nature portfolio

Peer Review File

ST-GEARS: Advancing 3D Downstream Research through Accurate Spatial Information Recovery

Reviewer #1 (Remarks to the Author):

The paper introduces ST-GEARS, a computational method for recovering spatial profiles in 3D Spatial Transcriptomics. It focuses on optimizing 'anchors' for aligning sections based on gene expression and structural similarity, employing Distributive Constraints for precision. ST-GEARS aligns sections, corrects distortions through Elastic Fields, and uses Gaussian Denoising for improved accuracy. Tested across various datasets, it outperforms existing methods in spatial information recovery. The study suggests enhancements in data preprocessing and efficient clustering. The paper was poorly written, and the motivation for the proposed method was not well justified. My major concerns are:

1. The method part (including Figure 1) is a disaster. The notations are not consistent and are poorly defined. My suggestion is to use clearer, simpler language to define the data, parameters, and model. Also, the authors need to break down the framework into a step-by-step process, sometimes incorporating diagrams or flowcharts to represent the methodological process, which can be particularly helpful for understanding the method.

2. Lack of novelty. 3D reconstruction is a well-explored topic for spatially resolved transcriptomics data, with numerous existing methods available in top journals as well as bioRxiv. Although the manuscript has highlighted some distinctive features of the proposed method, including formulating the problem using fused Gromov-Wasserstein (FGW) Optimal Transport (OT), elastic registration, Gaussian smoothing, and so on, the authors did not justify why these features are essential for the accurate 3D reconstruction. Furthermore, the proposed method is an ensemble of components existing in existing methods. While we know that complexity does not necessarily lead to a better method, the authors should provide a clear justification for the proposed method and the necessity of every component added to the method compared to existing methods.

3. Beyond comparing with existing methods, include a broader range of datasets, particularly those with complex tissue types or pathological samples, to demonstrate the versatility and robustness of ST-GEARS. In addition, it provides a more in-depth comparison with existing methods, including detailed case studies where ST-GEARS offers significant advantages or addresses specific challenges not met by other approaches.

4. Many spatial transcriptomics tools integrate the paired histology images. However, this method ignores this very important information, which will definitely limit its applicability.

5. Scalability issue. Given the well-known scalability issue, the proposed method is no exception. If the scalability issue is not addressed, the method will not be a useful tool for spatial transcriptomic data analysis. Explore the potential for computational optimization to reduce processing time, especially for large datasets. No computational cost was reported.

6. Real data application. This manuscript has limited results from real data analysis and all the figures are in low-resolution and poorly generated. Additionaly, the authors should elaborate on the potential long-term implications of ST-GEARS for the fields of developmental biology, oncology, and tissue engineering, including how it might influence future research directions or clinical applications.

Reviewer #2 (Remarks to the Author):

In this article, the authors proposed a novel three-dimensional tissue recovery approach ST-GEARS for spatial transcriptomic data. This approach can use multiple tissue sections to reconstruct their original three-dimensional morphology. ST-GEARS adopts fused Gromov-Wasserstein optimal transport scheme with innovative distributive constraints to enhance the anchor retrieval, and it sequentially performs rigid and elastic registrations under the guidance of anchors to achieve the section alignment and deformation correction. The authors also proved the validity of bi-sectional fields in eliminating distortions of sections. Diverse real applications not only exhibit the effectiveness of introducing distributive constraints in the anchor retrieval and employing the elastic registration, but also highlight the overall satisfactory reconstruction performance of ST-

GEARS. In my opinion, this paper addresses an important biological question in the field of spatial transcriptomic 3D recovery, but I have some comments that require more explanations or discussions by the authors.

Major comments:

1. In the method part, the authors mention that each spot has its cell type. For example, on line 653, "spots with same cell types". It is possible that one spot may have several cells from different types, so I am wondering how to determine the cell type annotation for such spots with heterogeneous cells?

2. The inputs of ST-GEARS include mRNA expression, spatial coordinates, and approximate grouping information, with the latter explicitly specified as rough clustering or annotation. For datasets lacking biological annotations, what degree of "roughness" is acceptable? Additionally, how significantly does the clustering result impacts the final 3D recovery performance? The authors may incorporate some numerical experiments to explain these aspects.

3. The distributive constraints are not applied to the mouse brain dataset due to the vast variations in cell types across sections. Does it imply that distributive constraints can only be employed when cell types across sections are almost identical? Please provide some guidance for users in which cases we should use the distributive constraints.

4. Excess zeros could be observed in next-generation-sequencing-based ST data, but the ST-GEARS approach does not consider this explicitly. Do the zero proportions influence the 3D recovery result of ST-GEARS?

Minor comments:

1. Line 48: "Visum" should be corrected to "Visium".

2. Mathematical notations should be consistent throughout the manuscript. For example, Line 589 introduces $X_A \in \mathbb{R}^{n_A,2}$, whereas Line 603 presents $X_{i,:}^{(A)}$.

3. Some notations are repeatedly used. For example, ``W" in Lines 597, 638, and 686 have distinct meanings. And in Lines 601 and 618, the authors use C_A with different meanings.

4. In Figure 4 panel a, "S-GEARS (rigid result)" should be "ST-GEARS (rigid result)"?

Reviewer #3 (Remarks to the Author):

The manuscript titled "ST-GEARS: Advancing 3D Downstream Research through Accurate Spatial Information Recovery" tackles the challenge of accurately reconstructing the 3D morphology of tissue sections from their in situ spatial transcriptomics data. Current approaches to 3D spatial reconstruction suffer from significant inaccuracies due to their failure to account for experiment-induced distortions or their sole reliance on gene expression data without incorporating structural information. This results in discrepancies between reconstructed and actual in vivo cell locations, affecting downstream analyses. ST-GEARS introduces an innovative approach that utilizes optimized anchors between sections based on both expression and structural similarities. It incorporates Distributive Constraints into the optimization process, enhancing the precision of anchor retrieval. The method employs elastic fields for distortion correction and Gaussian Denoising for data quality improvement, significantly advancing the accuracy of spatial information recovery. By providing a more accurate method for reconstructing the 3D spatial profiles of tissue sections, ST-GEARS enables a deeper understanding of biological processes at the tissue, cell, and gene levels. Its ability to precisely recover spatial information supports more reliable downstream analyses, potentially unlocking new insights in developmental biology, organogenesis, and disease pathology, and fueling biological discoveries.

The method employs elastic fields within the Fused Gromov-Wasserstein (FGW) framework to

correct experimental distortions by mathematically modeling the deformation that tissue sections undergo during experimental procedures. Elastic fields are used to represent how each point in the tissue is displaced or transformed, allowing for the adjustment of the spatial coordinates of gene expression data. This process involves calculating the optimal transformation that minimizes the difference between the distorted experimental data and the expected undistorted state, effectively 'undoing' the distortions and aligning the data more accurately with its original, undistorted configuration. This step is critical for ensuring that the reconstructed 3D spatial information accurately reflects the true morphology of the tissue.

I found the problem the authors tackle is very challenging and has a profound impact on our understanding of tissue 3D structure and cellular environment. The method discussed in this study presents a comprehensive strategy that aligns tissue slices by addressing limitations and gaps that were not resolved by existing approaches. Generally, I feel this is an important and useful methodology for the community. Yet, I have several major concerns that prevent me from recommending the paper in its current form (See below).

Major concerns:

1) The manuscript's benchmarking framework, while inclusive of comparisons with GPSA, PASTE, and PASTE2, can be significantly enhanced by integrating STAlign and SLAT into the comparative analysis. The addition of STAlign, renowned for its precision in slice-to-slice spatial alignments, would provide a critical evaluation of ST-GEARS in terms of alignment accuracy and efficiency. Furthermore, although SLAT does not directly offer 3D reconstruction solutions, its inclusion could provide valuable insights into pairwise slice alignment capabilities. This broader benchmarking spectrum is essential for a comprehensive assessment, offering a clearer picture of ST-GEARS's technological advancements and its comparative effectiveness within the rapidly evolving field of spatial transcriptomics. Expanding the benchmarking to include these methods would not only highlight ST-GEARS's unique contributions but also help identify areas for further methodological refinement and development, ensuring its competitive edge and utility in addressing complex biological questions. (Also why GPSA benchmarking is missing for several sets in the study).

2) The authors' efforts in demonstrating ST-GEARS' performance across multiple real datasets are commendable, showcasing its practical application and robustness. However, the inherent limitation of ground truth in these datasets poses a challenge for systematic benchmarking. To address this, a recommendation for further strengthening the manuscript is to include benchmarking against simulated datasets. By artificially manipulating slices through rotation, scaling, cropping, and adding noise, the authors could generate controlled conditions to rigorously test and quantitatively compare ST-GEARS' performance. This approach would allow for a more precise evaluation of its capabilities in handling various distortions and noise levels, providing a comprehensive benchmark that underscores its accuracy and efficiency in spatial reconstruction.

3) A critical weakness in ST-GEARS may lie in its computational complexity, particularly when processing large-scale datasets. The method's advanced features, such as optimized anchor alignment and elastic field application for distortion correction, could demand significant computational power and memory, impacting its efficiency. This aspect may limit its accessibility for researchers with limited computational resources or extend processing times for voluminous datasets. As the authors are undoubtedly aware, the volume of single-cell spatial transcriptomics data is ever-increasing, with datasets growing in scale and complexity. Therefore, it is imperative to ensure that the ST-GEARS algorithm can be efficiently applied to large-scale single-cell spatial datasets, as this is crucial for its broader adoption and practical utility in cutting-edge research. To address this concern and ensure the method's scalability, we kindly request that the authors provide a comprehensive analysis of both time and memory complexity in their manuscript. Such an analysis would not only serve as a testament to the method's computational efficiency but also provide valuable insights for researchers who may be considering its application on large-scale datasets. By presenting a detailed breakdown of time and memory requirements, the authors can demonstrate the method's ability to handle substantial datasets without compromising performance.

4) While the manuscript does provide evidence of ST-GEARS' application to various tissue types,

another significant weakness that should be addressed relates to the potential for overfitting or hyperparameter sensitivity. The method incorporates multiple complex steps, including anchor selection, distributive constraints, Gaussian denoising, and elastic field modeling, each involving specific parameter choices. A potential weakness lies in the possibility that the performance of ST-GEARS is highly dependent on the fine-tuning of these parameters. If the method is sensitive to the choice of parameters, it could lead to overfitting on certain datasets or challenges in reproducibility across different research groups. To mitigate this concern, it would be beneficial for the authors to provide a comprehensive sensitivity analysis that explores how variations in parameter settings impact the results. Additionally, recommendations or guidelines for parameter selection, based on the authors' extensive experience with the method, would aid users in achieving optimal outcomes. By addressing this weakness and offering insights into the robustness of ST-GEARS with respect to parameter choices, the authors can enhance the method's usability and reliability, ensuring that it can be successfully applied by a wider range of researchers without the risk of unintended biases or overfitting issues.

Reviewer #3 (Remarks on code availability):

pros: easy installation and comes with an example jupyte note with test datasets.

cons: no detailed description of APIs (functions and modules, what are the functions of each method and their parameters)

Reviewer 1

The paper introduces ST-GEARS, a computational method for recovering spatial profiles in 3D Spatial Transcriptomics. It focuses on optimizing 'anchors' for aligning sections based on gene expression and structural similarity, employing Distributive Constraints for precision. ST-GEARS aligns sections, corrects distortions through Elastic Fields, and uses Gaussian Denoising for improved accuracy. Tested across various datasets, it outperforms existing methods in spatial information recovery. The study suggests enhancements in data preprocessing and efficient clustering. The paper was poorly written, and the motivation for the proposed method was not well justified. My major concerns are:

We thank the reviewer for the careful observation and insightful suggestions. We have conducted thorough studies based on your questions and have revised corresponding sessions of our manuscript such as methods and innovation studies in red color. We hope this edition and the reply below will address your concerns.

1. The method part (including Figure 1) is a disaster. The notations are not consistent and are poorly defined. My suggestion is to use clearer, simpler language to define the data, parameters, and model. Also, the authors need to break down the framework into a step-by-step process, sometimes incorporating diagrams or flowcharts to represent the methodological process, which can be particularly helpful for understanding the method.

Thanks for your careful reading and insightful comments. We noticed the consistency problem of notations and have revised the notations to be consistent and clearly defined. We have also modified our languages on definition of data, parameters, and model to be simpler and clearer. We appreciate your advice of breaking down the framework into step-by-step process and have revised our Figure 1 based on incorporating a diagram to represent the process.

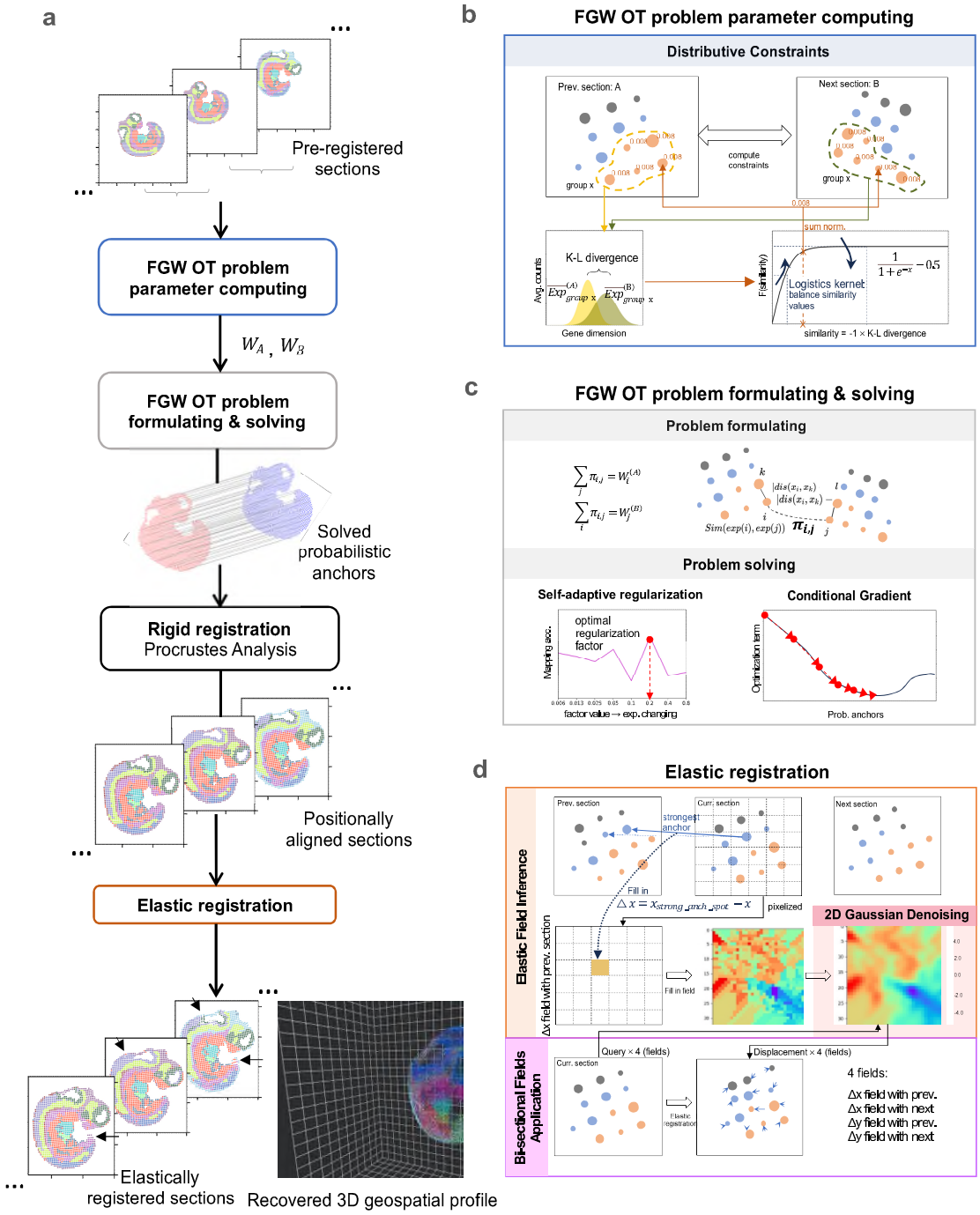


Figure 1: **Three-Dimensional (3D) Spatial Transcriptomics (ST) Geospatial profile recovery with ST-GEARS**. (a) The automatic pipeline of ST-GEARS which recovers ST-GEARS 3D *in vivo* spatial information by ordered steps including Fused Gromov Wasserstein (FGW) Optimal Transport (OT) problem parameter computing, problem formulating and solving which outputs probabilistic anchors across sections, rigid registration through Procrustes Analysis which solves optimal positional alignment using the anchors, and finally elastic registration. The input of the method is Unique molecular identifier (UMI) counts and location of each spot measured by ST technology, along with their annotations or cross-section clustering result. And the output of the method is recovered 3D in vivo spatial information of the experimented tissue, or sample. (b) FGW OT problem parameter computing, which assigns nonuniform weights to spots in preparation for future problem formulating, based on cross-sectional

similarity of annotation types or clusters. (c) FGW OT problem formulating, whose setting aims to solve probabilistic anchors joining spots with highest *in situ* proximity, through optimizing the combination of gene expression and structural similarity**Error! Reference source not found.**. FGW OT problem solving, which is implemented based on Conditional Gradient (CG) method, leading to retrieved probabilistic anchors. (d) Elastic registration, which utilizes the anchors again to compute and denoise distortion fields which guides the elimination of distortions, then applies the fields bi-sectionally to positionally aligned sections, leading to the recovered 3D *in vivo* spatial information.

2. Lack of novelty. 3D reconstruction is a well-explored topic for spatially resolved transcriptomics data, with numerous existing methods available in top journals as well as bioRxiv. Although the manuscript has highlighted some distinctive features of the proposed method, including formulating the problem using fused Gromov-Wasserstein (FGW) Optimal Transport (OT), elastic registration, Gaussian smoothing, and so on, the authors did not justify why these features are essential for the accurate 3D reconstruction. Furthermore, the proposed method is an ensemble of components existing in existing methods. While we know that complexity does not necessarily lead to a better method, the authors should provide a clear justification for the proposed method and the necessity of every component added to the method compared to existing methods.

Thanks for your observations and your comments on our method. We dissemble 3D spatial profile recovery into multiple missions that are connected head to tail, hence ST-GEARS is an ensemble of multiple method components. In some mission such as anchors computing, we have referred to current approaches such as PASTE in the optimization solver; however, to largely enhance anchors accuracy, we introduced Distributive Constraints for the first time. This innovation utilizes cell type component information to assign different weight to cells in anchors computation, hence increases mapping accuracy and the final registration accuracy.

Besides the above components, some modules are completely innovated by ST-GEARS such as elastic registration module. We apologize for not explaining this relationship in a clear enough manner. Elastic registration minimizes differences between distorted experimental data and the expected undistorted state, hence ‘undo’ the distortions. It is implemented based on the innovative design and combination of three operations including elastic field inference, 2D Gaussian denoising and bi-sectionally fields application. The module enables ST-GEARS to be the first fully automatic registration method that can correct distortion.

In the section of Enhancement of anchor retrieval accuracy through Distributive Constraints, and section of Recovery of in situ shape profile through elastic registration, we respectively provide the necessity justification of Distributive Constraints and Elastic registration. Large improvements can be witnessed by comparing results of ST-GEARS to the same method with either component dis-included while other operations unchanged (Fig. 2, Fig. 3, Supplementary Fig. 1, Supplementary Fig. 6).

In application cases, the comparison between including and dis-including Elastic Registration is conducted as well. Obvious improvements can be seen by results of ST-GEARS adopting Elastic

Registration (Supplementary Fig. 13, Supplementary Fig. 16, Supplementary Fig. 17, Supplementary Fig. 18).

Based on your suggestion, we further conducted several studies to dis-include and include Distributive Constraints. In Mouse hippocampus application, increasement of mapping accuracy can be seen on each section pairs of by adopting Distributive Constraints compared to not adopting it (Fig. R1).

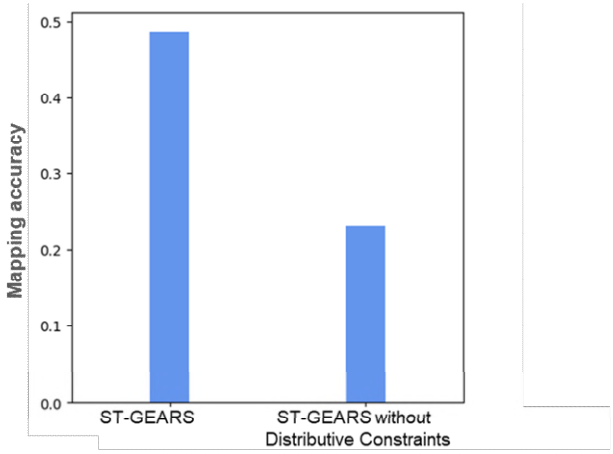


Figure R1: **Distributive Constraints in ST-GEARS enhances mapping accuracy of Mouse hippocampus.** Shown here is the comparison of mapping accuracy of ST-GEARS result, and result of ST-GEARS with Distributive Constraints (DC) excluded, when registering Mouse hippocampus data. The accuracy index is higher upon DC adopted.

On Drosophila embryo, adopting Distributive Constraints enhances mapping accuracy across all section pairs, as well (Fig. R2). For the final registration result, the experimental flaw on the 15th section is witnessed to be fixed by incorporating Distributive Constraints, while the flaw is left unfixed when dis-including Distributive Constraints (Fig. 5c, Fig. R3).

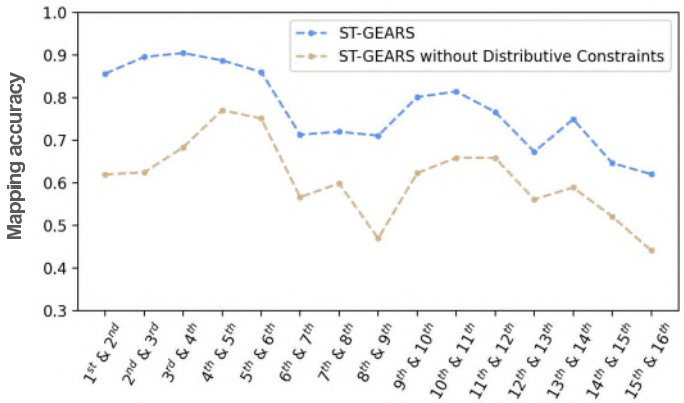


Figure R2: **Mapping accuracy of ST-GEARS with and without Distributive Constraints on registration of Drosophila embryo.** The accuracy index is higher upon DC adopted.

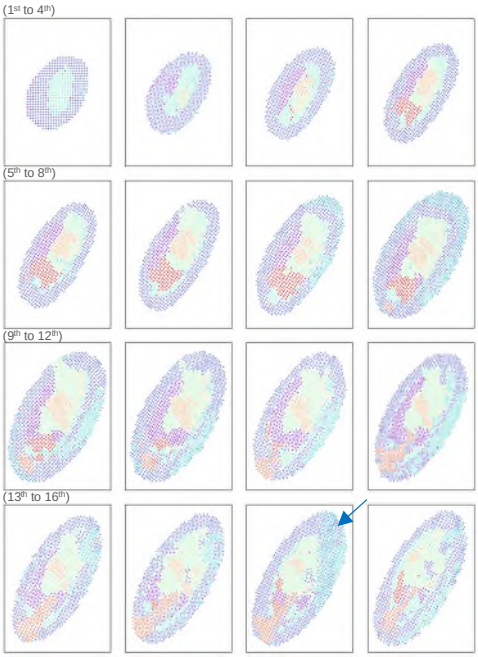


Figure R3: **Individual sections of Drosophila embryo generated by ST-GEARS with Distributive Constraints excluded.** The blue arrow points to section area where experimental flaw remains unfixed by the method.

We have included the above studies as modifications of our manuscript to clarify the novelty of our method, specifically in section of Application to sagittal sections of Mouse hippocampus, and section of Application to 3D reconstruction of Drosophila embryo. We hope the above reply and modifications can address your comments and questions.

3. Beyond comparing with existing methods, include a broader range of datasets, particularly those with complex tissue types or pathological samples, to demonstrate the versatility and robustness of ST-GEARS. In addition, it provides a more in-depth comparison with existing methods, including detailed case studies where ST-GEARS offers significant advantages or addresses specific challenges not met by other approaches.

We appreciate your observations and the suggestion, and have included pathological datasets to understand ST-GEARS' effect on pathological datasets.

We applied PASTE, PASTE2, GPSA, STalign and ST-GEARS onto the registration of Squamous cell carcinoma (SCC) sections (Fig. R4). By GPSA, various spots on different locations share exactly same coordinates, characterizing a failed registration. Across the other 4 methods, the edge of sections corresponds best by ST-GEARS (Fig. R5), indicating that our method addresses the distortion correction challenge best across all methods. In quantification of the comparison, across all methods generating anchors, ST-GEARS achieved significantly higher mapping accuracy than PASTE and PASTE2 (Fig. R6). Above results complement the applicability and advantages of ST-GEARS through diverse sample types using pathological data.

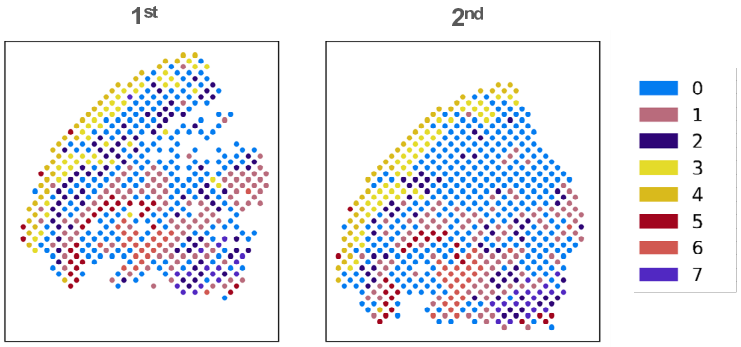


Figure R4: **Squamous cell carcinoma (SCC) sections before registration.**

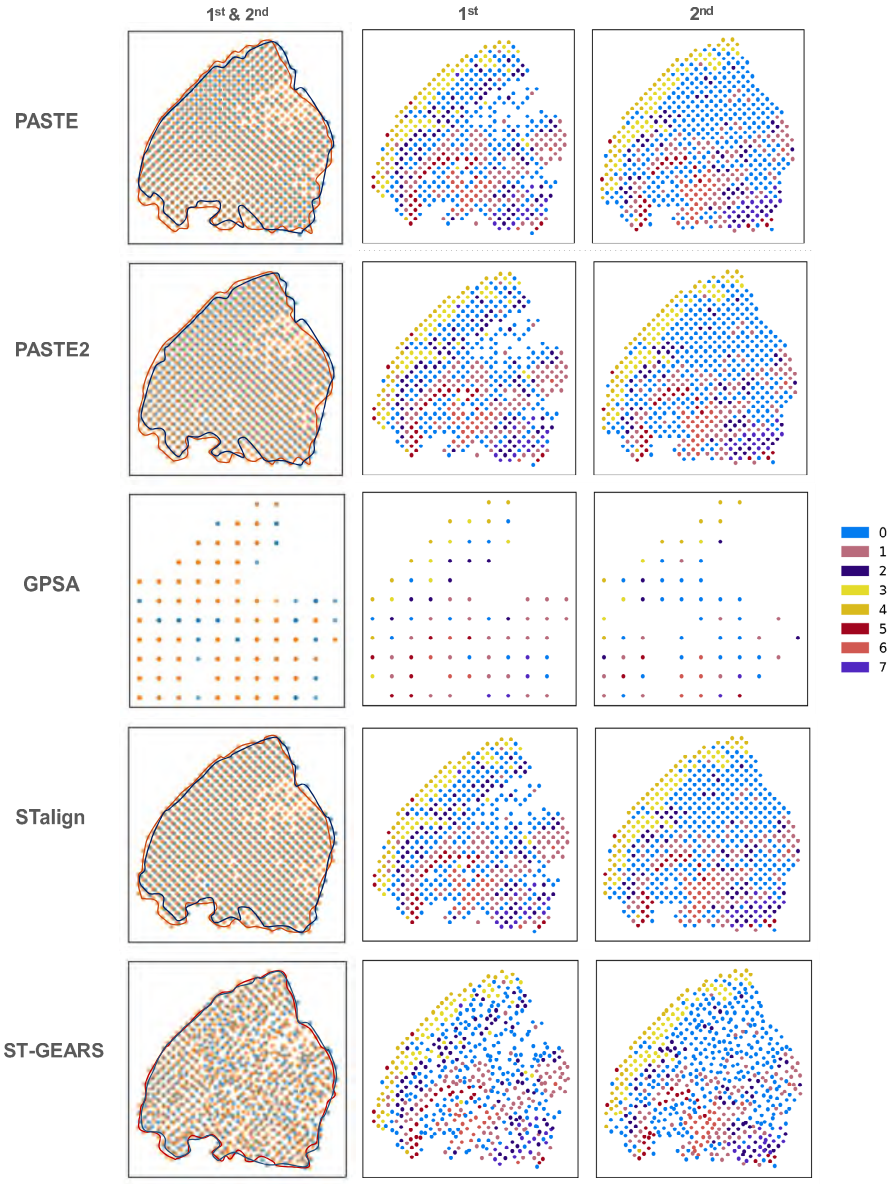


Figure R5: **Registration result of PASTE, PASTE2, GPSA, STalign and ST-GEARS on Squamous cell carcinoma (SCC) data.** Presented from top to bottom are results by PASTE, PASTE2, GPSA, STalign and ST-GEARS. The 1st column shows the overlapping of registered sections, with blue scatters

representing bin spots of the 1st section, and orange spots representing the 2nd section. Edge of the 1st section indicated by spots positions was highlighted by blue line, while edge of the 2nd was highlighted by line of orange. The 2nd and the 3rd column visualize registered results of the 1st and the 2nd section, respectively. Various spots share the exact same position by registration result of GPSA. Among all methods except from GPSA with obvious error, edge of sections corresponds best by ST-GEARS.

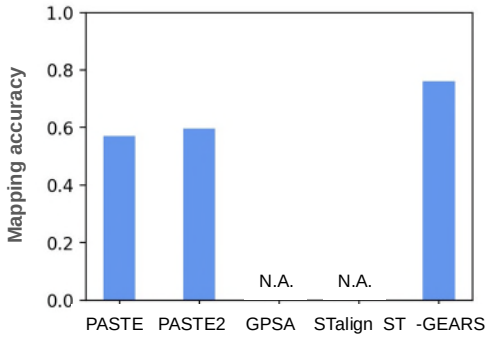


Figure R6: **Mapping accuracy of PASTE, PASTE2, GPSA, STalign, and ST-GEARS on registration of Squamous cell carcinoma (SCC) data.** The index is not applicable on GPSA and STalign due to their lack of mapping information in results. Among methods with mapping results, ST-GEARS achieves the highest accuracy.

4. Many spatial transcriptomics tools integrate the paired histology images. However, this method ignores this very important information, which will definitely limit its applicability.

Thanks for the observation and comments on histology image integration. When designing ST-GEARS, we did consider the possibility of enhancing registration through incorporating histology images. Hence, to explicitly study if integrating histology image into the method can enhance our registration precision, we compared ST-GEARS with an attempt of involving the image information into the method. Specifically, we first manually registered ST data of human dorsolateral prefrontal cortex (DLPFC) to its Hematoxylin (HE) histology image (Fig. R7), then integrated similarity term between histology images into registration of ST-GEARS, as the same approach adopted by PASTE2. By integrating and not integrating histology information, the mapping accuracy remains the same up to the 3rd decimal place across all section pairs (Fig. R8). Based on the in-variance of accuracy by integrating histology image or not, the image information has not been involved by method of ST-GEARS. The result is probably because context information of images introduced by cell variance such as cell size variance and cell shape variance are already embedded in gene expression.

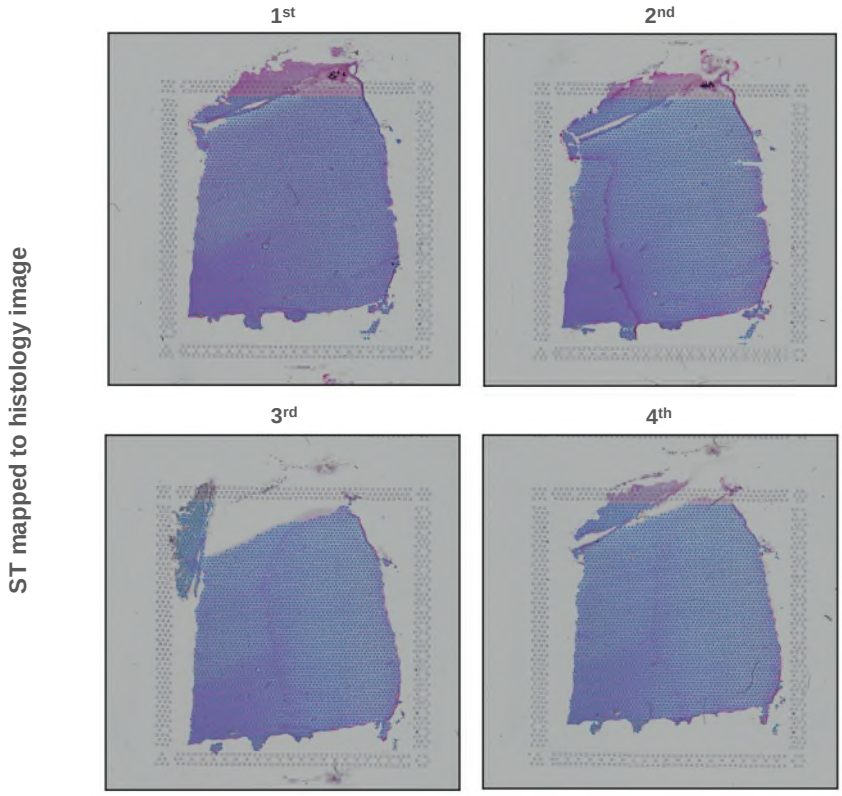


Figure R7: **Spatial Transcriptomics spots of Dorsolateral Prefrontal Cortex (DLPFC) mapped to corresponding Hematoxylin (HE) histology image.** The 1st row shows the mapped data of 1st and 2nd sections, and the 2nd row shows the 3rd and 4th sections.

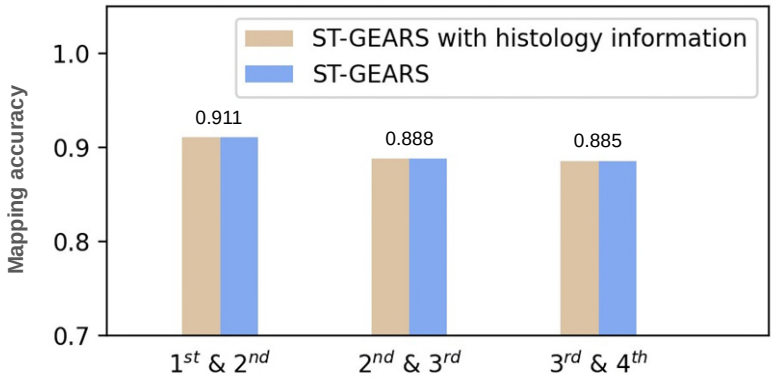


Figure R8: **Mapping accuracy of ST-GEARS with and without histology information integrated when registering Dorsolateral Prefrontal Cortex (DLPFC) data.** Up to the third decimal place, the accuracy index remains the same between options on each pair of sections.

In the practical perspective, different types of histology images are adopted across different types of tissues in ST studies, hence a universal registration framework across different histology types

remains a challenge to be established. For example, Single-Stranded DNA (ssDNA) images are the usually adopted histology modality in brain ST experiments involving histology while Hematoxylin (HE) images are usually preferred in study of tumors. Furthermore, the huge modality difference between histology images and ST dataset itself poses challenge for methods to automatically integrate and register without introducing manual labor. Hence, we would like to extend this part of research in our future work.

5. Scalability issue. Given the well-known scalability issue, the proposed method is no exception. If the scalability issue is not addressed, the method will not be a useful tool for spatial transcriptomic data analysis. Explore the potential for computational optimization to reduce processing time, especially for large datasets. No computational cost was reported.

Thanks for reminding and for your suggestions. We agree on the importance of scalability for methods to be adopted across diverse scenarios, especially across data sizes. Across different methods in the field of 3D spatial profile recovery, scalability has been an issue not well addressed. Confronting exceedingly large dataset, OT-based methods such as PASTE suffer from the exponential memory and time consumption in constructing adjacency matrix describing gene expression and structural similarity. PASTE2 exacerbates the consumption by attempting different regularization factors while saving all intermediate results in memory. STalign transforms ST registration to registering image with large number of iterative steps, which introduces challenges on both time and memory consumption. Adopting deep Gaussian process, GPSA also solves registration result iteratively which requires large amount of time and requires extensive memory in its network construction.

To deal with scenario with large data size, we introduce Granularity adjusting as a computational optimization to assist ST-GEARS. And we recommend users to turn on this option when over 3000 spots are present in each section. In granularity adjusting, section area is first binned, with spots squared by each pixel summarized into one single spot, leading to a ST data with coarser resolution than original data. When summarizing within each grid, Unique molecular identifier (UMI) counts of spots is summed to, and the most frequent annotation type or cluster is labelled to the generated one spot. Then ST-GEARS is applied onto the coarser version of data, outputting a registered dataset with coarse resolution. Finally, to recover the original resolution as registration result, the original resolution data can be interpolated into the pre-registered and registered coarse dataset, leading to registration result in original resolution (Fig. R9). The conduction code of binning and interpolation method has been updated to GitHub repository of ST-GEARS.

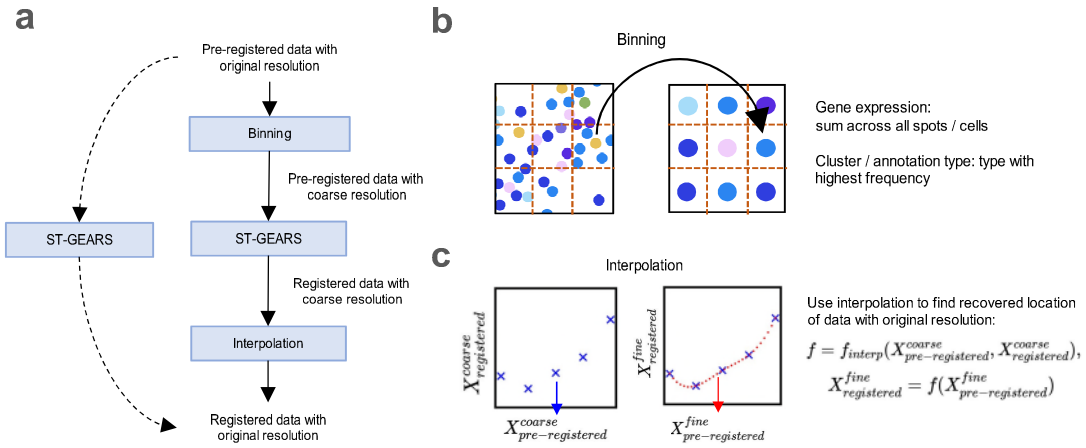


Figure R9: **Granularity adjusting option recommended for datasets with over 3000 spots.** (a) The automatic process of granularity adjusting includes binning the section area, which leads to bin sets that are much less compared to original spots, running the bin sets data with ST-GEARS, leading to registered data with the coarse resolution, and eventually interpolating original resolution data into the coarse one, outputting the registered data with original resolution. (b) In binning step, section area is gridded, with spots squared by each pixel summarized into one single spot. (c) In interpolating step, registered data with original resolution is solved by interpolating the pre-registered original resolution data into pre-registered (output by binning step) and post-registered (output by ST-GEARS step) coarse resolution data.

The strategy enables higher computational efficiency in both anchors computation and geospatial correction, without compromising accuracy of the reconstructed result. For example, we applied granularity adjusting on Mouse hippocampus dataset and ran ST-GEARS on original dataset as well as binned dataset adopting bin size of respectively 30 and 40 μm. Both time and memory by ST-GEARS are much lower on binned dataset than on original resolution (Fig. R10). Lower computational cost was achieved by higher bin size. By comparing registration result through granularity adjusting and direct registration, the coefficient of determination (R^2) remains over 0.98 across all coordinates and sections, on both bin sizes (Fig. R11, Fig. R12), indicating that accuracy is not compromised by granularity adjusting option.

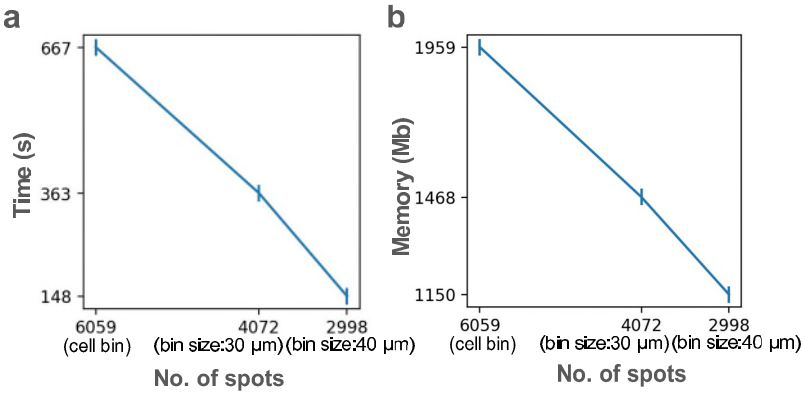


Figure R10: **Time and peak memory consumption by ST-GEARS, and number of spots of different resolutions of Mouse hippocampus.** The different resolution data includes dataset in original resolution

of cell bin, the binned data with bin size of 30 μm and the binned data with bin size of 40 μm.

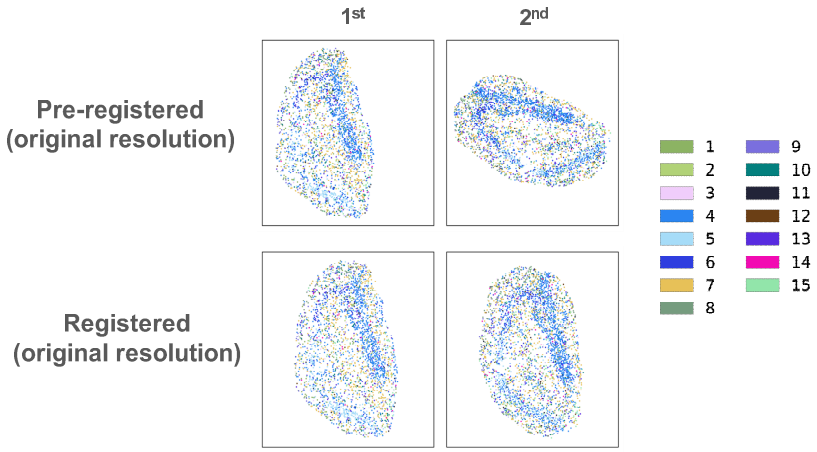


Figure R11: **Pre-registered and post-registered Mouse hippocampus dataset with original resolution.** The 1st row shows pre-registered dataset, and the 2nd row shows registration result of ST-GEARS directly using original resolution, without granularity adjusting adopted. The 1st column represents the 1st section, while the 2nd column represents the 2nd.

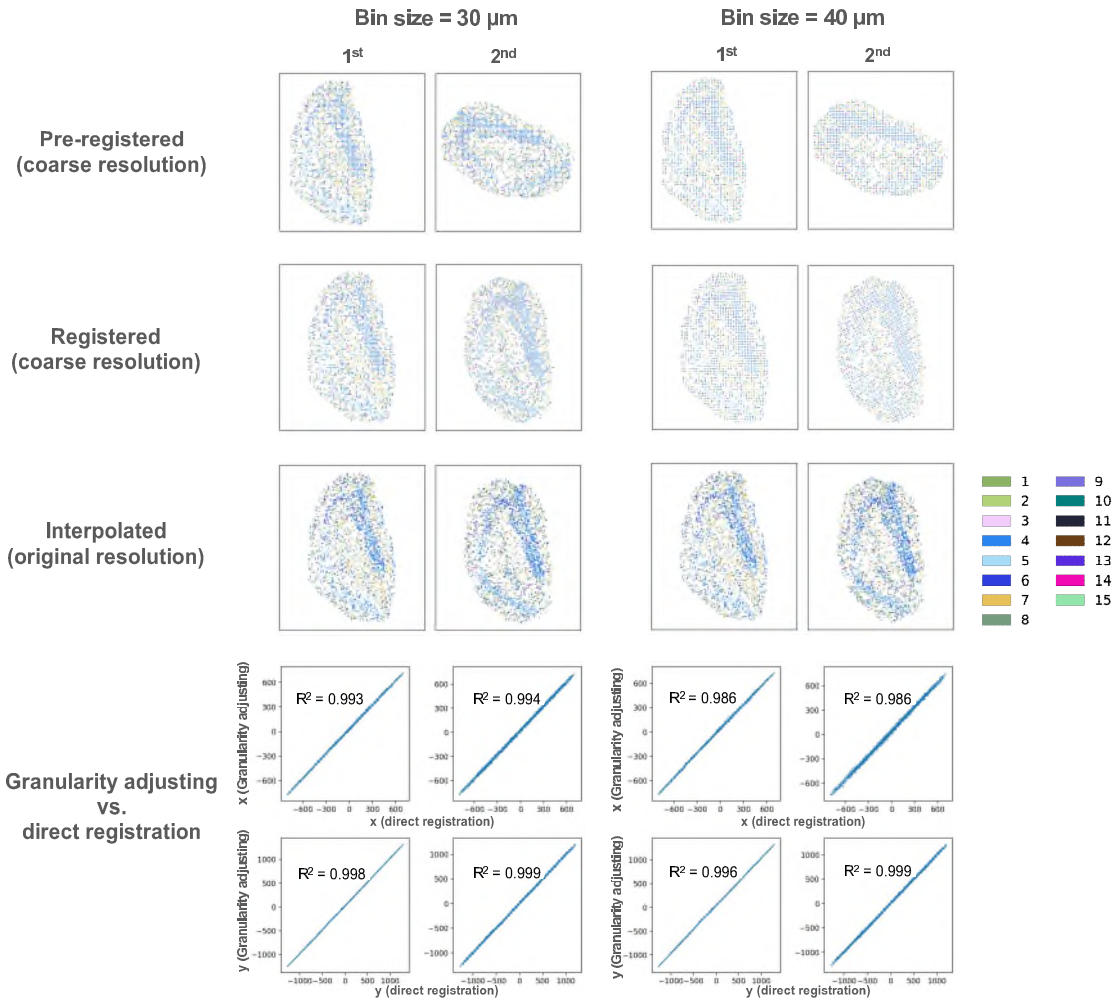


Figure R12: **Registration of Mouse hippocampus dataset with granularity adjusting strategy, in different bin sizes.** From top to bottom presented are binned yet pre-registered datasets, registered binned datasets, interpolation result of dataset with original resolution, and comparison between position of cells by process of granularity adjusting and direct registering through ST-GEARS. Two rows are included in the comparison plots, with the 1st row showing comparison of x coordinates and the 2nd row showing comparison of y coordinates. The 1st and 2nd columns are respectively results of 1st and 2nd sections by bin size of 30 μm in binning step of granularity adjusting, while the 3rd and 4th columns are result of the same two sections by bin size of 40 μm. In comparison between position of cells by process of granularity adjusting and direct registering, the coefficient of determination (R^2) remains over 0.98 across all coordinates, sections, on both bin sizes.

We also compared both time and memory cost of methods including PASTE, PASTE2, GPSA, STalign and ST-GEARS when registering identical datasets. ST-GEARS used least peak memory when registering Mouse brain (Fig. R13) and used second least memory on Mouse hippocampus and Drosophila embryo, while using almost the same memory as the method with highest ranking. In terms of time efficiency, ST-GEARS is among the top two methods on two out of the three applications, with over 10 times lower time cost than PASTE2.

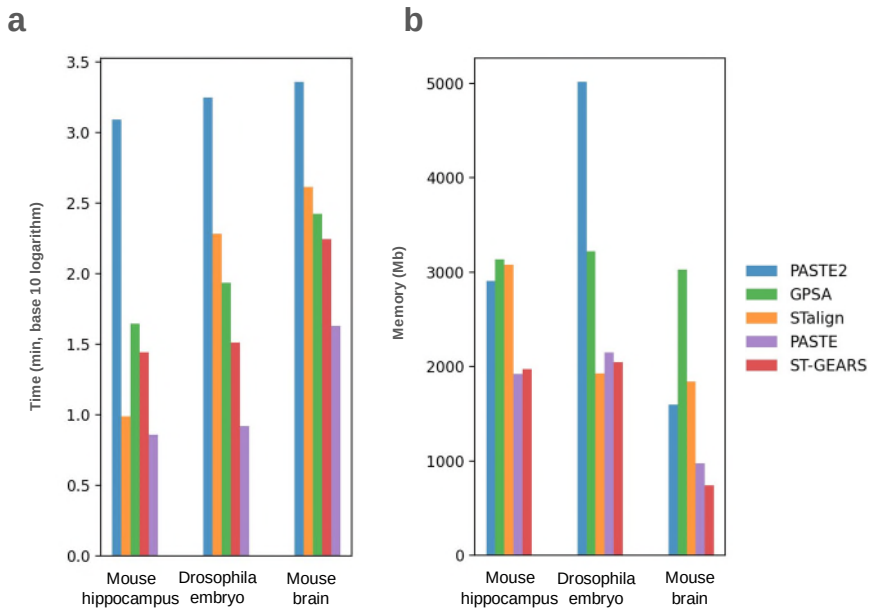


Figure R13: **Time and peak memory consumption of PASTE, PASTE2, GPSA, STalign and ST-GEARS, respectively on Mouse hippocampus, Drosophila embryo and Mouse brain datasets.**

We have included above explanations and studies of granularity adjusting as modification to section of Granularity Adjusting in our Supplementary materials. We have also added the scalability study of time and memory to application sections of our manuscript. We hope above explanations and studies address your questions and comments.

6. Real data application. This manuscript has limited results from real data analysis and all the figures are in low-resolution and poorly generated. Additionaly, the authors should elaborate on the potential long-term implications of ST-GEARS for the fields of developmental biology, oncology, and tissue engineering, including how it might influence future research directions or clinical applications.

Thanks for your comments on applications.

In terms of the resolution problem, we apologize for the inadequate resolution and have uploaded high-resolution figures along with our revised version.

With respect to real data analysis, we applied ST-GEARS on reconstruction of 3 datasets and analyzed gene and cells distribution by our method. On Mouse hippocampus, we found region-specific cell types including DG, Neurogenesis, subiculum, CA1, CA2 and CA3 have almost identical distribution on both sections after registration (Fig. 4c), indicating those cell types are accurately registered. On Drosophila embryo dataset, we found ST-GEARS produced most accurate location of midgut, without any extruding regions as produced by other methods (Fig. 5d). Marker gene Cpr56F and Osi7 show closest approximation to hybridization evidence by ST-GEARS among all methods (Fig. 5e). On Mouse brain application, 7 cell types within cortex layers including L2/3

IT, L4/5 IT, L5 ET, L5 IT, L6 CT, L6 IT, L6b are all correctly reconstructed (Fig. 6d).

Above findings indicate ST-GEARS produces accurate reconstructions of cell and gene, which is essential for revealing key variational regions in fields including developmental biology, oncology, and tissue engineering. Located in small areas, the regions precisely regulate process of developments and disease. Hence, by above results, ST-GEARS is shown to provide fundamental tool basis for accurate analysis of 3D ST data.

Reviewer 2

In this article, the authors proposed a novel three-dimensional tissue recovery approach ST-GEARS for spatial transcriptomic data. This approach can use multiple tissue sections to reconstruct their original three-dimensional morphology. ST-GEARS adopts fused Gromov-Wasserstein optimal transport scheme with innovative distributive constraints to enhance the anchor retrieval, and it sequentially performs rigid and elastic registrations under the guidance of anchors to achieve the section alignment and deformation correction. The authors also proved the validity of bi-sectional fields in eliminating distortions of sections. Diverse real applications not only exhibit the effectiveness of introducing distributive constraints in the anchor retrieval and employing the elastic registration, but also highlight the overall satisfactory reconstruction performance of ST-GEARS. In my opinion, this paper addresses an important biological question in the field of spatial transcriptomic 3D recovery, but I have some comments that require more explanations or discussions by the authors.

We thank the reviewer for the positive feedback and constructive critiques. We have significantly improved ST-GEARS based on your professional comments and suggestion. All significant modifications are marked in red color in the revised manuscript. We hope this edition will address your concerns.

1. In the method part, the authors mention that each spot has its cell type. For example, on line 653, "spots with same cell types". It is possible that one spot may have several cells from different types, so I am wondering how to determine the cell type annotation for such spots with heterogeneous cells?

Thank you for the careful observation and for bringing up the issue.

We apologize for using the phrase ‘cell type’ incorrectly. In all related cases we were trying to express ‘annotation type or clustering result of the spot’. Since various resolutions are introduced by different ST techniques, spots consisting of heterogeneous cells are sometimes present in the datasets. When dealing with such datasets, it is still the clustering result or annotation type of each spot that is required by ST-GEARS. We have revised this phrase across our Manuscript and Supplementary materials.

The annotation type or clustering information are used in Distributive Constraints section of ST-GEARS. Using the grouping information, ST-GEARS assigns different weight to spots or cells based on their similarity of gene expression hence identity. For Spatial Transcriptomics (ST) sections with annotation type or clustering information, once similarity information across spots or cells is embedded in the annotation type or clustering information, the dataset is acceptable input of ST-GEARS. For example, human dorsolateral prefrontal cortex (DLPFC) data with annotated tissue types (Fig. 2) and Mouse hippocampus data (Supplementary Fig. 11) with clustering information were both successfully registered by ST-GEARS.

2. The inputs of ST-GEARS include mRNA expression, spatial coordinates, and approximate grouping information, with the latter explicitly specified as rough clustering or annotation. For datasets lacking biological annotations, what degree of "roughness" is acceptable? Additionally, how significantly does the clustering result impacts the final 3D recovery performance? The authors may incorporate some numerical experiments to explain these aspects.

Thanks for your insightful question and comments. For datasets with clustering information of each spot or cell, ST-GEARS has a relatively loose requirement of its roughness extent. As the cluster information assists Distributive Constraints module in assigning different weight to observations based on their similarity of gene expression and identity, once the similarity information is embedded in the clustering, the roughness level is acceptable.

To explicitly demonstrate the acceptance range of roughness and its impact on method performance, we applied ST-GEARS to register 3 different SCC sections with varying clustering numbers. We respectively applied 3, 6, 9, 15, 18 and 21 as number of clusters in K-means clustering (Fig. R14) and used the same published cluster labels to measure mapping accuracy. In the published version, the SCC sections have 12 clusters. We found ST-GEARS correctly registers all sections across different roughness level of clustering and the mapping accuracy remains almost constant (Fig. R15). This indicates that once the clustering information includes the similarity messages, the results are least impacted; and the clustering result almost doesn’t impact the final 3D recovery performance.

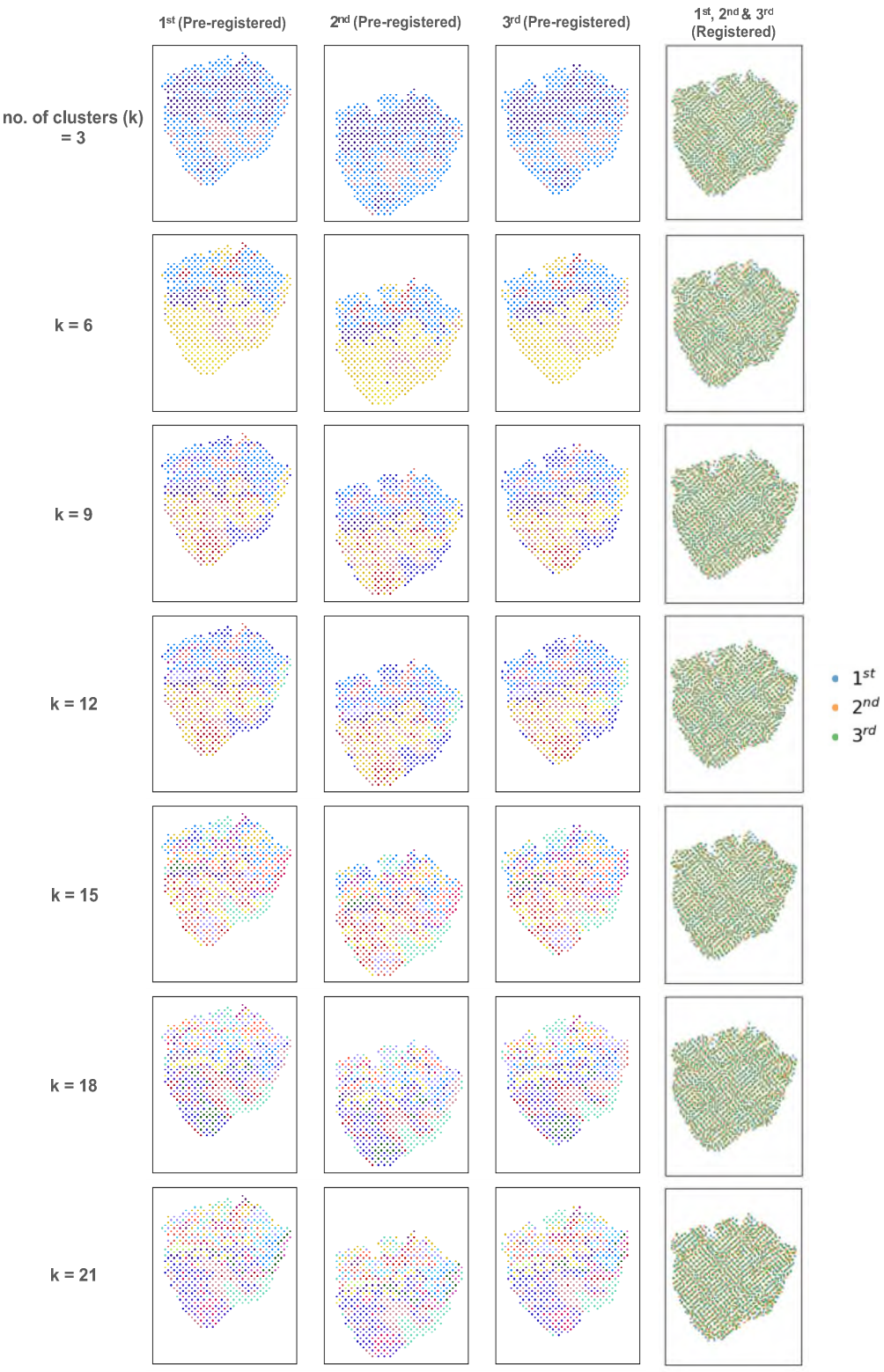


Figure R14: **Pre-registered and post-registered Squamous Cell Carcinoma (SCC) sections with different number of clusters as inputs.** From the top to bottom rows represented are sections with cluster number of 3, 6, 9, 12, 15, 18, 21. The 1st, 2nd and the 3rd column respectively shows clustering results of the 1st, 2nd and the 3rd section which are pre-registered, and the 4th column shows the overlapping of all three sections.

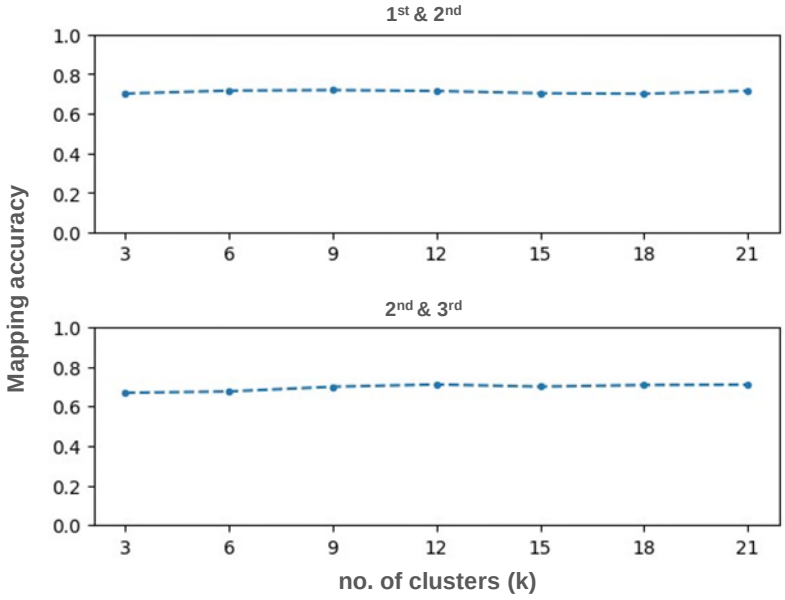


Figure R15: **Mapping accuracy of ST-GEARS on registering Squamous Cell Carcinoma (SCC) data with different number of clusters as inputs.** The 1st row shows the mapping accuracy of anchors between the 1st and 2nd sections, while the 2nd row represents accuracy of anchors between the 2nd and 3rd sections. The clustering was implemented by method of K-means. The mapping accuracy of data with different number of clusters is measured with the same published cluster labels.

3. The distributive constraints are not applied to the mouse brain dataset due to the vast variations in cell types across sections. Does it imply that distributive constraints can only be employed when cell types across sections are almost identical? Please provide some guidance for users in which cases we should use the distributive constraints.

We appreciate your observation and constructive suggestion.

In its anchors computation where Distributive Constraints is involved, ST-GEARS resolves anchors between pair of closest sections. For each section pair, sections are not required to have almost identical gene expression, or almost identical annotation type or cluster distribution. Certain variance is accepted on both spot / cell level and grouping level.

For guidance of using Distributive Constraints, we suggest users to calculate probabilistic distribution of number of spots on different clusters or annotations for each section (Fig. R16), then measure Kullback-Leibler (KL) divergence of the distribution between closest section pairs (Fig. R17). If the maximum KL divergence remains below 1, Distributive Constraints is suggested to be adopted, since the annotation type or cluster distribution is close between every section pair. However, if the maximum KL divergence exceeds 1, users are encouraged to try ST-GEARS without Distributive Constraints.

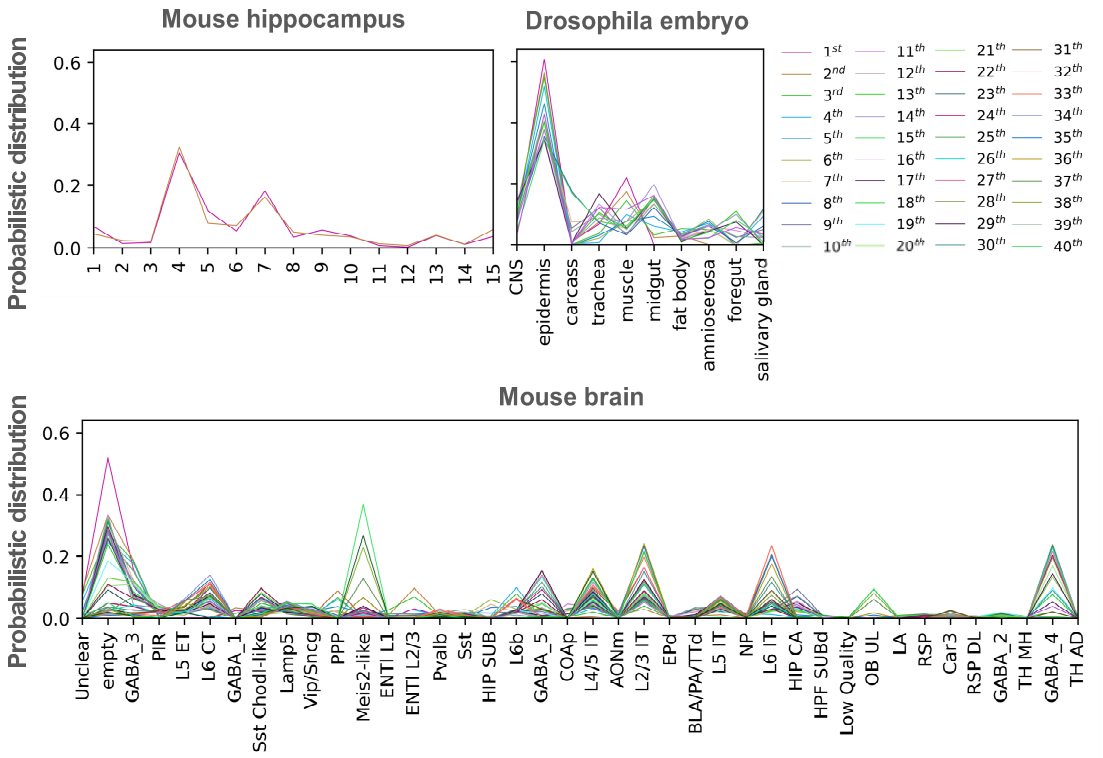


Figure R16: **Probabilistic distribution of proportion of spots on clustering / annotation types, of Mouse hippocampus, Drosophila embryo and Mouse brain datasets.** Each polyline was drawn based on distribution information of a section.

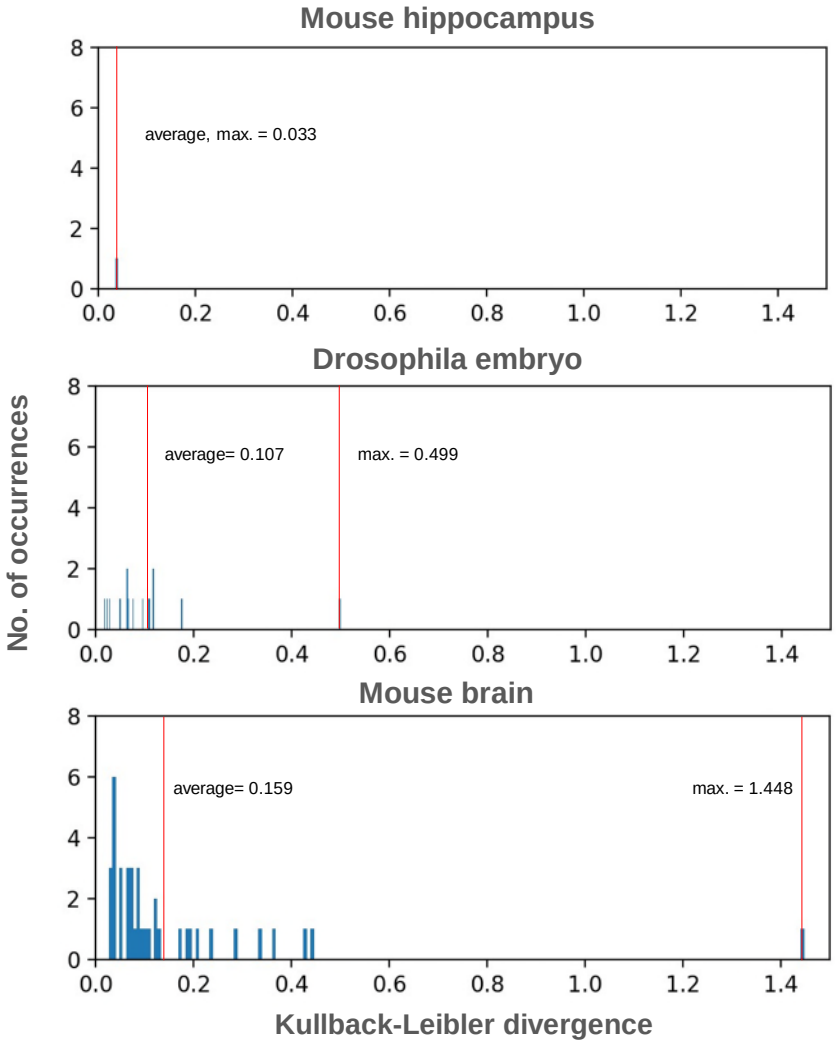


Figure R17: **Histogram of Kullback-Leibler (KL) divergence of probabilistic distribution of no. of spots on different clusters or annotations between closest section pairs, respectively of Mouse hippocampus, Drosophila embryo and Mouse brain datasets.** The position of average and maximum KL divergence was marked by red vertical lines and the respective values were labeled in black.

For example, in our application cases, the maximum KL divergence remains below 1 for both Mouse hippocampus and Drosophila embryo (Table R1), and Distributive Constraints was adopted in registration. While in Mouse brain dataset, the value reached 1.448 and the option was not adopted during registration. Annotation type distribution obviously varies across sections of Drosophila embryo (Fig. R16), yet Distributive Constraints is still adopted, leading to a successful registration. This indicates that ST-GEARS does not require identical grouping distribution to employ Distributive Constraints. The code for calculating maximum KL divergence has been uploaded to GitHub repository of ST-GEARS.

Table R1: Maximum **Kullback-Leibler (KL) divergence of Probabilistic distribution of no. of spots between closest section pairs and whether the Distributive Constraints was adopted in cases of Mouse hippocampus, Drosophila embryo and Mouse brain datasets.**

Application cases	Max. KL divergence	Distributive Constraints adopted
Mouse hippocampus	0.033	True
Drosophila embryo	0.499	True
Mouse brain	1.448	False

We have modified our section of **Reconstruct samples with different constraints settings** in Supplementary material to include the above guidance for adopting Distributive Constraints. And we have included our setting on this option in the material as well.

4. Excess zeros could be observed in next-generation-sequencing-based ST data, but the ST-GEARS approach does not consider this explicitly. Do the zero proportions influence the 3D recovery result of ST-GEARS?

Thanks for the insights on gene expression data sparsity and its influence. Through our study on applications of ST-GEARS on dataset with varying zero proportions, excess zeros have almost no influence on method performance.

In our study, we applied ST-GEARS on datasets sequenced by different methods, including Mouse brain dataset sequenced by Barcoded Anatomy Resolved by Sequencing (BARseq) which belongs to In situ sequencing (ISS), Mouse hippocampus data sequenced by Slide-seq, Drosophila embryo and larva data sequenced by Stereo-seq, and DLPFC data sequenced by Visium technology. Slide-seq, Stereo-seq and Visium all belong to Next-generation sequencing (NGS) technology, and excess zero proportion is observed on all 4 datasets (Fig. R17). On datasets both with and without excess zeros, ST-GEARS successfully reconstructed all sections (Fig. 2, Supplementary Fig. 8, Fig. 4, Fig. 5, Fig. 6). We further studied the change of mapping accuracy with different sparsity, and did not observe clear correlation between zero proportion and the accuracy results (Fig. R18), with coefficient of determination (R^2) being less than 0.2. Above results indicate the stability of ST-GEARS across different sparsity levels.

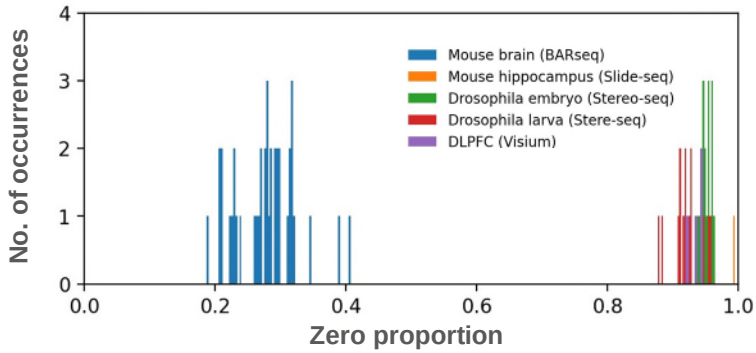


Figure R17: **Histogram of proportion of zero in expression matrix of Mouse brain, Mouse hippocampus, Drosophila embryo, Drosophila larva and Dorsolateral Prefrontal Cortex (DLPFC) sections.**

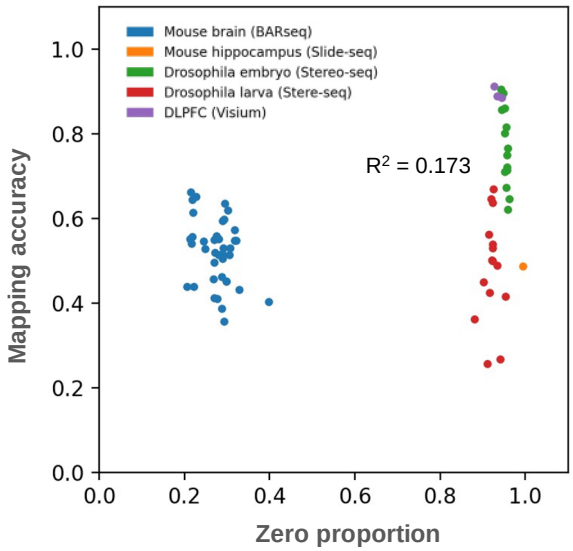


Figure R18: **Mapping accuracy and zero proportion of Mouse brain, Mouse hippocampus, Drosophila embryo, Drosophila larva and Dorsolateral Prefrontal Cortex (DLPFC) sections.** the coefficient of determination (R^2) was labeled in black.

Minor comments:

1. Line 48: "Visum" should be corrected to "Visium".

Thanks for pointing out. We have revised the corresponding spelling.

2. Mathematical notations should be consistent throughout the manuscript. For example, Line 589 introduces $X_A \in \mathbb{R}^{n_A,2}$, whereas Line 603 presents $X_{i,:}^{(A)}$.

Thanks for your careful observation. $X_A \in R^{n_A,2}$ introduced in line 589 (634 in revised version) and $X_{i,:}^{(A)}$ in line 603 (650 in revised version) are actually consistent. The subscript i of $X_{i,:}^{(A)}$ represents spot or cell index, and it is specifically labelled in $X_{i,:}^{(A)}$, to show projection of sectional spatial information of each spot $X_{i,:}^{(A)}$, to spot-wise distance measure $C_{i,j}^{(A)}$, where $C_{i,j}^{(A)} = dis(X_{i,:}^{(A)}, X_{j,:}^{(A)})$. In $X_A \in R^{n_A,2}$, the subscript i is not labelled due to the absence of need to show any projection. A is labelled as superscript of $X_{i,:}^{(A)}$ yet subscript of $X_A \in R^{n_A,2}$, because the subscript of $X_{i,:}^{(A)}$ is already used to label spot index i. Hence A is moved to the superscript and added parentheses, to represent it is the same matrix as X_A.

3. Some notations are repeatedly used. For example, ``W" in Lines 597, 638, and 686 have distinct meanings. And in Lines 601 and 618, the authors use C_A with different meanings.

Thanks for your careful observation and the very constructive suggestion. we apologize that same notation W was used to indicate distinct meaning in line 686 (671 in revised version) compared to line 597 (645 in revised version) and 638 (701 in revised version). We have revised the notation from W to $Width$ to represent size of the column dimension of elastic fields in all corresponding lines.

We also apologize for using C_A repeatedly for different meaning in line 601 (649 in revised version) and 618 (678 in revised version) and have changed the notations to differentiate them two.

Exceptionally, in line 597 (645 in revised version) and 638 (701 in revised version), W was actually differentiated by superscript to indicate their difference in meaning. In line 597 (645 in revised version), W appeared in $W_i^{(A)}$ and $W_j^{(B)}$. Subscript are taken by i and j, for us to specify projection between spot index with adjacency matrix π in $\sum_j \pi_{i,j} = W_i^{(A)}$. And hence section code A and B are moved to superscript and added parentheses to show that the matrices are actually W_A and W_B. $W_i^{(A)}$ and $W_j^{(B)}$ indicate constraints values of section A and B calculated by Distributive Constraints. In 638 (701 in revised version) and 639 (702 in revised version), $W_{\{i|C_i^{(A)}=c_i\}}^{(A\_raw)}$ and $W_{\{i|C_i^{(B)}=c_i\}}^{(B\_raw)}$ have the subscript occupied with the same reason, hence A_{raw} and B_{raw} appear as superscripts. $W_{\{i|C_i^{(A)}=c_i\}}^{(A\_raw)}$ and $W_{\{i|C_i^{(B)}=c_i\}}^{(B\_raw)}$ indicate constraints values of section A and B *before normalization*. As letter W is used in both $W_i^{(A)}$ and $W_{\{i|C_i^{(A)}=c_i\}}^{(A\_raw)}$ to represent constraint values, A and A_{raw} are used to specify their difference in respect to

normalization.

4. In Figure 4 panel a, "S-GEARS (rigid result)" should be "ST-GEARS (rigid result)"?

We appreciate for pointing this out and have revised the corresponding figure.

Reviewer 3

The manuscript titled "ST-GEARS: Advancing 3D Downstream Research through Accurate Spatial Information Recovery" tackles the challenge of accurately reconstructing the 3D morphology of tissue sections from their in situ spatial transcriptomics data. Current approaches to 3D spatial reconstruction suffer from significant inaccuracies due to their failure to account for experiment-induced distortions or their sole reliance on gene expression data without incorporating structural information. This results in discrepancies between reconstructed and actual in vivo cell locations, affecting downstream analyses. ST-GEARS introduces an innovative approach that utilizes optimized anchors between sections based on both expression and structural similarities. It incorporates Distributive Constraints into the optimization process, enhancing the precision of anchor retrieval. The method employs elastic fields for distortion correction and Gaussian Denoising for data quality improvement, significantly advancing the accuracy of spatial information recovery. By providing a more accurate method for reconstructing the 3D spatial profiles of tissue sections, ST-GEARS enables a deeper understanding of biological processes at the tissue, cell, and gene levels. Its ability to precisely recover spatial information supports more reliable downstream analyses, potentially unlocking new insights in developmental biology, organogenesis, and disease pathology, and fueling biological discoveries.

The method employs elastic fields within the Fused Gromov-Wasserstein (FGW) framework to correct experimental distortions by mathematically modeling the deformation that tissue sections undergo during experimental procedures. Elastic fields are used to represent how each point in the tissue is displaced or transformed, allowing for the adjustment of the spatial coordinates of gene expression data. This process involves calculating the optimal transformation that minimizes the difference between the distorted experimental data and the expected undistorted state, effectively 'undoing' the distortions and aligning the data more accurately with its original, undistorted configuration. This step is critical for ensuring that the reconstructed 3D spatial information accurately reflects the true morphology of the tissue.

I found the problem the authors tackle is very challenging and has a profound impact on our understanding of tissue 3D structure and cellular environment. The method discussed in this study presents a comprehensive strategy that aligns tissue slices by addressing limitations and gaps that were not resolved by existing approaches. Generally, I feel this is an important and useful methodology for the community. Yet, I have several major concerns that prevent me from recommending the paper in its current form (See below).

Thanks for the positive feedback of the overall method and its tackling of problems not solved by current approaches. And we appreciate the reviewer for insightful summary of each module of ST-GEARS including their designs and biological meanings. We have significantly improved ST-GEARS based on all your suggestions and questions. All significant modifications are marked in red in the revised manuscript. We hope this edition will address your concerns.

1. The manuscript's benchmarking framework, while inclusive of comparisons with GPSA, PASTE, and PASTE2, can be significantly enhanced by integrating STAlign and SLAT into the comparative analysis. The addition of STAlign, renowned for its precision in slice-to-slice spatial alignments, would provide a critical evaluation of ST-GEARS in terms of alignment accuracy and efficiency. Furthermore, although SLAT does not directly offer 3D reconstruction solutions, its inclusion could provide valuable insights into pairwise slice alignment capabilities. This broader benchmarking spectrum is essential for a comprehensive assessment, offering a clearer picture of ST-GEARS's technological advancements and its comparative effectiveness within the rapidly evolving field of spatial transcriptomics. Expanding the benchmarking to include these methods would not only highlight ST-GEARS's unique contributions but also help identify areas for further methodological refinement and development, ensuring its competitive edge and utility in addressing complex biological questions. (Also why GPSA benchmarking is missing for several sets in the study).

Thanks for your suggestion towards broader benchmarking spectrum. We have conducted registration of current application datasets with further inclusion of STalign and SLAT.

Since SLAT is a method generating mappings across spots or cells between sections, we focused on comparing its mapping accuracy with other mapping-involved methods including PASTE, PASTE2 and ST-GEARS, to understand its potential in pairwise slice alignment. Identical to the analysis we have conducted on anchors of other methods, we tagged each spot of human dorsolateral prefrontal cortex (DLPFC) section with the annotation of its mapping spot by SLAT with highest probability. We then compared this result to the tagged spot’s original annotation (Fig. 2a, Supplementary Fig. 1). SLAT generated similar annotation with ST-GEARS, excepts that it mapped multiple spots to spots from different tissue layers, particularly of spots located on layer 2, 4 and 6. The slightly inferior mapping is probably because of SLAT’s optimization framework: though it embeds spatial information in its graph neural network to propagate expression features, structural consistency term is not involved in its loss function in alignment solving, causing random spots on separate layers to be mis-aligned. Consequently, SLAT resulted in mapping accuracy that is higher than PASTE (Fig. 2b), yet slightly lower than PASTE2 and ST-GEARS. Similar results were witnessed on other 2 section pairs of DLPFC (Supplementary Fig. 1). ST-GEARS remains the method with highest mapping accuracy in the comparison.

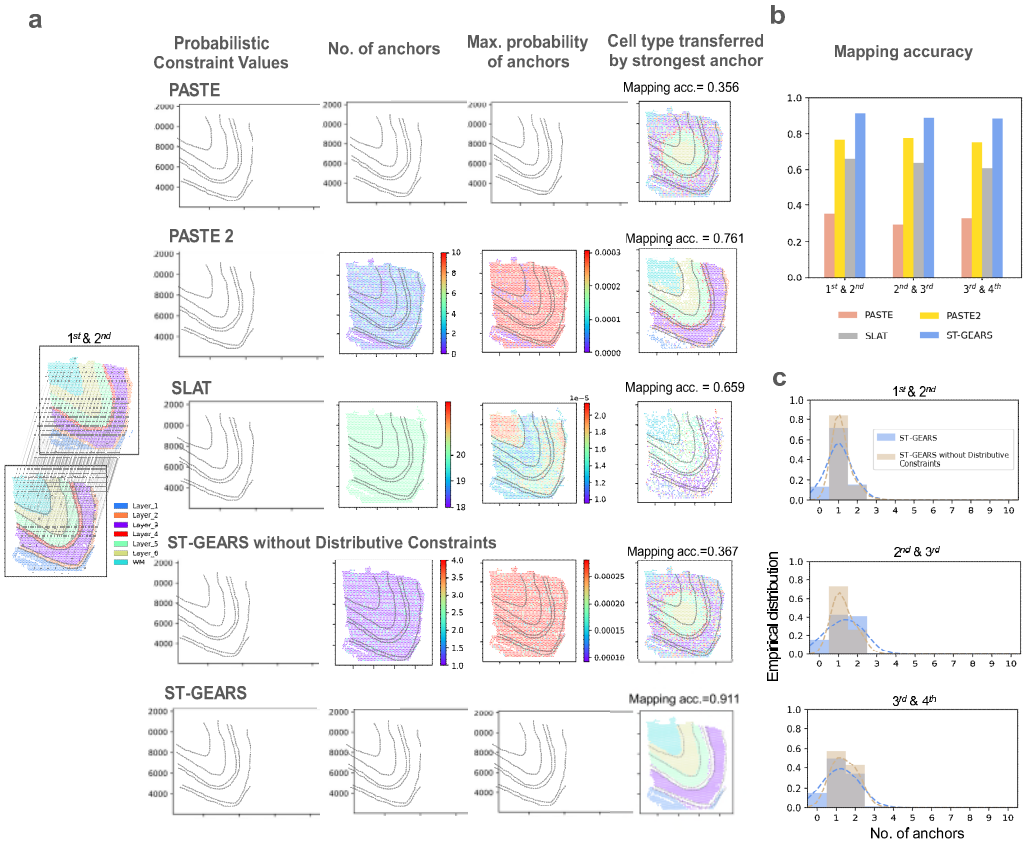
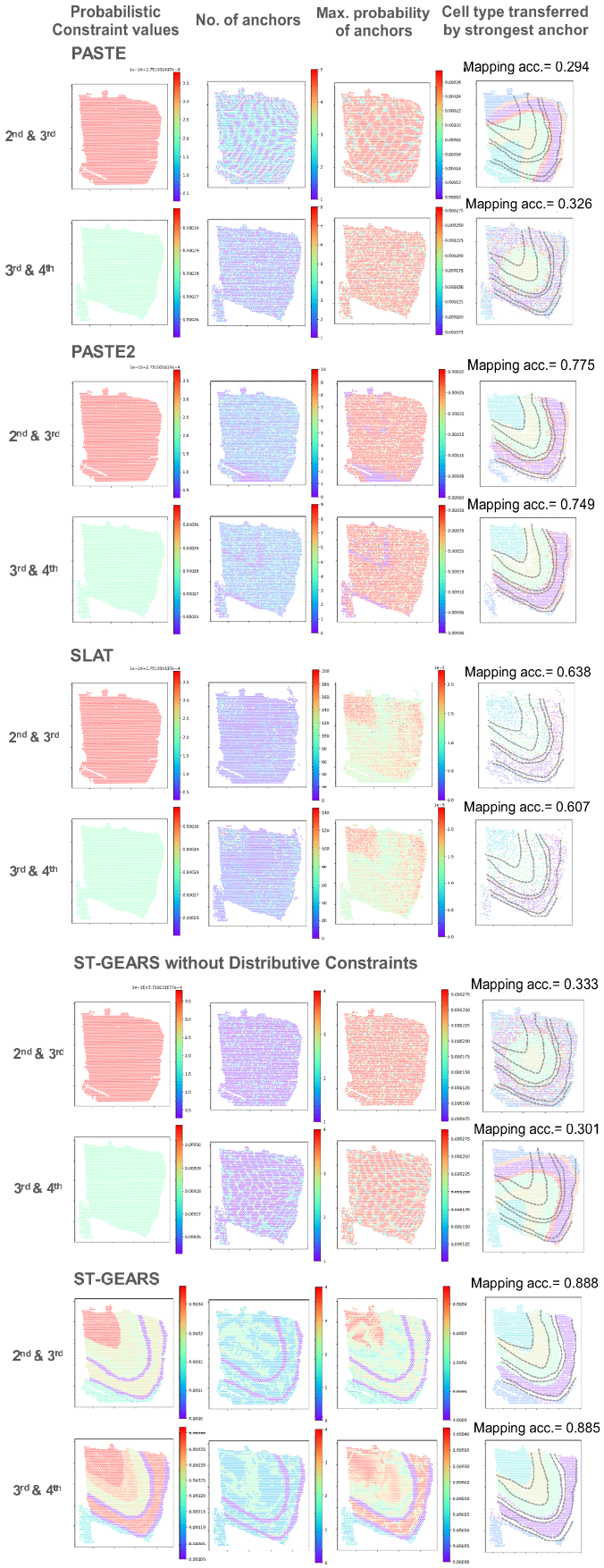


Figure 2: **Demonstration of anchors accuracy by ST-GEARS.** (a) (from left to right) 1st and 2nd human dorsolateral prefrontal cortex (DLPFC) section of patient #3 by Maynard et al.**Error! Reference source not found.** with their provided annotations and our anchors showcase, (of the same section pair) probabilistic constraints settings in Optimal Transport (OT) problem formulating, no. of anchors computed on each spot, max. anchor probability value computed of each spot, and annotated cell type mapped back to spots through computed anchors; (from top to bottom) respectively by PASTE, PASTE2, SLAT, ours without distributive constraints setting, and ours. The distinction of different cell types on the 1st section is marked by dotted lines. Mapping accuracy is marked alongside respective cell type mapping visualizations. (b) Mapping accuracy measured on anchors of sections pairs used in (b) by PASTE, PASTE2, SLAT, and ST-GEARS. (c) Comparison of no. of anchors histograms between ST-GEARS and ST-GEARS without distributive constraints, of sections pairs of 1st & 2nd, 2nd & 3rd, and 3rd & 4th sections. The Probability Density Function (PDF) estimated by Gaussian kernel was plotted in

dotted lines with the same color of histograms, to highlight the distribution differences.



Supplementary Fig. 1: **Distributive emphasis of different cell types of the 2nd to 4th section of DLPFC causes the advanced anchor accuracy.** From the 1st to the 4th column are presented probabilistic constraints settings in problem formulating, no. of anchors computed on each spot, max. anchor probability value computed of each spot, and annotated cell types on the next sections mapped back to its previous sections through computed anchors, with mapping accuracy marked. The distinction of different cell types on the sections are marked by dotted lines. 1st and 2nd row show analysis results of PASTE, 3rd and 4th show results of PASTE2, 5th and 6th row show results of SLAT, 7th and 8th row show results of ST-GEARS without distributive constraints settings, and 9th and 10th rows show results of ST-GEARS with distributive constraints settings. In the results of each method, the upper row presents result of the 2nd and the 3rd sections, while the lower row presents results of the 3rd and 4th sections. As ST-GEARS adopts distributive constraints, it generates relatively more and higher probabilistic anchors on cell types with higher expression consistency across sections, and hence it produces anchors with higher mapping accuracy.

Considering its registration function, we applied STalign on all registration application cases including Mouse hippocampus, Drosophila embryo and Mouse brain.

On Mouse hippocampus dataset, though fixing rotational misalignment to some degree, STalign left an angle between two slices in registration result (Fig. 4). This may be due to the method's processing of ST data into images which completely relies on gene expression abundance to decide pixel intensities. On the sagittal section of Mouse hippocampus, the abundance difference between regions may not provide sufficient structural information required by registration. By quantification, its MSSIM is lower than ST-GEARS, which remains more accurate registration method then PASTE, PASTE2, GPSA and STalign.

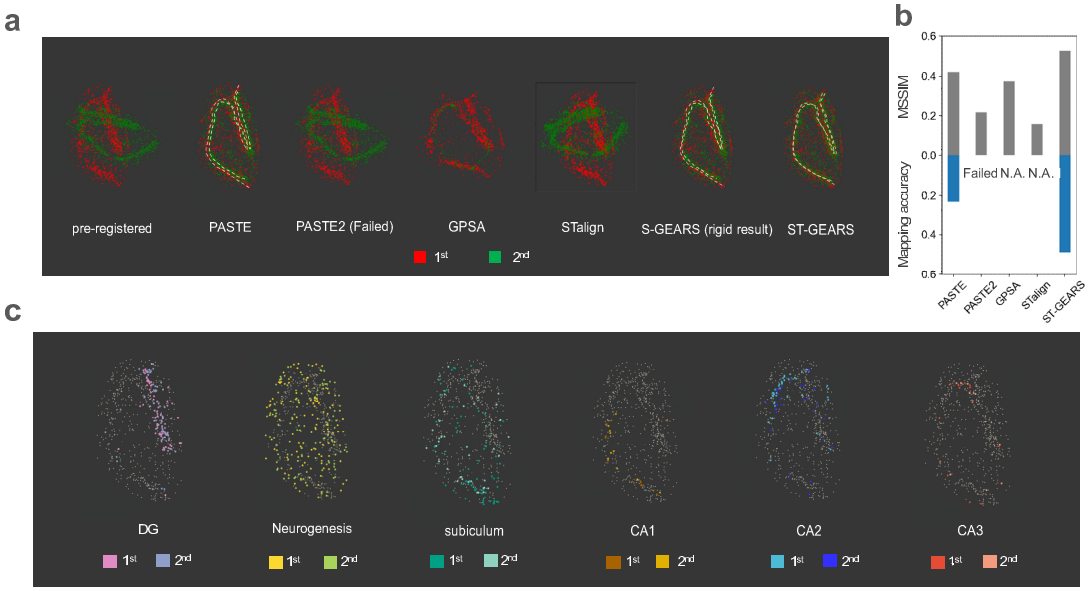


Figure 4: **Registration of Mouse hippocampus, respectively by PASTE, PASTE2, GPSA, STalign and ST-GEARS.** (a) Stacked projections of Cornu Ammonis (CA) fields and dentate gyrus (DG), of pre-registered and registered result of Mouse hippocampus sagittal sections with 10 μm distance, respectively by PASTE, PASTE2, GPSA, STalign and ST-GEARS. (b) A comparison of both MSSIM and Mapping accuracy of the 2 registered sections, across PASTE, PASTE2, GPSA, STalign and ST-GEARS. (c) Stacked projections of region-specific cell types including DG, Neurogenesis, subiculum, CA1, CA2 and CA3, registered by ST-GEARS. Each column highlights the stacked projection of a single cell type.

On Drosophila embryo dataset, STalign correctly aligned most sections(Fig. 5c). However, by comparing area changes with SI-STD-DI quantification of the complete section, and 3 individual tissues, STalign generated a result with less tissue area smoothness than ST-GEARS (Fig. 5b). This indicates STalign did not adequately correct distortion without support of enough structural messages in its generated image. By structural similarity analysis, ST-GEARS remains the method with highest MSSIM in 5 out of the 6 structurally consistent pairs (Fig. 5a). Analyzing each registered section, STalign produced a wrong flipping on the 13th section along A-P axis (Fig. 5c), probably caused by loss of intra-structural information in its image generating process. Stacking the projections back to 3D, a mistaken regionalization of foregut, caused by the wrong flipping, was circled in orange (Fig. 5d). In contrast to all compared methods, ST-GEARS avoided all their mistakes in its results (Fig. 5c, Fig. 5d). On gene level, high expression of Cpr56F was not witnessed on either anterior or posterior end on STalign's result in contrast to the hybridization evidence (Fig. 5f). And Osi7 is not obviously highly expressed on the top right outermost region by STalign's result as pointed by purple arrow. Hence, on both cell and gene level, ST-GEARS remains the most effective registration method towards addressing complex biological questions.

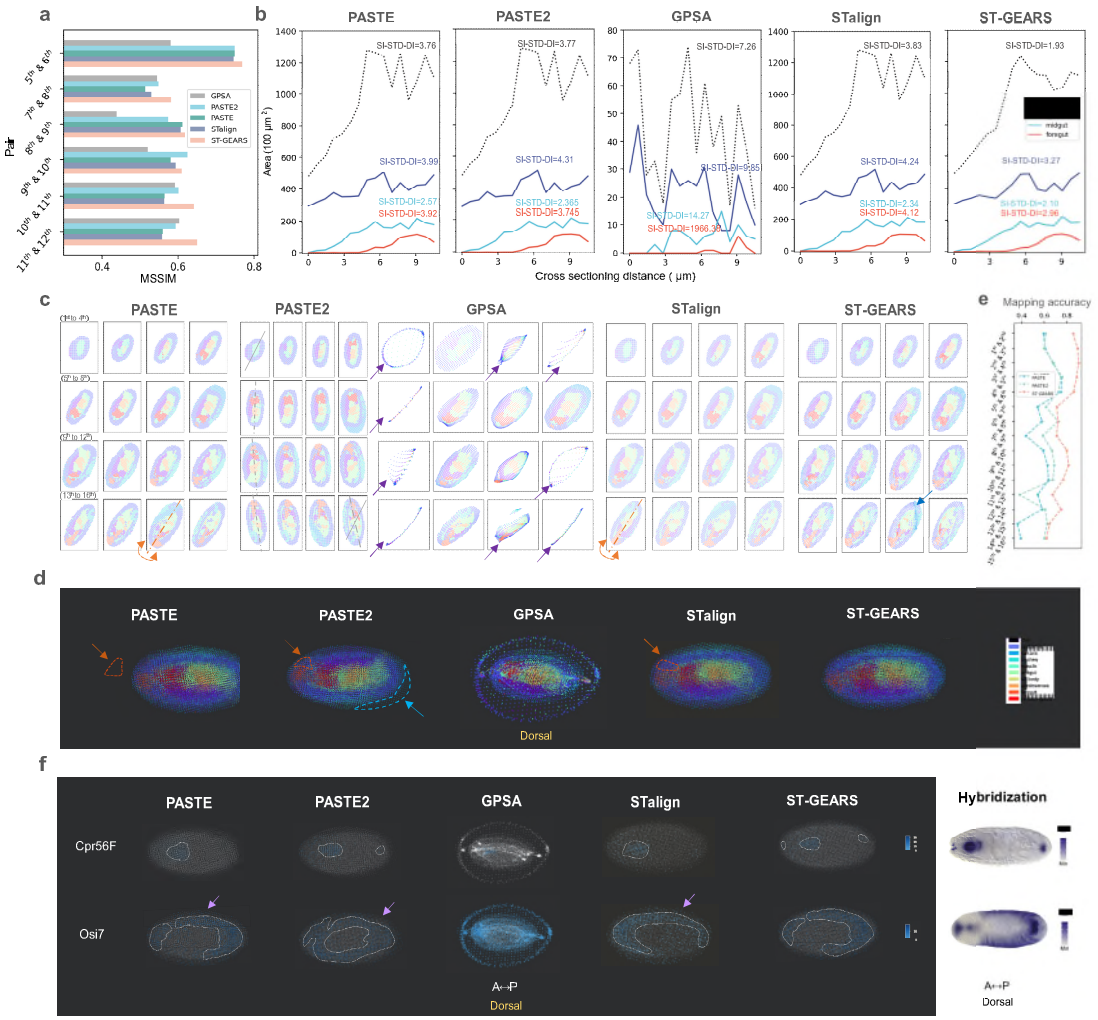


Figure 5: **Three-Dimensional (3D) reconstruction of Drosophila Embryo, respectively by PASTE, PASTE2, GPSA, STalign and ST-GEARS.** (a) A comparison of Mean Structural Similarity (MSSIM), of section pairs that are structurally consistent from Drosophila Embryo (E14-16h), between reconstruction results of PASTE, PASTE2, GPSA, STalign and ST-GEARS. (b) A comparison of area changes of 3 tissues and complete body of Drosophila Embryo, along cross-sectioning direction, between reconstruction result of PASTE, PASTE2, GPSA, STalign and ST-GEARS. Standard Deviation of Differences (SI-STD-DI) is marked alongside each curve to quantify the smoothness. (c) Reconstructed individual sections with recovered spatial location of each spot. In result of PASTE, the incorrect flipping on the 15th section was highlighted in orange. In result of PASTE2, gradual rotations were marked by the 1st, 5th, 9th, 13th and 16th sections' approximate symmetry axis whereas symmetry axis of the 1st section was replicated onto the 16th for angle comparison. In result of GPSA, mistakenly distorted sections were marked by purple arrows. In result of STalign, the incorrect flipping on the 13th section was highlighted in orange. In result of ST-GEARS, the fix of dissecting area on the 15th section was marked by a blue arrow. (d) Dorsal view of 3D reconstructed Drosophila embryo by PASTE, PASTE2, GPSA, STalign and ST-GEARS. The inaccurate regionalization of midgut was circled and pointed with arrow in orange. The resulted extruding part of single section by PASTE2 was circled and pointed in blue. (e) Mapping accuracy of all section pairs by PASTE, PASTE2 and ST-GEARS. (f) By dorsal view, regionalization of marker gene Cpr56F and Osi7 by PASTE, PASTE2, GPSA, STalign and ST-

GEARS, and their comparison with hybridization result from Berkeley Drosophila Genome Project (BDGP) database. The gathering expression regions were highlighted by dotted lines.

On Mouse brain dataset, though most sections were correctly aligned, 7 rotational misalignments were generated by STalign (Supplementary Fig. 19e). As ST-GEARS correctly aligned all 40 sections (Supplementary Fig. 19f), and reached highest median MSSIM score with the smallest variation (Fig. 6b), it remains the method with highest registration accuracy on the dataset.

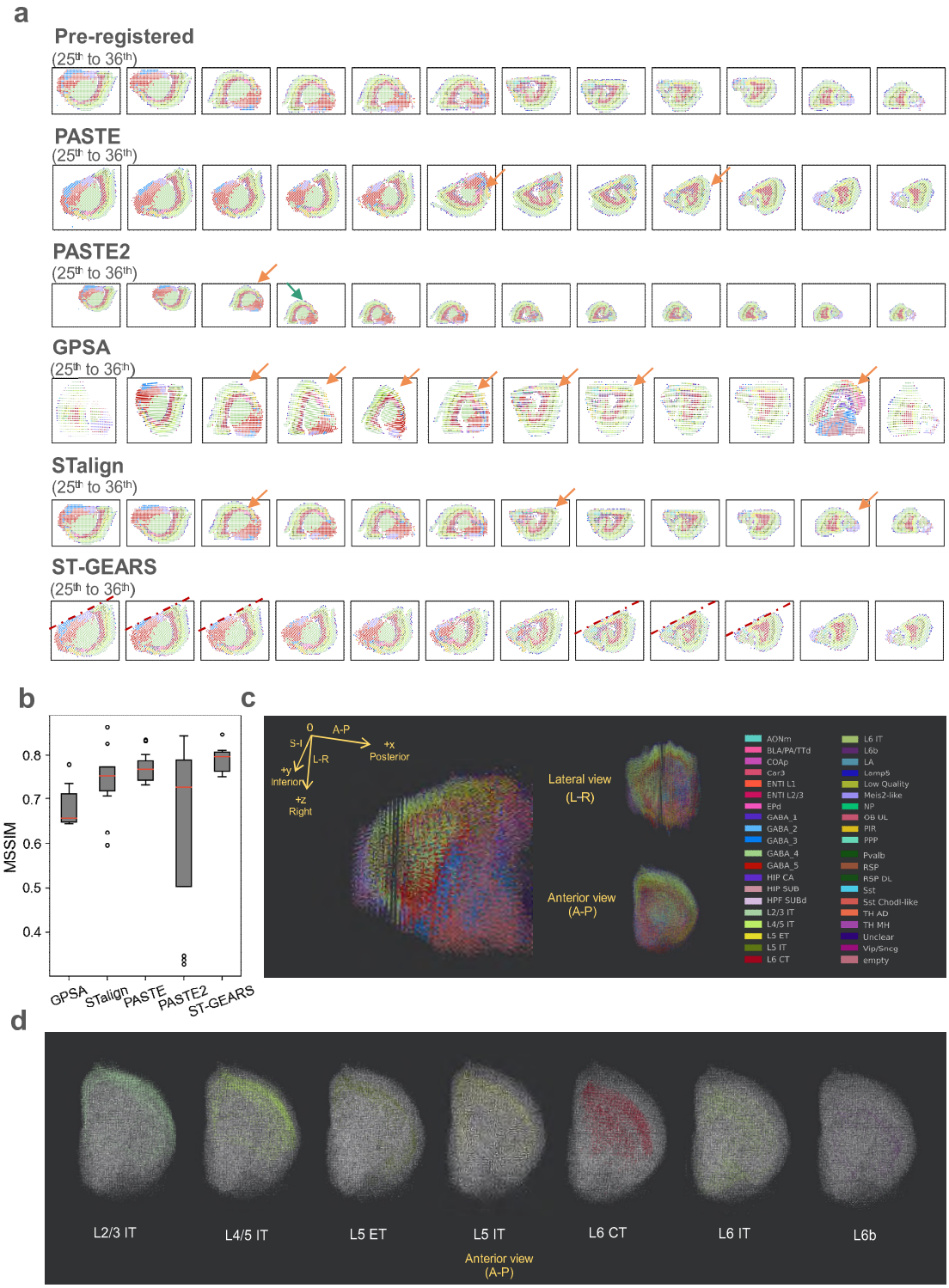


Figure 6: **Three-Dimensional (3D) reconstruction of Mouse Brain, respectively by PASTE, PASTE2, GPSA, STalign and ST-GEARS.** (a) Reconstructed individual sections with recovered spatial location of each spot from the 25th to 36th section. Positional misalignments are marked by arrows of green, and angular misalignments are marked by arrows of orange. Visible cutting lines by ST-GEARS are marked by dotted lines. (b) A comparison of Mean Structural Similarity (MSSIM) score of section pairs that are structurally consistent, between result of PASTE, PASTE2, GPSA, STalign and our method. The red lines positions show median score; the box extends from the first quartile (Q1) to the third quartile (Q3) of

scores; the lower whisker is at the lowest datum above Q1 – 0.5*(Q3-Q1), and the upper whisker is at the highest datum below Q3 + 0.5*(Q3-Q1); scores out of whiskers range are marked by circles. (c) Perspective, Lateral and Anterior view of reconstructed mouse brain hemisphere. (d) Anterior view of layer cell types distribution of reconstructed mouse brain hemisphere.

GPSA is not included in either mapping accuracy or elastic registration part of study, though studied in all 3 reconstruction application cases. Since GPSA directly adjusts spatial information of distinct sections onto Common Coordinate System (CCS), it doesn't involve intermediate or final result of mapping. Hence it was not included in mapping accuracy part of study (Fig. 1). In elastic registration session (Fig. 2), we focused on analyzing the effect of Elastic Registration component of ST-GEARS by comparing the result of our method with result when elastic operation is dis-included. Since GPSA generates final registration result in a concrete step, elastic operation cannot be taken away from the method, leaving a rigid result to be compared like ST-GEARS. Hence, we did not include GPSA in this part of study. In all application cases including Mouse hippocampus, Drosophila embryo and Mouse brain, we included GPSA, and measured its final registration accuracy using MSSIM and tissue area curve smoothness, equally with other methods.

We have modified our manuscript in red to include above results of SLAT and STalign, specifically in sections of Application to sagittal sections of Mouse hippocampus, Application to 3D reconstruction of Drosophila embryo and Application to Mouse brain reconstruction. We hope the reply and respective modifications address your comments

2. The authors' efforts in demonstrating ST-GEARS' performance across multiple real datasets are commendable, showcasing its practical application and robustness. However, the inherent limitation of ground truth in these datasets poses a challenge for systematic benchmarking. To address this, a recommendation for further strengthening the manuscript is to include benchmarking against simulated datasets. By artificially manipulating slices through rotation, scaling, cropping, and adding noise, the authors could generate controlled conditions to rigorously test and quantitatively compare ST-GEARS' performance. This approach would allow for a more precise evaluation of its capabilities in handling various distortions and noise levels, providing a comprehensive benchmark that underscores its accuracy and efficiency in spatial reconstruction.

Thanks for the insightful suggestion. We totally agree the ground truth provided by simulation dataset offers valuable perspective, of understanding our method's capability, accuracy, and stability. Hence, we conducted the following experiment. We manipulated a Mouse primary motor cortex section through rotating, transforming, cropping and distorting it using different scales, then analyzed accuracy result of ST-GEARS in registering the synthetic section with the original one.

Benefitting from the ground truth information of correspondence between cells, we adopted as accuracy index the correspondence fraction, which measures percentage of cells that are correctly connected by anchors. We also calculated mean distance error of cells over average cell distance to measure the error in scale of distance. Meanwhile, mapping accuracy adopted in other parts of the study is also included.

Across different scales of rotation, and different scales and direction of translation, ST-GEARS reaches full score of 1 on both correspondence fraction and mapping accuracy (Fig. R19, Fig. R20). Its mean distance error over avg. cell distance remains less than 1×10^{-10}, almost reaching error of 0. Upon different percentage of cropping, correspondence fraction and mapping accuracy of ST-GEARS remain close to 1, indicating ST-GEARS' high performance and stability to cropping operation. The distance error over avg. cell distance remains less than 0.6 which means cell level alignment is achieved. Correspondence fraction and mapping accuracy of ST-GEARS remains close to 1 across different scales of distorting noise, which was generated by incremental kernel variance of Gaussian Process (GP) warping. At the same time mean distance error over avg. cell distance remains less than 0.7 indicating cell level alignment. By different manipulations, ST-GEARS presented high performance and stability despite of operating scales.

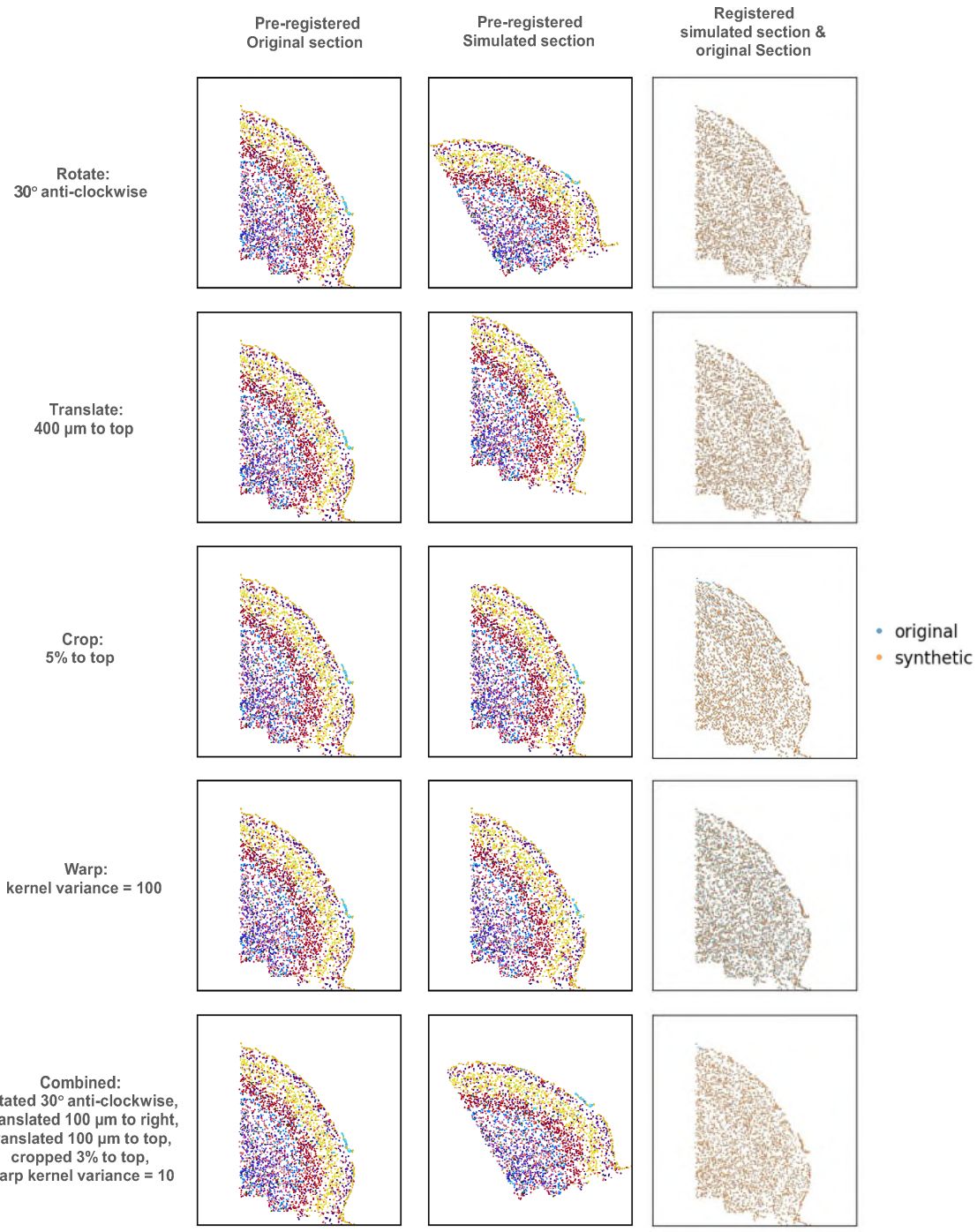


Figure R19: **Pre-registered and post-registered simulated sections with different manipulations.** From the top to bottom rows represented are manipulations of rotation, translation, cropping, warping and combined operation of above. The 1st and 2nd column shows pre-registered original sections and simulated section, and the 3rd column shows overlapping of registered sections.

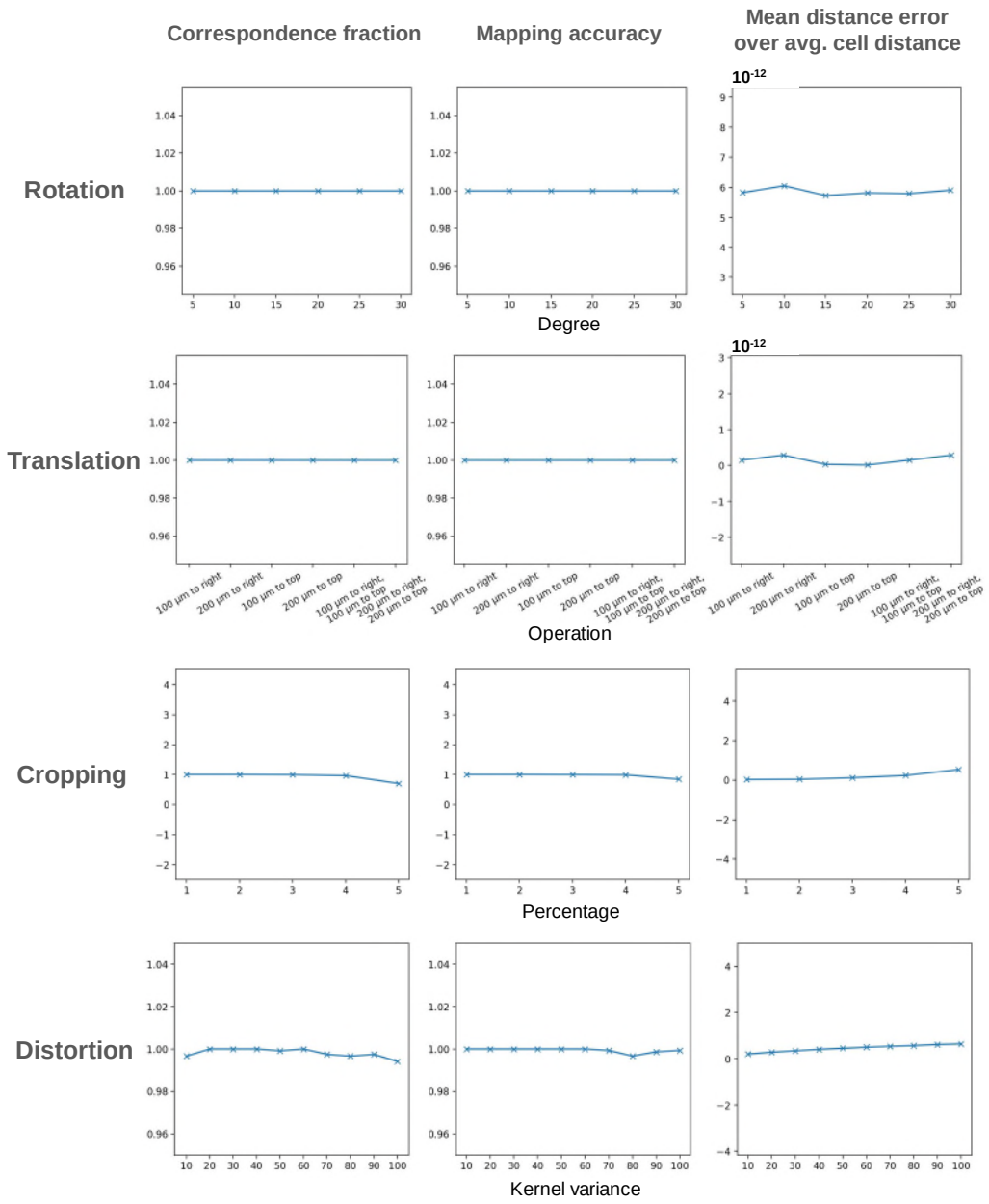


Figure R20: **Accuracy of ST-GEARS on simulated datasets.** From top to bottom presented are accuracy of ST-GEARS on simulated datasets with different levels of rotation, translation, cropping and distortion. From left to right presented are accuracy result based on index of correspondence fraction, mapping accuracy and mean distance error over average cell distance.

To further understand our method's performance in combined operation which often occurs in real world, we applied onto the section with operation including rotation of 30°, translation to right of 100 µm and to top of 100 µm, cropping fraction of 3% on top, and GP warping with kernel variance of 10 (Fig. R19). By registering the simulated dataset, ST-GEARS reaches 0.999 in corresponding fraction and 1 in mapping accuracy (Table R2). The mean distance error over avg. cell distance is less than 0.5.

Table R2: **Accuracy of ST-GEARS on simulated datasets with one section generated based on combined operation of the other one, including rotation, translation, cropping and distortion.**

Correspondence fraction	Mapping accuracy	Mean distance error over avg. cell distance
0.999	1.000	0.463

3. A critical weakness in ST-GEARS may lie in its computational complexity, particularly when processing large-scale datasets. The method's advanced features, such as optimized anchor alignment and elastic field application for distortion correction, could demand significant computational power and memory, impacting its efficiency. This aspect may limit its accessibility for researchers with limited computational resources or extend processing times for voluminous datasets. As the authors are undoubtedly aware, the volume of single-cell spatial transcriptomics data is ever-increasing, with datasets growing in scale and complexity. Therefore, it is imperative to ensure that the ST-GEARS algorithm can be efficiently applied to large-scale single-cell spatial datasets, as this is crucial for its broader adoption and practical utility in cutting-edge research. To address this concern and ensure the method's scalability, we kindly request that the authors provide a comprehensive analysis of both time and memory complexity in their manuscript. Such an analysis would not only serve as a testament to the method's computational efficiency but also provide valuable insights for researchers who may be considering its application on large-scale datasets. By presenting a detailed breakdown of time and memory requirements, the authors can demonstrate the method's ability to handle substantial datasets without compromising performance.

We appreciate your reminding and for the kind requests. We agree on the importance of controlling computational complexity, for our method to be adopted across diverse scenarios, especially on large-scale datasets.

As a comprehensive analysis of both time and memory complexity, we compared time and memory consumption of PASTE, PASTE2, GPSA, STalign and ST-GEARS across all application datasets. We find ST-GEARS reached least peak memory when registering Mouse brain (Fig. R13), and used second least memory on Mouse hippocampus and Drosophila embryo, with the data almost the same as the least one. In terms of time consumption, ST-GEARS is within the top two efficient methods on two out of the three applications, saving over 10 times of time cost than PASTE2.

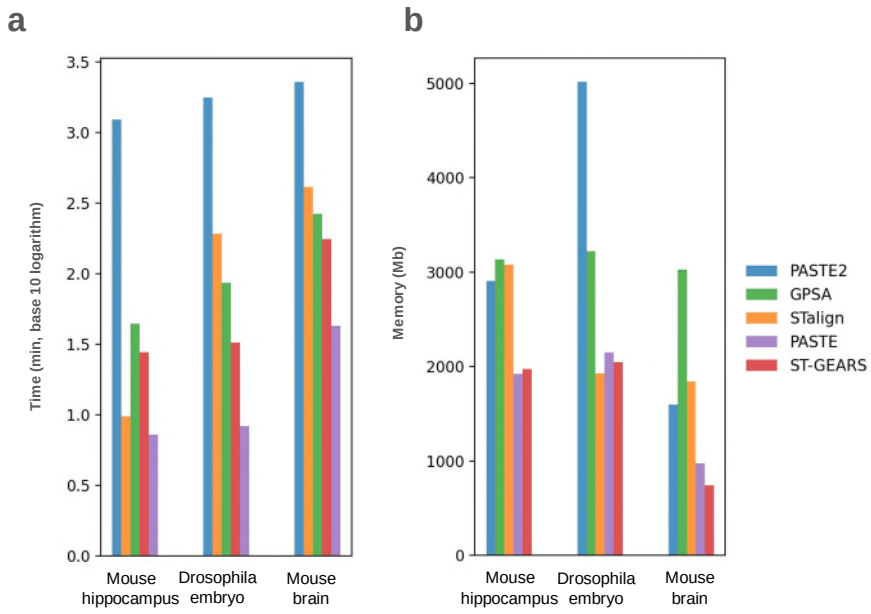


Figure R13: **Time and peak memory consumption of PASTE, PASTE2, GPSA, STalign and ST-GEARS, respectively on Mouse hippocampus, Drosophila embryo and Mouse brain datasets.**

To deal with the large data size problem, we introduce Granularity adjusting as a computational optimization to assist ST-GEARS. We recommend users to turn on this option upon over 3000 spots in each section.

In granularity adjusting, section area is binned first, with spots squared by each pixel summarized into one single spot, leading to a ST data with coarser resolution than original data. When summarizing within each grid, Unique molecular identifier (UMI) counts of spots is added together then transferred to the generated one spot, and the most frequent annotation type or cluster information is labelled to the spot as well. Then ST-GEARS is applied onto the coarser version of data, outputting a registered dataset with the coarser resolution. Finally, to recover the original resolution in final registration result, the original resolution data is interpolated into the coarse dataset on both pre-registered and registered version, leading to registration result in original resolution (Fig. R9). The conduction code of binning and interpolation method has been updated to GitHub repository of ST-GEARS.

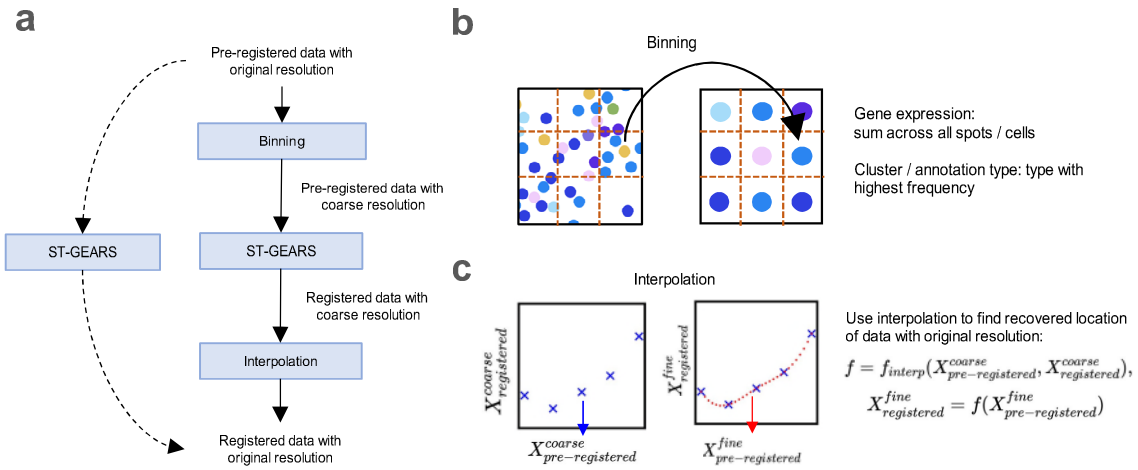


Figure R9: **Granularity adjusting option recommended for datasets with over 3000 spots.** (a) The automatic process of granularity adjusting includes binning the section area, which leads to bin sets that are much less compared to original spots, running the bin sets data with ST-GEARS, leading to registered data with the coarse resolution, and eventually interpolating original resolution data into the coarse one, outputting the registered data with original resolution. (b) In binning step, section area is gridded, with spots squared by each pixel summarized into one single spot. (c) In interpolating step, registered data with original resolution is solved by interpolating the pre-registered original resolution data into pre-registered (output by binning step) and post-registered (output by ST-GEARS step) coarse resolution data.

The granularity adjusting strategy enables higher computational efficiency of registration, without compromising accuracy of the reconstructed result. For example, we adopted granularity adjusting on Mouse hippocampus dataset and applied ST-GEARS on original dataset as well as binned dataset, using bin size of respectively 30 and 40 μm. Both time and memory by ST-GEARS largely decreased on binned dataset than on original resolution (Fig. R10). The higher bin size causes less spots number hence lower computational cost. By comparing coordinates of registration result by adopting granularity adjusting or not, the coefficient of determination (R^2) remains over 0.98 across all dimensions and sections, on both bin sizes (Fig. R11, Fig. R12), indicating that registration accuracy is not compromised by granularity adjusting option.

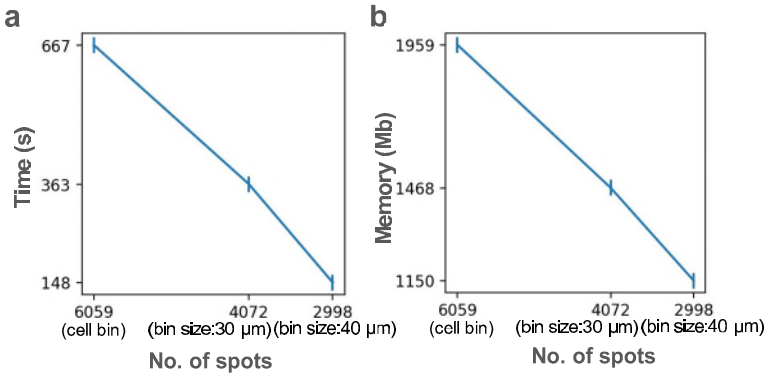


Figure R10: **Time and peak memory consumption by ST-GEARS, and number of spots of different resolutions of Mouse hippocampus.** The different resolution data includes dataset in original resolution

of cell bin, the binned data with bin size of 30 μm and the binned data with bin size of 40 μm.

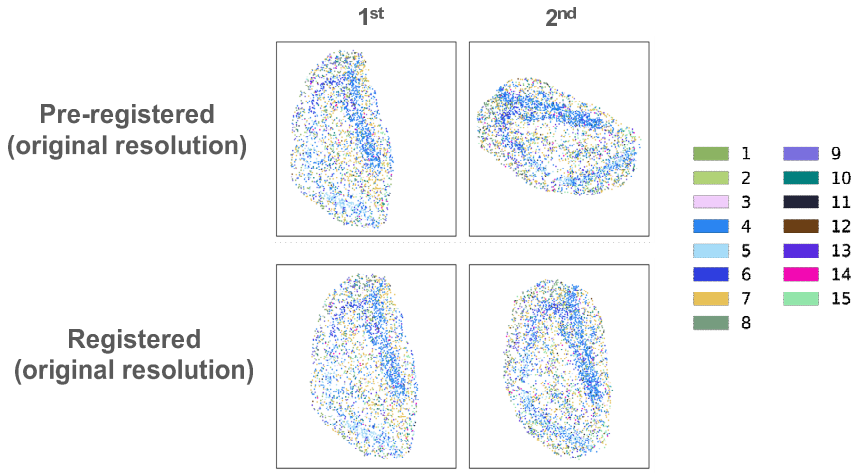


Figure R11: **Pre-registered and post-registered Mouse hippocampus dataset with original resolution.** The 1st row shows pre-registered dataset, and the 2nd row shows registration result of ST-GEARS directly using original resolution, without granularity adjusting adopted. The 1st column represents the 1st section, while the 2nd column represents the 2nd.

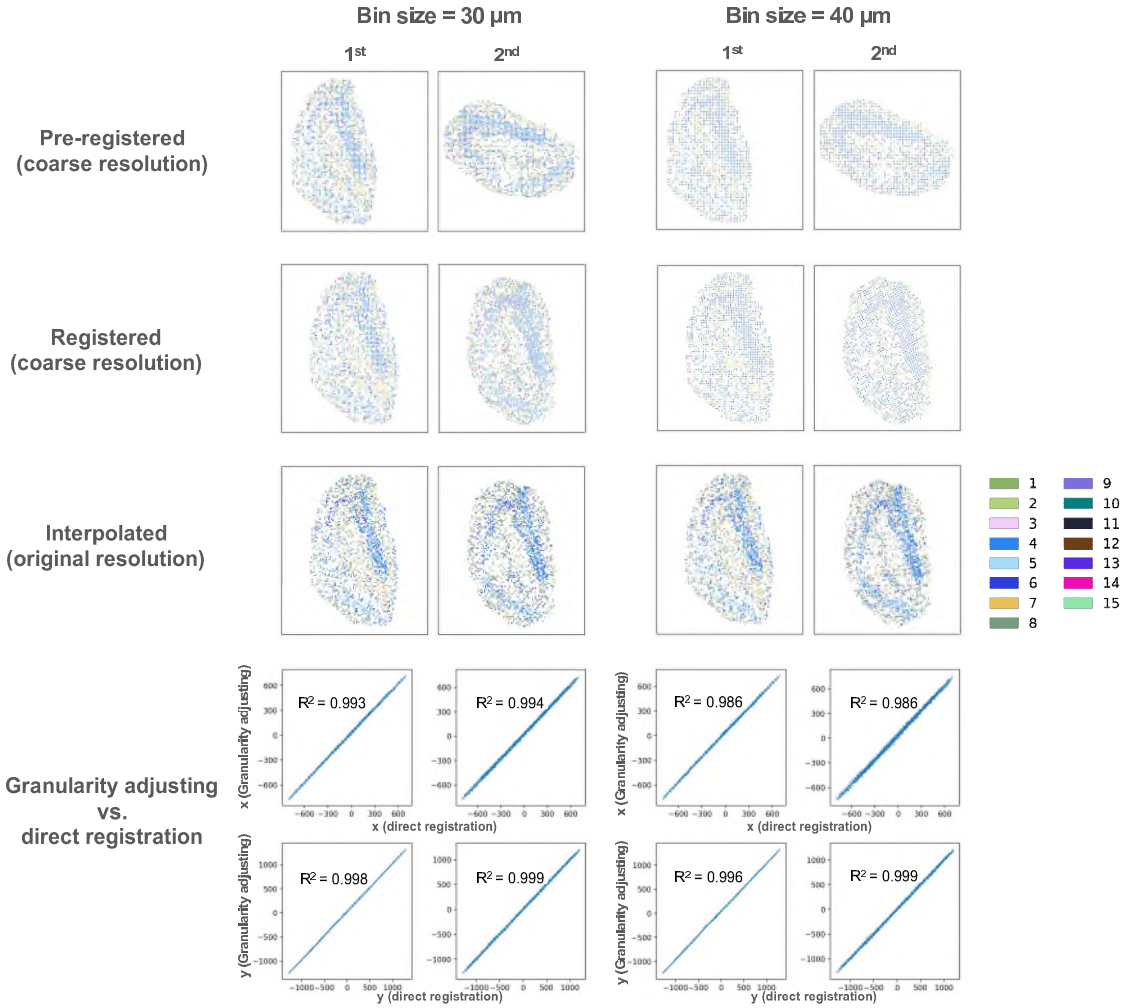


Figure R12: **Registration of Mouse hippocampus dataset with granularity adjusting strategy, in different bin sizes.** From top to bottom presented are binned yet pre-registered datasets, registered binned datasets, interpolation result of dataset with original resolution, and comparison between position of cells by process of granularity adjusting and direct registering through ST-GEARS. Two rows are included in the comparison plots, with the 1st row showing comparison of x coordinates and the 2nd row showing comparison of y coordinates. The 1st and 2nd columns are respectively results of 1st and 2nd sections by bin size of 30 μm in binning step of granularity adjusting, while the 3rd and 4th columns are result of the same two sections by bin size of 40 μm. In comparison between position of cells by process of granularity adjusting and direct registering, the coefficient of determination (R^2) remains over 0.98 across all coordinates, sections, on both bin sizes.

4. While the manuscript does provide evidence of ST-GEARS' application to various tissue types, another significant weakness that should be addressed relates to the potential for overfitting or hyperparameter sensitivity. The method incorporates multiple complex steps, including anchor selection, distributive constraints, Gaussian denoising, and elastic field modeling, each involving specific parameter choices. A potential weakness lies in the possibility that the performance of ST-GEARS is highly dependent on the fine-tuning of these parameters. If the method is sensitive to the choice of parameters, it could lead to overfitting on certain datasets or challenges in reproducibility

across different research groups. To mitigate this concern, it would be beneficial for the authors to provide a comprehensive sensitivity analysis that explores how variations in parameter settings impact the results. Additionally, recommendations or guidelines for parameter selection, based on the authors' extensive experience with the method, would aid users in achieving optimal outcomes. By addressing this weakness and offering insights into the robustness of ST-GEARS with respect to parameter choices, the authors can enhance the method's usability and reliability, ensuring that it can be successfully applied by a wider range of researchers without the risk of unintended biases or overfitting issues.

Thanks for the observation and for raising up your concern. To provide optimum flexibility for users, we indeed exposed many parameters of ST-GEARS method as being tunable, such as regularization factor list and the start and end index from section list on which section is to registered. However, most of them are exposed for users to conveniently conduct the process up to their own will, not to tune the process for a successful result. For example, users can assign a smaller number of regularization factors in the factor list to save more time, and to register only a part of sections from their dataset according to their own requirements. Except from the special circumstances such as above, most of the parameters are suggested not to be adjusted. We are sorry for the confusion! Meanwhile, default values have been provided for most parameters, and the values are sufficient for the registration to be successfully conducted. In the experiments of this study, the default values were adopted as well.

Only 3 parameters need to be specifically assigned values by users, including ‘uniform weight’, ‘label_col’, and ‘pixel_size’, and required or suggested values are provided for each one of them. Thereinto, ‘uniform_weight’ is a parameter of anchors computing process, ‘label_col’ is adopted by anchors computing, rigid registration and elastic registration process, and ‘pixel_size’ belongs to elastic registration process. The value adopted in our experiments are listed as below (Table R3).

Table R3: **ST-GEARS parameters that need to be specified and their value assigned across applications.**

Dataset	Uniform_weight	label_col	pixel_size
Mouse hippocampus	False	‘annotation’	10
Drosophila embryo	False	‘annotation’	1
Mouse brain	True	‘annotation’	200

‘uniform_weight’ specifies whether the Distributive Constraints will be adopted in the registration. Value of False indicates the adoption of the setting, and vice versa. To decide if Distributive Constraints shall be adopted, we suggest users to calculate probabilistic distribution of spot or cell numbers from each section, labelled with different clusters or annotations (Fig. R16), then to measure Kullback-Leibler (KL) divergence of the distribution between closest section pairs (Fig. R17) and to find the maximum divergence value. The code for calculating the maximum KL divergence has been uploaded to GitHub repository of ST-GEARS. If the maximum divergence is

below 1, Distributive Constraints is suggested to be adopted, since the annotation type or cluster distribution is close between every section pair, and the appropriate weight can be calculated based on the grouping information across sections. However, if the maximum KL divergence exceeds 1, users are encouraged to try ST-GEARS without Distributive Constraints.

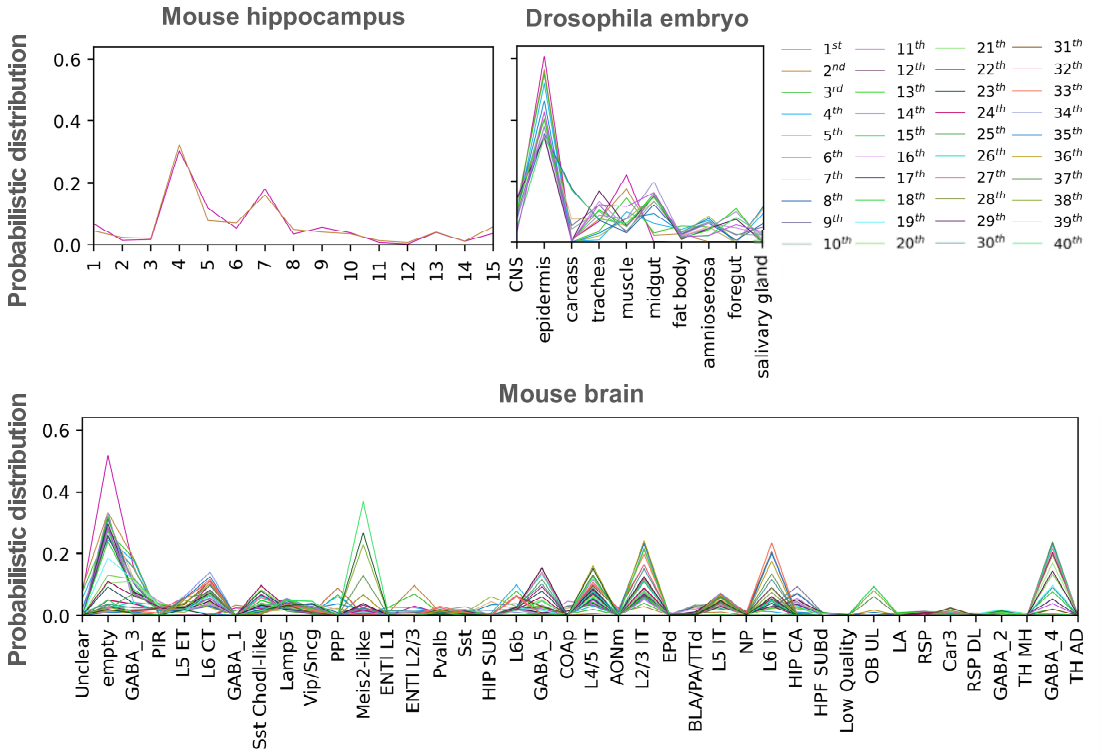


Figure R16: **Probabilistic distribution of proportion of spots on clustering / annotation types, of Mouse hippocampus, Drosophila embryo and Mouse brain datasets.** Each polyline was drawn based on distribution information of a section.

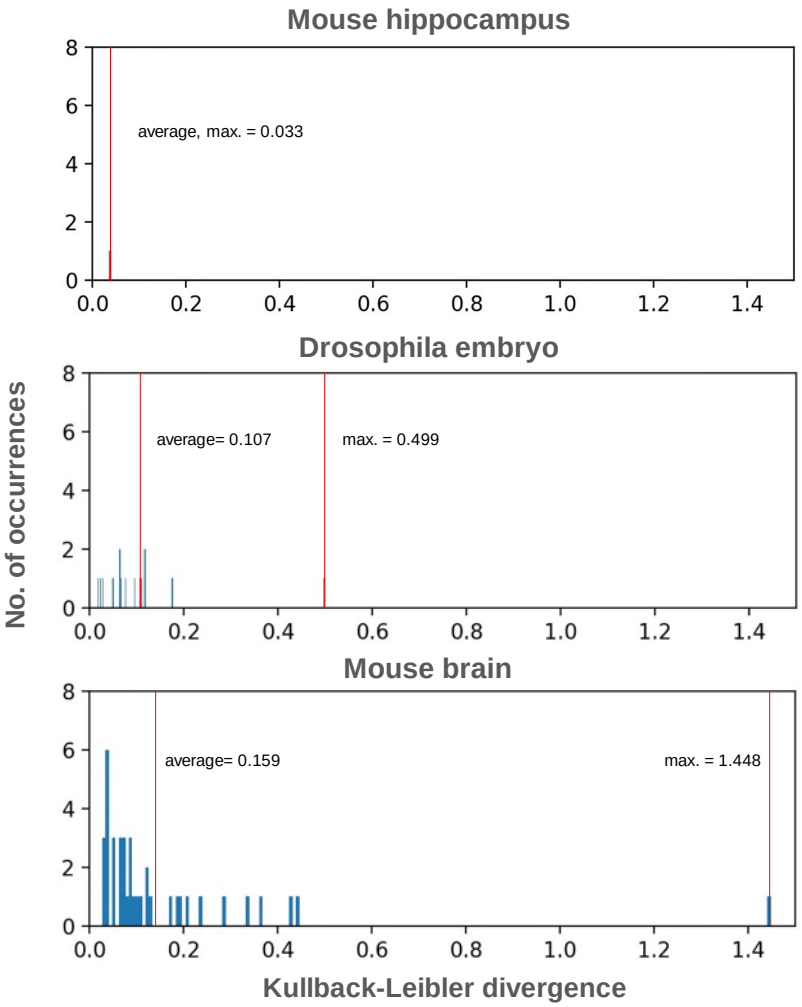


Figure R17: **Histogram of Kullback-Leibler (KL) divergence of probabilistic distribution of no. of spots on different clusters or annotations between closest section pairs, respectively of Mouse hippocampus, Drosophila embryo and Mouse brain datasets.** The position of average and maximum KL divergence was marked by red vertical lines and the respective values were labeled in black.

For example, in our application cases, the maximum KL divergence remains below 1 for both Mouse hippocampus and Drosophila embryo (Table R1), and Distributive Constraints was adopted in registration and hence 'uniform_weight' was set to False While in Mouse brain dataset, the value reached 1.448 and the option was not adopted during registration, hence 'uniform_weight' was set to True.

Table R1: Maximum **Kullback-Leibler (KL) divergence of Probabilistic distribution of no. of spots between closest section pairs and whether the Distributive Constraints was adopted in cases of Mouse hippocampus, Drosophila embryo and Mouse brain datasets.**

Application cases	Max. KL divergence	Distributive Constraints adopted
Mouse hippocampus	0.033	True
Drosophila embryo	0.499	True
Mouse brain	1.448	False

As input required by ST-GEARS is stored within anndata.AnnData, ‘label_col’ is required to be set to the name of column in .obs of AnnData, where clustering information or annotation type is stored. Hence, there is only one value to be adopted for this parameter for each dataset (Table R3).

‘pixel_size’ indicates step length on width and height when generating elastic field based on spatial coordinates data. The suggested value is the average distance between closest spots or cells (Table R3).

We have included above explanations, guidance of value assignments, and parameter values for applications in section of Parameter Settings in Supplementary materials.

(Remarks on code availability):
pros: easy installation and comes with an example jupyte note with test datasets.
cons: no detailed description of APIs (functions and modules, what are the functions of each method and their parameters)

Thanks for the positive comments on our package installation and example. We apologize for lack of adequate description of APIs and have uploaded the description to our GitHub repository, in file FunctionAPI.txt.

Reviewer #1 (Remarks to the Author):

I am very satisfied with the revised version of the paper. The authors have addressed all of my concerns, especially those arising from the model part. I only have two minor comments as follows.
1. Equation (1), I don't understand why (1-alpha)M_{AB}^2 in the first line equals to (1-alpha)<M_{AB}^2, pi> in the second line. Is this a typo?
2. My understanding is that the first term in Equation (1) measures the molecular similarity as KL doesn't consider any spatial information, while the second term solely measures spatial information. Alpha is the tuning parameter that controls the weight of spatial information. If so, it is very interesting to investigate the model performance given a range of alpha. The result as shown in Figure S24 should be reported to all the real datasets analyzed in the paper, not just the mouse brain dataset.

Reviewer #2 (Remarks to the Author):

The authors' responses addressed my concerns very well.

Reviewer #3 (Remarks to the Author):

The authors have provided a comprehensive and thoughtful response to the comments and suggestions that I raised. The revisions and detailed explanations have significantly strengthened the manuscript. The authors have effectively addressed each major concern, including the expansion of the benchmarking framework to include STAlign and SLAT, the use of simulated datasets to validate the robustness and accuracy of ST-GEARS, the detailed analysis of time and memory complexity along with the introduction of the granularity adjusting strategy, the provision of guidelines and recommendations for parameter selection to mitigate concerns about hyperparameter sensitivity and potential overfitting, and the addition of detailed API descriptions in the GitHub repository to improve accessibility and usability.

I have some minor comments to help further improve the manuscript.

1)Ensure that all legends are clear and provide sufficient detail for readers to understand the context and results without referring back to the text. Specifically, revise the figure legends to be more descriptive and add annotations to key figures to highlight important points or differences observed in the comparisons.

2)Verify that all mathematical notations are consistent throughout the manuscript and address any remaining inconsistencies. For example, the notations \(X_A \in \mathbb{R}^{n_A,2}\) in Line 589 (634 in revised version) and \(X_{i,:}^{(A)}\) in Line 603 (650 in revised version) should be clarified for consistency. Additionally, the notation \(W\) is used with distinct meanings in Lines 597 (645 in revised version), 638 (701 in revised version), and 686 (671 in revised version).

3) While the manuscript does provide a section discussing limitations, it would benefit from an expanded discussion explicitly addressing the potential for overfitting and hyperparameter sensitivity, as well as the scalability issue due to computational complexity. These are critical aspects that may impact the practical applicability and robustness of ST-GEARS, especially for large-scale datasets.

Overall, the revisions have significantly improved the manuscript. The detailed responses and additional experiments provided have addressed the major concerns. The final version of the manuscript should incorporate the minor suggestions mentioned above to further enhance its clarity and comprehensiveness.

Reviewer #3 (Remarks on code availability):

no detailed review of the code is necessary this time, as I reviewed it in the previous round and it looks good. The authors also added details for their APIs in this update, which addresses the concern that I raised from the last round of the review.

Reviewer #1 (Remarks to the Author):

I am very satisfied with the revised version of the paper. The authors have addressed all of my concerns, especially those arising from the model part. I only have two minor comments as follows.

Thanks for your positive feedback of our revision, and we appreciate all your suggestions and requirements which largely help strengthen this manuscript, especially in methods part. We have structured our response to your comments as follows, which will hopefully address your concerns.

1. Equation (1), I don't understand why (1-alpha)M_{AB}^2 in the first line equals to (1-alpha)<M_{AB}^2, pi> in the second line. Is this a typo?

The second line of equation (1) is actually derived from its first line. In the first line, the term $\langle(1-\alpha)M_{AB}^2 + \alpha L^2(C_A, C_B) \otimes \pi, \pi\rangle$ denotes matrix multiplication between term $(1-\alpha)M_{AB}^2 + \alpha L^2(C_A, C_B) \otimes \pi$ and term π. Since the multiplied term $(1-\alpha)M_{AB}^2 + \alpha L^2(C_A, C_B) \otimes \pi$ is essentially the summation of $(1-\alpha)M_{AB}^2$ and $\alpha L^2(C_A, C_B) \otimes \pi$, the multiplication operation can be derived into respective multiplication on each of the terms then summation of the multiplication results, whilst the constant variable $(1-\alpha)$ can be moved outside of multiplication operator as a constant variable, hence giving the second row: $argmin_{\pi \in \Pi(a,b)}((1-\alpha)\langle M_{AB}^2, \pi\rangle + \alpha\langle L^2(C_A, C_B) \otimes \pi, \pi\rangle)$.

To extract the structure of the above derivation and to further clarify the process, let A denote M_{AB}^2, and B denote $L^2(C_A, C_B) \otimes \pi$. The first row of equation (1) can be simplified as :

$$\pi = argmin_{\pi \in \Pi(a,b)}\langle((1-\alpha)A + \alpha B, \pi\rangle$$

And it can be derived as

$$\begin{aligned}\pi &= argmin_{\pi \in \Pi(a,b)}(\langle(1-\alpha)A, \pi\rangle + \langle\alpha B, \pi\rangle) \\ &= argmin_{\pi \in \Pi(a,b)}((1-\alpha)\langle A, \pi\rangle + \alpha\langle B, \pi\rangle)\end{aligned}$$

Replacing A and B with their original terms gives second row of equation (1).

2. My understanding is that the first term in Equation (1) measures the molecular similarity as KL doesn't consider any spatial information, while the second term solely measures spatial information. Alpha is the tuning parameter that controls the weight of spatial information. If so, it is very interesting to investigate the model performance given a range of alpha. The result as shown in Figure S24 should be reported to all the real datasets analyzed in the paper, not just the mouse brain dataset.

Thanks for your suggestion of investigating model performance given range of alpha. It is indeed that the first term in Equation (1) measures only molecular similarity, while the second term measures purely spatial similarity. To understand how the regularization factor α influences accuracy, we ran our method on all real datasets in our manuscript, across α that changes exponentially, to account for higher changing scale than a linear range.

We found that, ST-GEARS produces result with stable mapping accuracy across different α, on most real-world datasets except Drosophila larva and Mouse brain (Fig. R1, Supplementary Fig.

24). For Drosophila larva, the fluctuation is probably caused by changes of structural information across sections, such as tissue size changes from the 8th to the 11th section (Supplementary Fig. 3). While for Mouse brain, the mapping accuracy changes is mainly due to relative in-obvious similarity across anchored spots with high cross-sectional distances. To further clarify, distinct α values assign different weight hence different importance between structural similarity and expressional similarity. Hence, when either structural similarity changes drastically, or expressional similarity become less obvious, more noise is introduced in either one of them, causing the two terms the tendency to influence anchor results in different directions. In such way, the anchors accuracy tend to fluctuate when different weights are assigned.

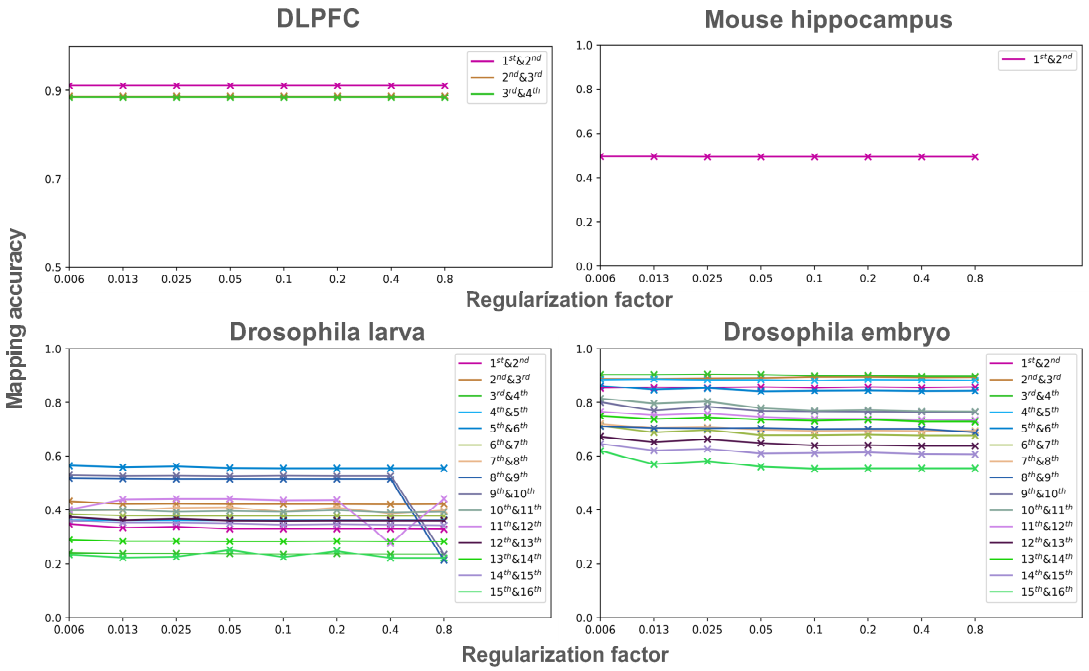
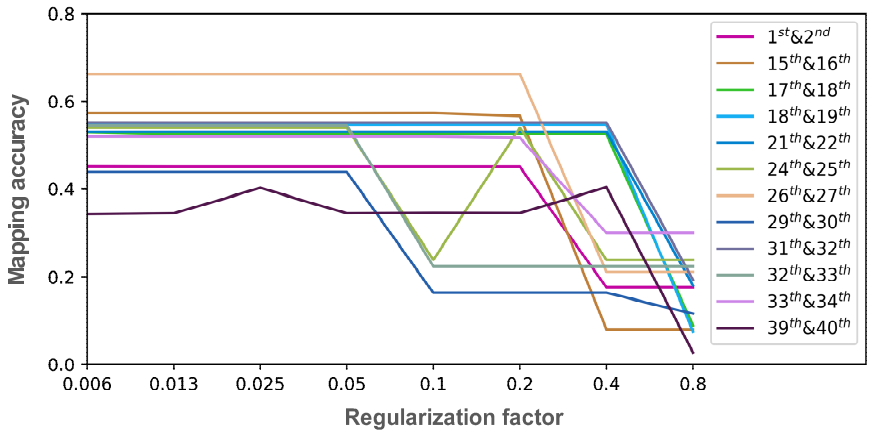
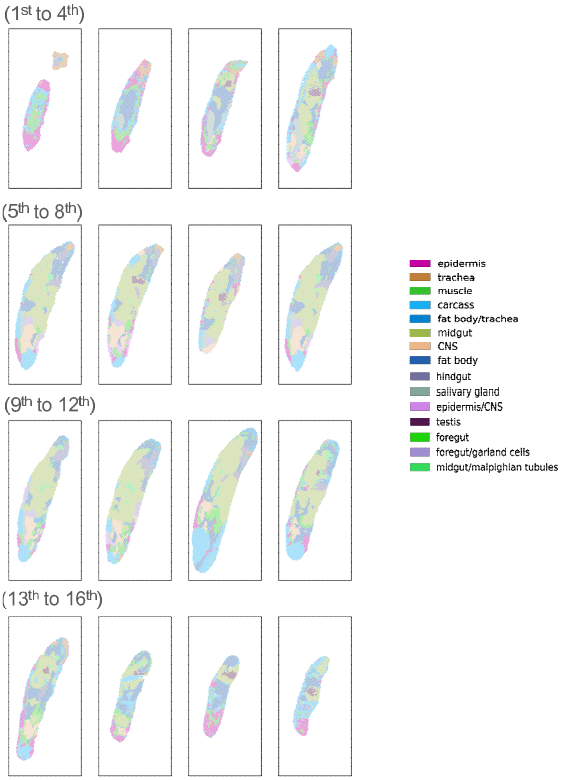


Figure R1: **Mapping accuracy changes with Regularization factor, on DLPFC, Mouse hippocampus, Drosophila larva and Drosophila embryo datasets.**



Supplementary Fig. 24: **Self-adaptive regularization generates regularization factor in Mouse brain**

registration that causes most advanced anchor accuracy. The figure shows changing mapping accuracy of anchors upon exponentially changing regularization factor. Section pairs with mapping accuracy changes over range of 0.1 are selected and plotted. ST-GEARS runs through the regularization factors, and adopts the factor that causes highest mapping accuracy to ensure a most advanced anchor accuracy.



Supplementary Fig. 3: **Individual sections of Drosophila larva generated by rigid registration of ST-GEARS.**

Please be noted that through Self-adaptive Regularization, the optimal α is selected by our method. The process essentially decides the optimal weight combination between structural and expressional similarity, leading to enhanced robustness of our method.

#2 (Remarks to the Author):

The authors' responses addressed my concerns very well.
We appreciate your positive feedback on our response. And thanks again for your constructive comments and suggestions during first round of revision, especially in terms of method's robustness and usability.

Reviewer #3 (Remarks to the Author):

The authors have provided a comprehensive and thoughtful response to the comments and suggestions that I raised. The revisions and detailed explanations have significantly strengthened the manuscript. The authors have effectively addressed each major concern, including the expansion of the benchmarking framework to include STAlign and SLAT, the use of simulated datasets to validate the robustness and accuracy of ST-GEARS, the detailed analysis of time and memory complexity along with the introduction of the granularity adjusting strategy, the provision of guidelines and recommendations for parameter selection to mitigate concerns about hyperparameter sensitivity and potential overfitting, and the addition of detailed API descriptions in the GitHub repository to improve accessibility and usability.

I have some minor comments to help further improve the manuscript.

Thanks for your positive feedback, and again for your insightful suggestions in benchmarking, computational complexity, parameters selection and code usability, in last round of review. We have structured our response to your comments in below, which will hopefully address your concerns. All significant modifications are marked in red in our revised manuscript.

1) Ensure that all legends are clear and provide sufficient detail for readers to understand the context and results without referring back to the text. Specifically, revise the figure legends to be more descriptive and add annotations to key figures to highlight important points or differences observed in the comparisons.

Thanks for your suggestions. We have ensured all legends provide sufficient details for readers to understand the context. We have revised the figure legends of Fig. 2, 3, 4, 5, 6 to be more descriptive, and have added annotation to Fig. 5b, which was missing in highlighting important differences. To further show sufficient details in Figure 6b, we added the n numbers of the boxplot, a clear definition of unit of study, and enough details about sample collection to distinguish between independent data points and technical replicates. Above revision are marked in red and we hope your concern has been addressed.

2)Verify that all mathematical notations are consistent throughout the manuscript and address any remaining inconsistencies. For example, the notations \(X_A \in \mathbb{R}^{n_A,2}\) in Line 589 (634 in revised version) and \(X_{i,:}^{(A)}\) in Line 603 (650 in revised version) should be

clarified for consistency. Additionally, the notation \(W\) is used with distinct meanings in Lines 597 (645 in revised version), 638 (701 in revised version), and 686 (671 in revised version).

Thanks for pointing out the notation problem. We have addressed all remaining inconsistencies in notation throughout the manuscript and have included explanations for clarity when necessary.

$X_{i,:}^{(A)}$ in line 652 (revised version) is actually the same notation as X_A in line 634 (revised version), and they both represent spatial location of spots on section A. To further clarify, when denoting $X_{i,:}^{(A)}$, section code A is moved to superscript, since subscript location is occupied by spot index i. Without this adjustment, $X_{i,:}^{(A)}$ would be denoted as $X_{Ai,:}$, which largely confuses readers when comprehending this matrix. To remind readers that $X_{i,:}^{(A)}$ and X_A indicate same meanings, a parenthesis is added to section code A. We noticed multiple subscript denoting adjustment in the Methods part and have included the explanation on each occurrence of the adjustment, marked in red.

We apologize that notation W was used with distinct meanings, and have revised notations to differentiate their representations, respectively in line 645, 708 and 709 (revised version). We have also checked throughout the manuscript for similar problems, and revised sessions including Distributive Constraints, Elastic Field Establishment and Bi-sectional Fields Application. All revisions are highlighted with red color.

3) While the manuscript does provide a section discussing limitations, it would benefit from an expanded discussion explicitly addressing the potential for overfitting and hyperparameter sensitivity, as well as the scalability issue due to computational complexity. These are critical aspects that may impact the practical applicability and robustness of ST-GEARS, especially for large-scale datasets.

Thanks for the suggestion of potential problem addressing. We have expanded our Discussion part to include the problem clarification and expectation of further studies, in terms of overfitting and hyperparameter sensitivity, as well as the scalability issue. The expanded discussion is marked in red.

Overall, the revisions have significantly improved the manuscript. The detailed responses and additional experiments provided have addressed the major concerns. The final version of the manuscript should incorporate the minor suggestions mentioned above to further enhance its clarity and comprehensiveness.

Thanks again for your review and comments.

Reviewer #3 (Remarks on code availability):

no detailed review of the code is necessary this time, as I reviewed it in the previous round and it looks good. The authors also added details for their APIs in this update, which addresses the concern that I raised from the last round of the review.

We appreciate your careful observation on our code and again, your previous suggestions.