cell type of a seqFISH+ cell using single cell data.

It is admittedly hard to know exactly what biases that are inherent to each method and to correctly capture these in synthetic data, one way to see how the model handles technical artifacts and batch effects would be to train it using single cell data from one publication and then predict the cell types in a single cell data set from a different publication (using the same subsampling strategy as before). Since the model itself, the dense fully connected neural network hasn't been tailored as to specifically work with only mouse brain data, the authors could choose any tissue where two such data sets exist. We are also keen on seeing this analysis, as neural networks have a tendency to overfit the data, while the same does not quite apply other methods used in the benchmarking (e.g., Seurat).

We are requesting this, as much of the model's validity relies on showing good performance on synthetic/semi-synthetic data where the ground truth is known, hence why a lot of rigour should be put into this analysis in order to establish confidence in the secondary results presented using the cell type calling as input.

[1]: https://www.nature.com/articles/s41587-021-00830-w
● Could the authors perhaps, e.g., as Supplementary Notes, provide more details on how the benchmarking analyses were performed. As of now we could not find this in the manuscript nor at the referenced github repository, without this information it's hard to evaluate whether a fair comparison has been made or not.
● Request: Continuing with the benchmarking, we would encourage the authors to include Tangram (https://github.com/broadinstitute/Tangram) in their comparison. The method shares many of its objectives with STANN and while not yet published, we believe it's likely that it will be once this manuscript reaches publication.
● Our final question regarding the benchmarking, which relates to the first, is if the same gene selection process (sPCA) and normalization process was applied to the data before analysis with Seurat and scPred? If not it's hard to say whether it's actually the data curation that is the crucial step in the authors' method or if it's the neural network that gives the increase in performance. To us it's important to disentangle what part of a method that actually provides improved performance, and initialization as well as normalization strategies are important aspects of this, see [1].

[1]: https://www.nature.com/articles/s41587-020-00809-z
● The authors state that six FOVs of seqFISH+ data were used, however when we look in Supplementary Table 4 there are 21 pairs and in total 7 FOV's listed (0-6). Is this a different data set, or a typo?
● Regarding the FOV independence analysis: First, the authors could be more clear with the fact that the chi2-tests are conducted on a pairwise basis. Second, instead of computing both p and q relative entropies, why don't the authors simply use the symmetric Jensen-Shannon divergence metric.[1]

[1]: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence
● Personally, we think that the application of multivariate kernels to assess cell type co-localization patterns is an elegant solution to the problem of working with point-pattern data (single cells). We also have one question, which is whether these density estimates potentially could be confounded by the general cell density? To further explain, could it be that there is an overlap in spatial location between two cell types that would imply a co-localization event, but that this is actually driven by the fact that certain regions are more populous than others and tend to host more cells. Perhaps, one could decompose the density estimates into two parts, one representing the general cell density and one the type specific density and then look at the correlation values between the later components? We are posing this as a question, as we are not sure of the answer, and welcome the authors insights.
● It would be interesting to see the same density estimate plots for the receptor and ligand pairs as for the cell type co-localization, this information could all be included in a single plot (e.g., by adding the receptor and ligand densities in different colors to a gray plot like that in 5C).
● We believe that the authors use the processed seqFISH+ data (where each transcript has been assigned to a cell), but seqFISH+ data also holds information on the exact position of every transcript. It would, as a complement to the above suggested density plots, be interesting to include an image of how the transcripts of ligands and receptors are located in neighboring cells, at least for some of the highlighted pairs; similar to what is done in Figure 4d in [1].

[1]: https://www.nature.com/articles/s41586-019-1049-y
● The authors' strategy to correct for false positives of long-range communications is clever, and seems

like a good approach. Perhaps they could generate synthetic spatial data to show that : (i) their interaction analysis works, and (ii) that the false positives are indeed caught. The large variation in receptor-ligand usage is interesting, but also something that I believe requires more validation to make sure it's not just a technical artifact from the computational methods used.
● Request: Since the dense fully connected network does not host any design elements specific to seqFISH+ data, we don't see an issue with applying the method to other spatial transcriptomics methods like MERFISH or ISS. To show that this strategy is robust we would encourage the authors to test it on more data sets from other platforms.
Discussion
● We would encourage the authors to discuss what aspects of their approach brings novel insights that existing methods like Tangram [1] and SVCA [2] cannot already provide. It would also be of interest to include a commentary on computational run-time in the benchmarking analysis, the authors are using a fairly small network but their approach also requires that several pre-processing steps are executed and they employ a form of cross validation, which I assume increases the run-time quite significantly.

The accuracy of the results is of course of highest importance, but run-time is an important aspect when it comes to a method's usefulness. Extremely computationally expensive methods are not always an option for smaller labs and also less attractive to include in a workflow that will be updated across several iterations.

[1]: https://www.biorxiv.org/content/10.1101/2020.08.29.272831v1
[2]: https://www.sciencedirect.com/science/article/pii/S2211124719311325?via%3Dihub

Methods

● In the section Cell type annotation in scRNA-seq data, the authors write:

● "[..} we first ranked genes based on the variance-to-mean ratio of their expression values and selected 4000 top highly variable genes."

This sounds similar to the procedure that scanpy tends to employ, and the suite is listed as being used for data normalization. We assume that the data normalization was performed before the annotation, and hence that the variable gene selection is also done in the scanpy suite, presumably using the "scanpy.pp.highly_variable_genes" function, if so we think it's apt the the authors state this in the text, as this function implements some correction and do not immediately just compute the variance-to-mean ratio.
● From the sentence "We implemented a multi-layer perceptron model and searched for its optimal architecture (using random initializations in terms of the number of hidden layers, the number of nodes in hidden layers, and the activation functions) using the TensorFlow framework." , it doesn't sound like the authors used a structured grid search approach (which is fine), but rather just sampled configurations randomly - what would be informative is to state from what sets of possible values these values were sampled, i.e., which were to possible activation functions, ranges of possible node sizes, and range of learning rates. Also, importantly, how many different evaluations were made, i.e. were 10 or 1000 models evaluated?

Reviewer #2: The manuscript present STANN, a computational solution - based on neural networks - able to predict cell type localization within a high-resolution spatial transcriptome map by the integration of single cell RNA-sequencing data.

Major comments

The manuscript focus on the integration of single-cell RNA sequencing data with either seqFISH+, or HSDT; both approaches providing spatial transcriptome readouts at a cellular resolution (or nearly). This being said, the authors argue that "current ST tools profile the transcriptional expression of only about half as many genes as scRNA-seq (1,000-10,000 compared to 20,000)" (page 3), or such statement might require to be nuanced. In fact, while the seqFISH+ strategy is bound to a total of 10 thousand interrogated genes due to methodological reasons; other ST approaches, including HSDT, depend on the sequencing depth in use for enhancing the interrogated number of genes; which in addition is also true for single-cell RNA sequencing assays.

Similarly, in page 6, the authors state that "scRNA-seq profiles the complete transcriptome"; completely forgetting that scRNA-seq assays follow a similar strategy than several ST assays; i.e. the capture of messenger RNA via a polyT sequence; followed by reverse transcription and a major step of material amplification prior NGS, which is systematically responsible for a bias on the interrogated transcripts. Furthermore, the sequencing coverage is strongly responsible for determining the "completeness" of the assessed transcriptome.

A last argument that might require to be discussed by the authors is the potential bias on scRNA-seq issued from the enzymatic cell dissociation process, which has been previously described as being a source of artifactual transcriptional response (van den Brink et al., 2017), but also due to the potential over-digestion of a fraction of the cells composing the tissue.

On the ground of these points, the relevance of STANN for integrating scRNA-seq and ST might require its validation in the context of "low resolution" ST data (e.g. Visium generated data, or even those issued from the first generation of DNA arrays described by the team of Dr. Lundeberg), which as consequence might provide higher sequencing depth levels per interrogated spatial region. While STANN has been compared in this article with tools like SEURAT or SCPRED, other tools like Stereoscope, SPOTlight or cell2location were recently shown to be applied for integrating "low resolution" ST maps with single-cell RNA-seq data.

Minor comments:

- Figure 3 might gain on significance if the authors could include the cell-type composition detected on seqFISH+ without the use of STANN. In fact, while the authors stated that only 30% of the known cell type markers are retrieved within such data, the SeqFISH+ article display a certain number of cell types, which might require to be compared with the STANN effort to evaluate the gain on using STANN over the

strategy used in the SeqFISH+ article for such cell-type classification.

- The authors explored the relevance of spatially variable gene regulatory networks implicated on defining a given cell-type and their role on their corresponding intercellular communication. Globally speaking this concept is of major interest, thus counting with strategies to reveal such spatial GRNs are more than welcome. This being said, this manuscript might gain on relevance if the authors could reveal the major gene co-regulatory network per cell types retrieved on each of the FOVs and their commonalities issued of their inter-cellular communication.

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Samee,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager. ***We hope to receive your files within 5 business days, but we recognize that the COVID-19 pandemic may challenge and limit what you can do. Please email me directly if this timing is a problem or you're facing extenuating circumstances.***

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

Title:

The current title is too long (when revising, please make sure it is 150 characters or less).

The title does not capture the conceptual advance of the paper. I suspect it could be more effective. The method you developed is not mentioned at all - integration of scRNA and Spatial Transcriptomics data, machine learning, cell type assignment should be mentioned in some form. I appreciate the motivation to include the downstream analyses enabled by STANN, but this draws focus away from the original contributions you provide in the paper. To capture the possible applications enabled by STANN, you might include, e.g. "…delineates brain tissue substructures" in the second half of the revised title.

As you re-consider your title, note that an effective title is easily found on Pubmed and Google. A trick for thinking about titles is this: ask yourself, "How would I structure a Pubmed search to find this paper?" Put that search together and see whether it comes up is good "sister literature" for this work. If it does, feature the search terms in your title. You also may wish to consider that PubMed is sensitive to small differences in search terms. For example, "NF-kappaB" returned ~84k hits as of March, 2018, whereas "NFkappaB" only returned ~8200. Please ensure that your title contains the most effective version of the search terms you feature.

Abstract:

Please write out what STANN stands for in full when you first use it.

The Abstract reads nicely, but is unfortunately too long. Please condense to 150 words or less.

Manuscript Text:

Please restructure your Introduction to place the biological context at the beginning, previous approaches and their limitations in the middle, and your rationale for how to overcome these as well as an overview of your study design at the end. There needs to be a clearer logical progression from what was done before, to what motivates your current paper, to how you plan to achieve your goals.

There is too much summary and interpretation of results in the Introduction – please replace this with an overview of the study design.

We do not allow supplemental text. Please incorporate Supplementary Note 1 into the main text in the Results section. The benchmarking is quite important and deserves to be placed in the main text.

Also:

- House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.), especially the Results section. These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. "Notably" is suitably neutral to use once or twice if absolutely necessary.
- We don't allow "priority claims" (e.g. new, novel, etc.). For a discussion of why, read: http://crosstalk.cell.com/blog/getting-priorities-right-with-novelty-claims, http://crosstalk.cell.com/blog/novel-insights-into-priority-claims.
- Please only use the word "significantly" in the statistical sense.

Figures and Legends:

- Please ensure that all figures included in your point-by-point response to the reviewers' comments are present within the final version of the paper, either within the main text or within the Supplemental Information.
- Please go over the final versions of the figures and legends with the following in mind:
    - When data visualization tools are used (e.g. UMAP, tSNE), please ensure that the dataset being visualized is named in the figure legend and, when applicable, its accession number is included.
    - When color scales are used, please define them, noting units or indicating "arbitrary units," and specify whether the scale is linear or log.
    - Ensure that every time you have used a graph, you have defined "n's" specifically and listed statistical tests within your Figure Legend.

STAR Methods:

Please format the methods section according to the STAR Methods guidelines. An additional note from Cell Systems: ***If you are using GitHub, please follow the instructions here to archive a "version of record" of your GitHub repo at Zenodo, then report the resulting DOI. Additionally, please note that the Cell Systems strongly recommends that you also include an explicit reference to any***

scripts you may have used throughout your analysis or to generate your figures within section 2 of the resource availability statement.

Thank you!

Reviewer comments:

Reviewer #2: The authors have satisfactorily addressed reviewer comments and made necessary changes.

RESPONSES TO COMMENTS FROM REVIEWER #1

We sincerely thank Reviewer #1 for their generous and helpful comments to improve our manuscript. We hope our revisions, as discussed in our point-by-point responses below, have adequately addressed Reviewer #1's comments and concerns.

Reviewer #1: Introduction

● *The authors state that:*
"However, current ST tools profile the transcriptional expression of only about half as many genes as scRNA-seq (1,000-10,000 compared to 20,000) (Vieth et al. 2019; Stuart and Satija 2019), an issue that can make it problematic to identify cell-types in ST datasets."

A statement which we would argue is partially incorrect, since several of the in situ capture-based methods like Slide-seq v{1,2} and Visium (as well as the first generation Spatial Transcriptomics platform, the predecessor to Visium) capture the near full-transcriptome (in theory all poly-adenylated transcripts). As the authors accurately point out, these in situ capture-based methods indeed operate at a pseudo-bulk level, but that does not change the fact that they profile the same population of genes as most scRNA-seq methods.

We thank the reviewer for suggesting this clarification. In the Introduction of our revised manuscript, we have now clarified the distinction between single-cell resolution ST (sc-ST, such as seqFISH+) and spot-based ST (spot-ST, such as Visium and Slide-seq) that can profile the same set of genes as most scRNA-seq methods but in a pseudo-bulk fashion in spots organized in a regular grid. The relevant text now reads as follows.

"Current ST technologies fall into two broad categories, and importantly, neither category profiles the transcriptome of single-cells. The spot-based ST technologies (spot-ST) use spots (or beads) organized in a regular grid where each spot captures the transcriptome of a variable number of cells (Liao et al. 2020; Stuart and Satija 2019). The commercially available Visium technology, for example, captures 5 to 10 cells (on average) per spot. Because of this "pseudo-bulk" nature of the spot-ST technologies, it becomes challenging to use these datasets to investigate the above questions that require locating single-cells in situ. In particular, although recent methods have used spot-ST data to compute the relative proportion of different cell-types in each spot (Cable et al. 2021; Elosua-Bayes et al. 2021; Andersson et al. 2020; Biancalani et al. 2021; Kleshchevnikov et al. 2020), since the number of cells in each spot is variable and is difficult to determine, the estimated cell-type composition of a given tissue region that comprises multiple spots is not as accurate as could be derived from single-cell resolution spatial data. For the same reason, it is challenging to compute the colocalization of cell-types or their intercellular communications from spot-ST data. One practical solution is to first make a binary presence-absence call in the spots for each cell-type using a predefined threshold on the cell-type's proportion per spot. On the one hand, it is unclear how to define this threshold and whether one should use a cell-type-specific threshold; on the other hand, the conclusions from such

binarized analyses would arguably be sub-optimal than those that a single-cell resolution ST data could offer.

In contrast to the spot-ST technologies, the single-cell ST (sc-ST) technologies record the location of single-cells. Such datasets are, in principle, more well-suited to locate the individual cell-types in situ and study their colocalization and intercellular communication with other cell-types. However, because of their technological design, current sc-ST technologies profile the transcriptional expression of only about half as many genes as commonly profiled by scRNA-seq and spot-ST (1,000-10,000 compared to 20,000) (Vieth et al. 2019; Stuart and Satija 2019), an issue that can make it problematic to identify cell-types in the sc-ST datasets. In particular, when the marker genes of different cell-types are absent in an sc-ST dataset, it is challenging to assign correct types to the cells in that dataset (Dumitrascu et al. 2021). Errors in cell-type assignment, in turn, may lead to inaccurate biological conclusions from an sc-ST data analysis."

● *In the context of in situ capture-based methods the authors state that these:*
"[..] collect "pseudo-bulk" transcriptome spatially-resolved groups of cells and it becomes challenging to investigate the above questions using these datasets."

While true that answering questions with respect to cell type distribution across the tissue is not as straightforward as with technologies providing single-cell or sub-cellular resolution, there are several tools that perform so called "spatial-decomposition" using single cell data as a reference, for example:
○ *stereoscope : https://www.nature.com/articles/s42003-020-01247-y*
○ *RCTD : https://www.nature.com/articles/s41587-021-00830-w*
○ *Tangram : https://www.biorxiv.org/content/10.1101/2020.08.29.272831v1*
○ *Cell2location :https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1*
○ *SPOTlight : https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab043/6129341*

To give an appropriate background to the context, we would be fit to at least acknowledge these methods and that computational solutions (of course with certain drawbacks) exist.

We thank the reviewer for this excellent suggestion to discuss the appropriate context of our work. As we mentioned in the above response, we have now clarified the distinction between single-cell resolution ST (sc-ST, such as seqFISH+) and spot-based ST (spot-ST, such as Visium and Slide-seq), and have acknowledged how some recent tools have used spot-ST data to detect a cell-type's presence in a spot and the cell-type composition of individual spots. We have also discussed why sc-ST is better suited to answer the questions we posed here and how the STANN model, the downstream analyses, and our benchmarking offer important insights toward delineating the principles of brain architecture and intercellular communication and for developing the necessary computational tools in this realm.

● *Speaking of the seqFISH+ data's consistency, the authors report that:*
"as a field of view (FOV) (Fig. 1C). The data were highly concordant between sample (Pearson correlation of 0.95) [..]"

To us it's not fully clear how this Pearson correlation was computed? Is it an average correlation value for the pairwise correlation between all six sections? If so, it would be appropriate to report the standard deviation for the 15 pairs. Also, was the data from each section collapsed to represent a form of "bulk" sample? Could the authors please elaborate a bit on this.

We apologize for creating this confusion. Eng et al. made this comment based on their pilot study of NIH/3T3 fibroblast cells (Eng et al., Nature 2019, PMID: 30911168) to demonstrate the efficacy of seqFISH+. We have now made this point clear in the Introduction section. The relevant text reads as follows.

"seqFISH+ profiles the transcriptional expression from single cells while retaining their spatial information, making it suitable for investigating the types of questions motivated above. Eng et al. also found their pilot seqFISH+ data to be highly reproducible; the data were concordant between two replicates (Pearson correlation of 0.95) (Eng et al. 2019) and they further validated the data with three other bulk and single-cell datasets from RNA-seq, smFISH (single-molecule FISH), and SPOT (RNA sequential probing of targets) (Pearson correlation values > 0.80) (Eng et al. 2019)."

● *The authors state that the seqFISH+ dataset profiles 33% of marker genes characterizing the different MOB cell-types, Looking at Supplementary Note 1 and Supplementary Table 1, it seems as if this number is calculated from a single data set (Tepe et al., 2018). While we fully agree that this high lights that marker genes are often lacking in some types of ST data, this is also highly dependent on the reference single cell data set. For example, would the number (33%) be the same had the single cell data from the site mousebrain.org been used instead? We would recommend the authors to revise their statement to be less strong (now saying that 33% marker genes in MOB are missing), to rather say that 33% of the genes in the particular data set was missing.*

We thank the reviewer for this suggestion. We repeated the analysis using mousebrain.org scRNA-seq data and found that 53% of the marker genes of this dataset are profiled in seqFISH+. We added this list in Supplementary Table S1. We note that, this increase in the number of marker genes (from 33% in Tepe et al. data to 53% in mousebrain.org data) is mainly because the mousebrain.org data features fewer clusters, i.e., cell-types (15 vs. 6). Furthermore, 53% is still a relatively small fraction. Therefore, as the reviewer suggested, we have now revised our statement to reflect the fact that these numbers (33% or 53%) depend on the specific dataset and the number of cell-types being analyzed.

● *In the last part of the introduction the authors write:*
"To our knowledge, this is the first such systematic attempt to delineate the principles of brain architecture and intercellular communication by harnessing the unique features of the ST and scRNA-seq technologies."

We would like to make the authors aware of the following publications:

○ "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography" (https://www.nature.com/articles/s42003-020-01247-y). See Figure 2B for an example of how single cell and spatial transcriptomics data was used to chart the brain structure.
○ "Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram" (https://www.biorxiv.org/content/10.1101/2020.08.29.272831v1.full.pdf). see Figures 1-5 for examples of how ow single cell and spatial transcriptomics are used to delineate the brain architecture.
○ "Robust decomposition of cell type mixtures in spatial transcriptomics" (https://www.nature.com/articles/s41587-021-00830-w. See Figure 4-6 for examples of how ow single cell and spatial transcriptomics are integrated to survey the brain structure and cell composition,
○ "Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis" (https://www.sciencedirect.com/science/article/pii/S2211124719311325?via%3Dihub), see Figure 4 for an example of how cell-cell interactions in the mouse brain are surveyed in the spatial data using
Perhaps the authors should be slightly less strong in their claims of novelty.

We thank the reviewer for this excellent suggestion to properly contextualize our work. We have now discussed how the manuscripts mentioned above have used spot-ST data to detect a cell-type's presence in a spot and the cell-type composition of individual spots, but how the nature of spot-ST data fundamentally limits the types of investigations we can do using these datasets. As suggested by the reviewer, we also report additional benchmarking of STANN that not only showed STANN's efficacy but also suggested that methods for integrating spot-ST and scRNA-seq data are likely to produce suboptimal results if applied to integrate sc-ST and scRNA-seq data (Supplementary Note 1, Discussion). Therefore, we now summarize our contribution as providing important insights into brain architecture and intercellular communication principles and for developing the necessary computational tools for similar investigations. The relevant text reads as follows.

"The STANN model and the downstream analyses featured in this work are motivated by a critical need for using sc-ST data to delineate the consistent and the variable aspects of the architecture and intercellular communication mechanisms at a single-cell resolution in complex tissue regions, such as MOB, beyond their conventional layer-based architectural description. Previous studies in this realm have used spot-ST data and described brain architecture in terms of different cell-types' presence and their proportions in individual spots (Andersson et al. 2020; Cable et al. 2021; Biancalani et al. 2021). However, as we discussed above, since the number of cells in each spot is variable and is difficult to determine, the estimated cell-type composition of a given tissue region that comprises multiple spots is not as accurate as could be derived from single-cell resolution spatial data. For the same reason, spot-ST data is not particularly suitable for computing colocalization of cell-types or their intercellular communications at single-cell resolution. Given that sc-ST data are better suited to answer these questions, we developed STANN to tackle the computational challenges associated with sc-ST data and studied MOB

architecture. Besides featuring new approaches to quantify the colocalization of cell-types and study the variation in intercellular communication, we also benchmarked STANN against alternative models. STANN outperformed the alternative methods in this benchmarking. The analyses also suggested that methods for integrating spot-ST and scRNA-seq data are likely to produce suboptimal results if applied to integrate sc-ST and scRNA-seq data (Supplementary Note 1, Discussion). Altogether, our work offers important insights into brain architecture and intercellular communication principles and for developing the necessary computational tools for similar investigations."

Results

● *The authors cite Kelley 1960 for the backpropagation algorithm, while true that Kelley presented some of the fundamental concepts used in the algorithm, we believe that the formulation and application of the backpropagation algorithm in neural networks most often is attributed to Rumelhart, Hinton and Williams (see: https://www.nature.com/articles/323533a0). Also, since the authors use Tensorflow for their implementation, it might be apt to cite the automatic differentiation scheme that the suite employs.*

We thank the reviewer for this excellent suggestion. In the revised version, we have referenced Rumelhart et al.'s work on the backpropagation algorithm (Rumelhart et al., Nature, 1986). We also cited the TensorFlow library back-end for backpropagation using automatic differentiation (Abadi et al., Proc. of OSDI, 2016).

● *The authors state that they used a ten-fold cross-validation scheme to make sure their model is accurate and generalizable, something we fully support and commend the authors for. However, in their description of the cross-validation setup (Supplementary Note 1):*

"In order to test the generalizability of STANN on unseen data, we performed a 10-fold cross validation. On each fold, the data is split into 90% training and 10% testing."

To us, this does not sound like a k-fold cross validation setup. The idea with the k-fold cross validation is to first split up the data into k-folds, and then train the model i on the (k-1) partitions that remain when the i:th fold is held out (to be used as test data). If, as the authors describe, random resampling is done at every iteration - there's a risk that similar test/train combinations of train/test partitions occur at each round.[1]

[1]: https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation

We apologize for this misleading description of our cross-validation scheme. We indeed followed the 10-fold cross validation scheme that the reviewer pointed out, where we first split the data into 10 folds (each containing about 10% of the data) and at the i-th iteration, we hold out the i-th fold as our test data. Specifically, we used the stratified 10-fold cross validation scheme from scikit-learn so that the folds preserve the percentage of samples for each class

(cell-type). We have now revised the text to avoid confusion.

● *In the main text the authors report the following accuracy numbers for training and test data: "The optimal STANN model showed an average accuracy of 99.55 ± 0.05 % on training data and 95.24% ± 0.31% on separately held-out test data."*

While in Supplementary Note 1, they state that:
"On each fold, the data is split into 90% training and 10% testing. Over the 10 folds, the model showed an average accuracy of 99.413 ± 0.059% on the training data and 95.150% ± 0.325% on the separately held-out test data"

While the numbers are similar, maybe the authors want to adjust these pairs of values to make sure that they agree.

We sincerely thank the reviewer for their careful reading of the manuscript and apologize for this mismatch. The two sets of results came from two different runs of the model. In the revised version of our manuscript, we report the same run's output consistently throughout the text.

● *Request: we would like to see how well STANN generalizes to prediction between different single cell datasets. To elaborate, while similar in their superficial form the expression data collected from seqFISH+ and scRNA-seq both host different biases. Thus the data sets still represent different modalities, see [1] for a discussion about platform effects. Hence, we believe that the task of predicting cell type identity in a "held out" partition of the same single cell data set as the model is trained on is significantly easier than to predict the cell type of a seqFISH+ cell using single cell data.*

It is admittedly hard to know exactly what biases that are inherent to each method and to correctly capture these in synthetic data, one way to see how the model handles technical artifacts and batch effects would be to train it using single cell data from one publication and then predict the cell types in a single cell data set from a different publication (using the same subsampling strategy as before). Since the model itself, the dense fully connected neural network hasn't been tailored as to specifically work with only mouse brain data, the authors could choose any tissue where two such data sets exist. We are also keen on seeing this analysis, as neural networks have a tendency to overfit the data, while the same does not quite apply other methods used in the benchmarking (e.g., Seurat).

We are requesting this, as much of the model's validity relies on showing good performance on synthetic/semi-synthetic data where the ground truth is known, hence why a lot of rigour should be put into this analysis in order to establish confidence in the secondary results presented using the cell type calling as input.

[1]: https://www.nature.com/articles/s41587-021-00830-w

We thank the reviewer for this excellent suggestion.

As the reviewer suggested, we collected two independent lung scRNA-seq samples from the Tabula Sapiens Consortium. We separately processed the datasets following standard approaches, namely scale-factor transformation and log-transformation, and used cell-type annotation provided by the Tabula Sapiens Consortium. For the STANN pipeline, we then quantile normalized the datasets and applied STANN to learn cell-type mapping from one sample and predicting the cell-types of the cells in the second sample. STANN provided an accuracy of 96.66%, compared to Seurat's 85.33%, scPred's 78.37%, and Tangram's 41.85%. We report these results in Supplementary Note 1, our revised supplementary text elaborating on the benchmarking analyses (we discussed more in our response to the next comment).

Furthermore, following the reviewer's suggestion in another comment, we applied STANN on MERFISH data. This provided us with another check of STANN's efficacy for learning cell-types from scRNA-seq data and making predictions on MERFISH data. STANN's accuracy of 87.62% in this case outperformed Seurat's 82.52%, scPred's 52.08%, and Tangram's 41.86%.

We again thank the reviewer for this suggestion and hope the above analyses address the reviewer's concern.

● *Could the authors perhaps, e.g., as Supplementary Notes, provide more details on how the benchmarking analyses were performed. As of now we could not find this in the manuscript nor at the referenced github repository, without this information it's hard to evaluate whether a fair comparison has been made or not.*

We again thank the reviewer for this helpful suggestion. We have now provided the details of our benchmarking in Supplementary Note 1. Briefly, we have benchmarked STANN against Seurat, scPred, and Tangram using the following three datasets and the procedure of each method published in their corresponding tutorial or manuscript.

1. Tepe et al.'s mouse olfactory bulb scRNA-seq data (Tepe et al., Cell Reports, 2018),
2. Two independent samples of Tabula Sapiens lung scRNA-seq data, and
3. MERFISH (Moffitt et al., Science, 2018) data.

● *Request: Continuing with the benchmarking, we would encourage the authors to include Tangram (https://github.com/broadinstitute/Tangram) in their comparison. The method shares many of its objectives with STANN and while not yet published, we believe it's likely that it will be once this manuscript reaches publication.*

We thank the reviewer for this suggestion. As we noted above and discussed in detail in Supplementary Note 1, we have added Tangram to our benchmarking. Tangram did not perform better than scPred, Seurat, and STANN in our benchmarking.

We wish to share some thoughts here, although the following requires testing Tangram in more settings, and we think Tangram authors will optimize the method as the manuscript goes

through the peer-review process. We think the above performance results are not entirely unexpected from Tangram, which is developed as a more general method for transferring annotations from a source scRNA-seq dataset (S) to a target ST dataset (G). To this end, Tangram learns a mapping matrix, M, by maximizing the similarity of gene expression distribution and cell-densities between M^TS and G. When the annotations to be transferred are cell-types from a source scRNA-seq data, Tangram's formulation essentially reduces to deconvolving the data, as the original manuscript mentions, "This corresponds to probabilistic mapping and can be interpreted as the mixture of cell types which best explain the in situ gene expression." Thus, Tangram can be directly applied on spot-ST data, which indeed requires computing the relative proportion of different cell-types in each spot. In the case of sc-ST data, since deconvolution is not necessary, Tangram needs to use specific assumptions in their formulation. In particular, for density at each sc-ST location, Tangram uses a uniform prior on the cell-types and we think, for sc-ST data, then the model becomes more complicated than necessary to optimize the correlation of gene expression distribution between M^TS and G. We think that Seurat's CCA (canonical correlation analysis) could achieve the same goal in a more straight-forward manner. We anticipate this could be the same case for other similarly complex models (e.g., cell2location) that were developed for more generally deconvolving spot-ST data, but for sc-ST data, needs to make specific assumptions and essentially optimizes a linear correlation structure. Importantly, in their preprint versions, neither Tangram nor cell2location compared itself against Seurat for any sc-ST data. In fact, neither of the two methods showed an application for transferring cell-type labels from scRNA-seq data to sc-ST data; for example, Tangram's application to MERFISH dataset was to increase gene throughput (predict the expression of more genes than profiled by MERFISH). We think as those manuscripts go through the peer-review process, the authors will add more insights on their applicability to sc-ST data for transferring cell-type labels.

Overall, our benchmarking suggested that methods for integrating spot-ST with scRNA-seq might be sub-optimal for integrating sc-ST with scRNA-seq, and one should use methods like STANN that were specifically developed for sc-ST data. To know the cell-types in a spatial dataset, although both sc-ST and spot-ST require integration with scRNA-seq, the methods require solving two different computational problems. In the case of sc-ST, a tool like STANN needs to find a mapping to cell-types from a fewer number of genes, even from genes that were not used in the first place to define the cell-types in the scRNA-seq data. In the case of spot-ST, since each spot contains a variable number of cells, the problem is to deconvolve the data into relative proportions of different cell-types. Since sc-ST data do not require deconvolution, when applied to sc-ST datasets, the algorithms developed for spot-ST data essentially optimize linear correlations of gene expression between the given sc-ST and scRNA-seq datasets. However, as our benchmarkings suggest, explicitly learning a non-linear high dimensional mapping function, as STANN does, is potentially more useful than that for sc-ST data.

● *Our final question regarding the benchmarking, which relates to the first, is if the same gene selection process (sPCA) and normalization process was applied to the data before analysis with Seurat and scPred? If not it's hard to say whether it's actually the data curation that is the crucial step in the authors' method or if it's the neural network that gives the increase in*

performance. To us it's important to disentangle what part of a method that actually provides improved performance, and initialization as well as normalization strategies are important aspects of this, see [1].
[1]: https://www.nature.com/articles/s41587-020-00809-z

We thank the reviewer for this interesting question about which step in STANN's pipeline potentially provides the most improvement in performance. We have now added this point in our Discussion section and share our thoughts below.

We note that, to map cell-types in an sc-ST dataset through integrating it with an scRNA-seq dataset, we need a function that computes the type of each sc-ST cell given its shared genes with the scRNA-seq dataset. Since deep neural networks are effective in learning high-dimensional and non-linear functions in data-driven fashion, we anticipate that the neural network component of our approach is key to its higher accuracy. However, STANN's class-imbalance aware loss function was also critical, especially given the high imbalance of cell counts in different cell-types of scRNA-seq datasets. In our exploratory runs, the supervised PCA (sPCA) consistently helped improve STANN's cross-validation accuracy by up to 5%, but without the class-imbalance aware loss function, our cross-validation accuracies would often drop by 10% to 15%.

On the point of preprocessing, all methods (Seurat, scPred, and Tangram) apply the common preprocessing steps, such as scale factor normalization and log transformation. As we discuss below, Seurat and scPred employs their own additional preprocessing, which presumably were optimized for their overall pipeline, and in our benchmarking, we retained the same steps as the original pipelines.

scPred normalizes the two input datasets using Harmony (Korsunsky et al., Nature Methods, 2019) to align them in a low-dimensional space. The Harmony algorithm takes a PCA embedding of the cells and their batch assignments, and returns a batch corrected embedding. Once normalized using Harmony, scPred performs a PCA based feature selection and then uses a support vector machine (SVM) classifier to assign types to the cells.

Seurat uses canonical correlation analysis (CCA) to integrate the two input datasets according to their shared correlation structure (Stuart and Satija, Nature Reviews Genetics, 2019). This is more akin to an unsupervised clustering approach with its own correlation-based strategy for feature selection. Tangram maximizes the cosine similarity between the predicted and the expected gene expression values. Thus, unlike scPred, Seurat and Tangram do not require an explicit normalization to make the two input datasets' gene expression distributions comparable.

Overall, we posit that the deep neural network and the class-imbalance aware loss function are the two most critical drivers of STANN's improved performance. However, as the reviewer pointed, the initial steps for making the two data distributions comparable are likely to impact any algorithm's accuracy. In our exploratory analyses and benchmarking runs, STANN's

pipeline did not require a sophisticated method like Harmony; the quantile normalization approach was sufficient. As we are learning from our other projects, Harmony has its strengths and weaknesses, and we think it is more appropriate to recommend that future works to integrate sc-ST with scRNA-seq should perform careful exploratory analyses to select the initial normalization approach from the available options such as quantile normalization, Harmony, or any similar tools.

● *The authors state that six FOVs of seqFISH+ data were used, however when we look in Supplementary Table 4 there are 21 pairs and in total 7 FOV's listed (0-6). Is this a different data set, or a typo?*

There are seven FOVs in this seqFISH+ data. We have now checked and corrected the numbers throughout our manuscript. We apologize for this typo and thank the reviewer for their careful read.

● *Regarding the FOV independence analysis: First, the authors could be more clear with the fact that the chi2-tests are conducted on a pairwise basis. Second, instead of computing both p and q relative entropies, why don't the authors simply use the symmetric Jensen-Shannon divergence metric.[1]*
[1]: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

We thank the reviewer for these great suggestions. We now clarified that the chi-squared tests were conducted across all pairwise comparisons of FOVs. We also agree that Jensen-Shannon divergence could be a more direct metric to report this pairwise comparison and now reported Jensen-Shannon divergence in Supplementary Table S4. The arithmetic and harmonic means of the KL divergence values and the Jensen-Shannon divergence all agree with the chi-squared tests.

● *Personally, we think that the application of multivariate kernels to assess cell type co-localization patterns is an elegant solution to the problem of working with point-pattern data (single cells). We also have one question, which is whether these density estimates potentially could be confounded by the general cell density? To further explain, could it be that there is an overlap in spatial location between two cell types that would imply a co-localization event, but that this is actually driven by the fact that certain regions are more populous than others and tend to host more cells. Perhaps, one could decompose the density estimates into two parts, one representing the general cell density and one the type specific density and then look at the correlation values between the later components? We are posing this as a question, as we are not sure of the answer, and welcome the authors insights.*

We thank the reviewer for this excellent question and finding the approach useful. To make this approach broadly applicable, it will be important to consider the point of non-uniform cell-densities. We wish to note that, in our inspection of Eng et al.'s seqFISH+ data, cell densities were not specifically high in any particular part of an olfactory bulb FOV (Fig. 4 and Extended Data Figure 10 of Eng et al., Nature, 2019).

We think partial correlation coefficient (CC) could handle the scenario of non-uniform cell densities. In particular, we would first take the KDE considering all cell-types in the FOV; i.e., we would compute the KDE of the data C = {(x_i, y_i, Z)} where (x_i, y_i) is the spatial coordinate of the i-th datapoint and Z is a binary variable indicating if there is a cell of any type at (x_i, y_i). Then, instead of taking the Pearson CC of the KDEs of two cell-types A and B, we could take the partial Pearson CC of the KDEs of A and B given the KDE of C. Using available statistical packages (Kim S, Commun Stat Appl Methods, 2015), we could also compute a p-value of the partial CC.

● *It would be interesting to see the same density estimate plots for the receptor and ligand pairs as for the cell type co-localization, this information could all be included in a single plot (e.g., by adding the receptor and ligand densities in different colors to a gray plot like that in 5C).*

We have now shown these density plots separately in Supplemental Figure S6. We note that, since not all cells of a given type express the receptor/ligand of interest, the density plot of receptor/ligand expressing cells and that of all cells may not have their highest densities in the same region and overlaying the two density plots complicates the visualization. Hence, we show the density plots separately.

● *We believe that the authors use the processed seqFISH+ data (where each transcript has been assigned to a cell), but seqFISH+ data also holds information on the exact position of every transcript. It would, as a complement to the above suggested density plots, be interesting to include an image of how the transcripts of ligands and receptors are located in neighboring cells, at least for some of the highlighted pairs; similar to what is done in Figure 4d in [1].*

[1]: https://www.nature.com/articles/s41586-019-1049-y

We again thank the reviewer for an interesting suggestion. Unfortunately, we found that the processed seqFISH+ data does not include transcript locations for olfactory bulb cells.

Since it would be complicated to make a mechanistic conclusion from this analysis, we assumed this was not a high-priority suggestion from the reviewer. Thus, we focused more on incorporating the other prioritized suggestions of the reviewers and did not delay the revision for collecting and processing the raw data for this analysis. We sincerely hope the reviewer will favorably consider this point.

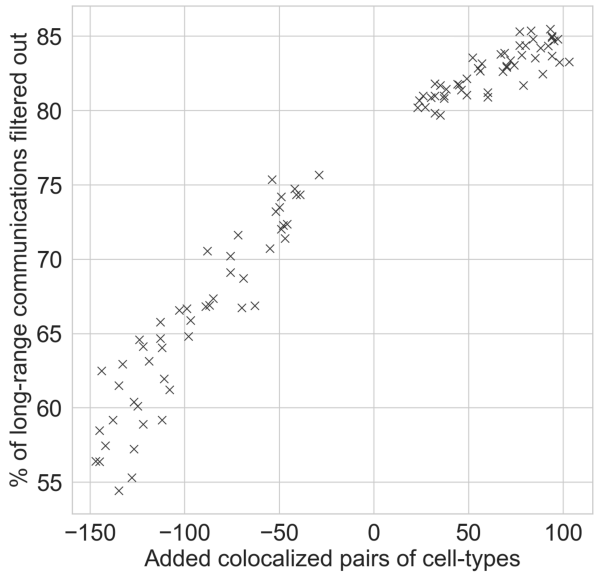
● *The authors' strategy to correct for false positives of long-range communications is clever, and seems like a good approach. Perhaps they could generate synthetic spatial data to show that : (i) their interaction analysis works, and (ii) that the false positives are indeed caught. The large variation in receptor-ligand usage is interesting, but also something that I believe requires more validation to make sure it's not just a technical artifact from the computational methods used.*

We thank the reviewer for this suggestion. We have now simulated spatial colocalization/separation patterns and checked if the number of long-range communications marked as false-positives by our approach increases or decreases as we introduce more spatial colocalization or separation, respectively.

We note that we can represent the spatial colocalization/separation pattern in an FOV using a graph where nodes represent cell-types and edges represent pairs of colocalized cell-types. The absence of an edge between two nodes represents a spatial separation of the corresponding cell-types.

Thus, we first compute seven graphs representing the spatial colocalization/separation patterns of cell-types in the seven FOVs of this seqFISH+ data. We then create 100 synthetic patterns of spatial colocalization/separation for the seven FOVs as follows. For 50 patterns, we increased spatial separation by randomly removing k edges from the seven graphs, where we sampled k uniformly from the range [10, 50% of the total number of edges in the seven graphs]. Similarly, 50 times we increased spatial colocalization by randomly adding k new edges to the seven graphs, again sampling k uniformly from the range [10, 50% of the number of node-pairs that did not have an edge in the original seven graphs].

For each of the 100 spatial colocalization/separation patterns generated above, we repeat our analysis and record the fraction of intercellular communications between spatially separate cell-types (i.e., long-range communications) that our approach marks as false-positives and filters out. This analysis showed that the number of false-positives marked by our approach increases or decreases as we simulate more spatial colocalization or separation, respectively, as we show in the following plot. Negative values in the X-axis denote removal of edges.



● *Request: Since the dense fully connected network does not host any design elements specific to seqFISH+ data, we don't see an issue with applying the method to other spatial*

transcriptomics methods like MERFISH or ISS. To show that this strategy is robust we would encourage the authors to test it on more data sets from other platforms.

We thank the reviewer for this suggestion. We have now included MERFISH in our benchmarking. STANN's accuracy of 87.62% in this case outperformed Seurat's 82.52%, scPred's 52.08%, and Tangram's 41.86%. We reported the details in Supplementary Note 1.

Discussion

● *We would encourage the authors to discuss what aspects of their approach brings novel insights that existing methods like Tangram [1] and SVCA [2] cannot already provide. It would also be of interest to include a commentary on computational run-time in the benchmarking analysis, the authors are using a fairly small network but their approach also requires that several pre-processing steps are executed and they employ a form of cross validation, which I assume increases the run-time quite significantly.*

The accuracy of the results is of course of highest importance, but run-time is an important aspect when it comes to a method's usefulness. Extremely computationally expensive methods are not always an option for smaller labs and also less attractive to include in a workflow that will be updated across several iterations.
[1]: https://www.biorxiv.org/content/10.1101/2020.08.29.272831v1
[2]: https://www.sciencedirect.com/science/article/pii/S2211124719311325?via%3Dihub

We thank the reviewer for this suggestion. For SVCA, we have now noted the following in the section titled "Widespread spatial variation in intercellular communication mechanisms give rise to spatially localized gene regulatory networks". The relevant text reads as follows.

"This concept of spatially localized GRNs has been noted in the literature (Yang, Fang, and Shen 2019) and aligns with previous observations on spatial variation of gene expression because of intercellular communication (Arnol et al. 2019). However, to our knowledge, the existence and role of spatially localized GRNs of different cell-types in mediating their intercellular communications have never been discussed."

In other words, while SVCA showed spatial variation in gene expression because of intercellular communication, the analysis did not take the different cell-types into account. The SVCA manuscript also mentions this as one of their motivating points: "In contrast to previous methods, our model directly uses the spatial coordinates and the gene expression profile of each cell as input, thereby avoiding the need to define discrete cell types ..."

For Tangram and other relevant works, we have now discussed the following in our Discussion section.

“As we have noted in the Introduction, our aim was to fill in a critical gap in this realm since previous studies have described brain architecture using spot-ST data in terms of the proportion of different cell-types in the spots (Andersson et al. 2020; Cable et al. 2021; Biancalani et al. 2021). However, since the number of cells in each spot of spot-ST data is difficult to determine and variable between spots, one cannot comment on a tissue region’s cell-type composition, colocalization of cell-types or their intercellular communications from spot-ST data as accurately as one could from sc-ST data. For those analyses, one practical solution is to first make a binary presence-absence call in the spots for each cell-type using a predefined threshold on the cell-type’s proportion per spot. On one hand, it is not clear how to define this threshold and whether one should use a cell-type specific threshold, on the other hand, the conclusions from such binarized analyses would arguably be sub-optimal than those that an sc-ST data could offer. That is why, we posited that sc-ST data are better suited to reveal the consistent and the variable aspects of the architecture and intercellular communication mechanisms in MOB beyond its layer-based architectural description, and developed STANN to tackle the associated computational challenges.”

We hope the above revisions address the reviewer’s suggestion.

We have now included a discussion on the runtime in Supplementary Note 1. We report the following.

“All benchmarking runs were conducted in a server equipped with 80x Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz processors and with 256GB of DDR4 memory. From the benchmarked methods, we found that STANN had similar runtime as Seurat and scPred and significantly lower memory usage than the three other benchmarked methods. Seurat and Tangram had the highest peak memory usage. STANN and Tangram were run on a single processor, Seurat and scPred were run on 10 processors to leverage functions that supported multithreading. Training a model with ~10K cells and ~150 genes (MERFISH benchmarking) and predicting on ~90K cells and ~150 genes took both STANN and Seurat runtime took < 20min, scPred and Tangram > 1hr. Training a model with ~10K cells and ~10K genes (Tabula Sapiens and MOB runs) and predicting in ~10K cells and ~10K genes took STANN, Seurat and scPred < 1hr and > 1hr for Tangram.”

Methods

● *In the section Cell type annotation in scRNA-seq data, the authors write:*
"[..} we first ranked genes based on the variance-to-mean ratio of their expression values and selected 4000 top highly variable genes."
This sounds similar to the procedure that scanpy tends to employ, and the suite is listed as being used for data normalization. We assume that the data normalization was performed before the annotation, and hence that the variable gene selection is also done in the scanpy suite, presumably using the "scanpy.pp.highly_variable_genes" function, if so we think it's apt the the authors state this in the text, as this function implements some correction and do not

immediately just compute the variance-to-mean ratio.

We thank the reviewer for this suggestion. We indeed performed the highly variable gene selection using “scanpy.pp.highly_variable_genes” and clarified this on the manuscript. The relevant text reads as follows.

“Following the common steps for cell-type annotation in scRNA-seq data (Zheng et al. 2017; Butler et al. 2018), we ran Scanpy’s highly_variable_genes function (v1.5.0) and selected the top 4000 genes based on the variance-to-mean ratio of their expression values.”

● *From the sentence "We implemented a multi-layer perceptron model and searched for its optimal architecture (using random initializations in terms of the number of hidden layers, the number of nodes in hidden layers, and the activation functions) using the TensorFlow framework." , it doesn't sound like the authors used a structured grid search approach (which is fine), but rather just sampled configurations randomly - what would be informative is to state from what sets of possible values these values were sampled, i.e., which were to possible activation functions, ranges of possible node sizes, and range of learning rates. Also, importantly, how many different evaluations were made, i.e. were 10 or 1000 models evaluated?*

We apologize for our unclear description. In the Methods section of our revised manuscript, we have now elaborated on our hyperparameter search. The relevant text reads as follows.

“We implemented hyperparameter optimization using KerasTuner (O’Malley et al. 2019). Specifically, we used the hyperband algorithm (L. Li et al. 2018) which performsdoes a computationally efficient random search through adaptive resource allocation and early stopping. We evaluated 2000 random models with varying the following parameters within the shown ranges or sets.

Dense layers neurons: Min: 10, Max: 500
Activation functions: Relu, Sigmoid and Tanh
Learning rates: Min: 1e-4, Max: 1e-2, and
Optimizers: Adam, SGD, and RMSprop.”

RESPONSES TO COMMENTS FROM REVIEWER #2

Reviewer #2: The manuscript present STANN, a computational solution - based on neural networks - able to predict cell type localization within a high-resolution spatial transcriptome map by the integration of single cell RNA-sequencing data.

We thank Reviewer #2 for their careful reading of our manuscript and thoughtful comments. We hope our responses and revisions, as discussed below, have adequately addressed their comments and concerns.

Major comments

The manuscript focus on the integration of single-cell RNA sequencing data with either seqFISH+, or HSDT; both approaches providing spatial transcriptome readouts at a cellular resolution (or nearly). This being said, the authors argue that "current ST tools profile the transcriptional expression of only about half as many genes as scRNA-seq (1,000-10,000 compared to 20,000)" (page 3), or such statement might require to be nuanced. In fact, while the seqFISH+ strategy is bound to a total of 10 thousand interrogated genes due to methodological reasons; other ST approaches, including HSDT, depend on the sequencing depth in use for enhancing the interrogated number of genes; which in addition is also true for single-cell RNA sequencing assays.

We thank the reviewer for this suggestion. In the Introduction of this revised manuscript, we have elaborated on the distinction between spot-based ST (spot-ST) and single-cell resolution ST (sc-ST). We hope this elaboration captures the nuances that the reviewer noted above. We have also discussed how the nature of spot-ST data fundamentally limits the types of investigations we can do using these datasets. The relevant text reads as follows.

“Current ST technologies fall into two broad categories, and importantly, neither category profiles the transcriptome of single-cells. The spot-based ST technologies (spot-ST) use spots (or beads) organized in a regular grid where each spot captures the transcriptome of a variable number of cells (Liao et al. 2020; Stuart and Satija 2019). The commercially available Visium technology, for example, captures 5 to 10 cells (on average) per spot. Because of this “pseudo-bulk” nature of the spot-ST technologies, it becomes challenging to use these datasets to investigate the above questions that require locating single-cells in situ. In particular, although recent methods have used spot-ST data to compute the relative proportion of different cell-types in each spot (Cable et al. 2021; Elosua-Bayes et al. 2021; Andersson et al. 2020; Biancalani et al. 2021; Kleshchevnikov et al. 2020), since the number of cells in each spot is variable and is difficult to determine, the estimated cell-type composition of a given tissue region that comprises multiple spots is not as accurate as could be derived from single-cell resolution spatial data. For the same reason, it is challenging to compute the colocalization of cell-types or their intercellular communications from spot-ST data. One practical solution is to first make a binary presence-absence call in the spots for each cell-type using a predefined threshold on the cell-type’s

proportion per spot. On the one hand, it is unclear how to define this threshold and whether one should use a cell-type-specific threshold; on the other hand, the conclusions from such binarized analyses would arguably be sub-optimal than those that a single-cell resolution ST data could offer.

In contrast to the spot-ST technologies, the single-cell ST (sc-ST) technologies record the location of single-cells. Such datasets are, in principle, more well-suited to locate the individual cell-types in situ and study their colocalization and intercellular communication with other cell-types. However, because of their technological design, current sc-ST technologies profile the transcriptional expression of only about half as many genes as commonly profiled by scRNA-seq and spot-ST (1,000-10,000 compared to 20,000) (Vieth et al. 2019; Stuart and Satija 2019), an issue that can make it problematic to identify cell-types in the sc-ST datasets. In particular, when the marker genes of different cell-types are absent in an sc-ST dataset, it is challenging to assign correct types to the cells in that dataset (Dumitrascu et al. 2021). Errors in cell-type assignment, in turn, may lead to inaccurate biological conclusions from an sc-ST data analysis."

Similarly, in page 6, the authors state that "scRNA-seq profiles the complete transcriptome"; completely forgetting that scRNA-seq assays follow a similar strategy than several ST assays; i.e. the capture of messenger RNA via a polyT sequence; followed by reverse transcription and a major step of material amplification prior NGS, which is systematically responsible for a bias on the interrogated transcripts. Furthermore, the sequencing coverage is strongly responsible for determining the "completeness" of the assessed transcriptome.

We again thank the reviewer for these suggestions. We have now removed the wording of "complete" transcriptome. We rather noted that under the current standard practices, both the spot-ST and scRNA-seq technologies have been shown to profile ~20000 genes.

A last argument that might require to be discussed by the authors is the potential bias on scRNA-seq issued from the enzymatic cell dissociation process, which has been previously described as being a source of artifactual transcriptional response (van den Brink et al., 2017), but also due to the potential over-digestion of a fraction of the cells composing the tissue.

On the ground of these points, the relevance of STANN for integrating scRNA-seq and ST might require its validation in the context of "low resolution" ST data (e.g. Visium generated data, or even those issued from the first generation of DNA arrays described by the team of Dr. Lundeberg), which as consequence might provide higher sequencing depth levels per interrogated spatial region. While STANN has been compared in this article with tools like SEURAT or SCPRED, other tools like Stereoscope, SPOTlight or cell2location were recently shown to be applied for integrating "low resolution" ST maps with single-cell RNA-seq data.

We apologize for any confusion, but STANN has been developed for single-cell resolution ST (sc-ST) data. Although both sc-ST and spot-based ST (spot-ST; referred above as "low

resolution" ST by the reviewer) data require integration with scRNA-seq, the integration methods require solving two different computational problems.

In the case of sc-ST, a tool like STANN needs to find a mapping to cell-types from a fewer number of genes, even from genes that were not used in the first place to define the cell-types. In the case of spot-ST, since the number of cells in each location ("spots" or "beads") is variable and is difficult to determine, the problem is to deconvolve the data into relative proportions of different cell-types. This is beyond the scope of the tools developed for sc-ST data. However, as the reviewer has pointed, we agree that the study would benefit from more benchmarking of STANN. We believe that, from the same reasoning Reviewer #1 has suggested us to apply STANN on other sc-ST data (such as MERFISH) and to add another state-of-the-art tool called Tangram to compare against STANN on seqFISH+ data. We have performed both analyses and reported the results in Supplementary Note 1. We also believe that, because of this fundamental difference between the two computational problems, Reviewer #1 did not suggest us to apply STANN on low-resolution spot-ST data. But again, we sincerely thank the reviewer for suggesting us to perform additional benchmarking. As we show in Supplementary Note 1 and discuss in the Discussion section, these new analyses not only showed STANN's efficacy but also suggested that methods for integrating spot-ST and scRNA-seq data are likely to produce suboptimal results if applied to integrate sc-ST and scRNA-seq data.

Minor comments

- Figure 3 might gain on significance if the authors could include the cell-type composition detected on seqFISH+ without the use of STANN. In fact, while the authors stated that only 30% of the known cell type markers are retrieved within such data, the SeqFISH+ article display a certain number of cell types, which might require to be compared with the STANN effort to evaluate the gain on using STANN over the strategy used in the SeqFISH+ article for such cell-type classification.

We respectfully note that this comparison could be misleading. Although the seqFISH+ article (Eng et al., Nature, 2019) studied mouse olfactory bulb, they used the mouse cortical cell scRNA-seq data from Tasic et al. (Nature Neuroscience, 2016) as their reference. Besides the fact that this data was collected from a different brain section, the data had fewer than 1,800 cells raising concerns about the accuracy of cluster assignment. However, our data set, from Tepe et al. (Cell Reports, 2018), was specifically from olfactory bulb and had ~10K cells. Furthermore, Eng et al. did not benchmark their support-vector machine (SVM) based algorithm for cell-type prediction. We wish to note that, scPred is another SVM-based approach (but more sophisticated than Eng et al.'s algorithm) and did not show superior performance in our benchmarking. Overall, considering there are issues with both the reference scRNA-seq data and the approach of SeqFISH+, we are afraid that it might not be appropriate to perform a comparison taking seqFISH+'s cell-type labeling as ground truth.

However, we completely agree with the reviewer about the need for benchmarking STANN. Thus, in the revision, we have benchmarked STANN and other competing methods using the

Tabula Sapiens scRNA-seq data, where the methods were trained on one scRNA-seq sample and used to predict the type of cells in a held-out sample. From Tabula Sapiens, we took the ground truth information of cell-types in this separately held-out sample and scored the models for their predictive accuracy. STANN outperformed the competing methods, suggesting the model's generalizability. We hope these additional analyses alleviate the reviewer's concern.

- The authors explored the relevance of spatially variable gene regulatory networks implicated on defining a given cell-type and their role on their corresponding intercellular communication. Globally speaking this concept is of major interest, thus counting with strategies to reveal such spatial GRNs are more than welcome. This being said, this manuscript might gain on relevance if the authors could reveal the major gene co-regulatory network per cell types retrieved on each of the FOVs and their commonalities issued of their inter-cellular communication.

We thank the reviewer for this excellent suggestion. In our revised manuscript and Supplementary Figure S8, we have now discussed the idea of the major gene regulatory networks (GRNs) per cell-type. The relevant text reads as follows.

"The above analyses not only revealed spatially variable GRNs regulating receptors and ligands and how variation in receptor-ligand usage could refine cell-subtypes, it also enabled us to identify the spatially consistent up-stream regulators of cell-type specific marker genes (Fig. S8). Consistent with the literature, we found that certain upstream regulators are key for overall cell-type specific functionality. For example, we found Rorb is an upstream regulator of astrocytes' marker genes across all FOVs. This is consistent with previous studies reporting a major role for Rorb in astrocyte maturation (Clarke et al. 2021). Similarly, we found that Sox10 regulates the olfactory ensheathing cell marker genes, as was previously reported (Barraud et al. 2013); and Larp1 regulates the markers of neuronal granule cells -- Larp1 has been associated previously with neuronal proliferation and differentiation (Gower-Winter et al. 2013)."

We hope the reviewer finds our new analysis and the discussion useful.