TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses

Eric D. Sun[1], Rong Ma[2], and James Zou[1,*]

[1]Department of Biomedical Data Science, Stanford University

[2]Department of Statistics, Stanford University

[*]jamesz@stanford.edu

Spatial transcriptomics capture high-resolution spatial distributions of RNA transcripts within biological systems, yet whole-transcriptome profiling of genes at single-cell resolution remains a challenge. To address this limitation, spatial gene expression prediction methods have been developed to infer the spatial expression of unmeasured transcripts, but the quality of these predictions can vary greatly across different contexts. Poor predictions of gene expression in even a small subset of cells or genes can manifest in misleading downstream analyses. As such, there is a need for uncertainty-aware procedures for utilizing predicted spatial gene expression profiles. Here we present TISSUE (Transcript Imputation with Spatial Single-cell Uncertainty Estimation) as a general framework for estimating uncertainty for spatial gene expression predictions and providing uncertainty-aware methods for downstream inference. Leveraging conformal techniques, TISSUE provides well-calibrated prediction intervals for predicted expression values. Moreover it improves downstream analyses to consistently reduce false discovery rates for differential gene expression analysis, improve clustering and visualization of predicted spatial transcriptomics, and improve the performance of predictive models trained on imputed gene expression profiles. We have made TISSUE publicly available as a flexible wrapper method for existing spatial gene expression prediction methods to assist researchers with implementing uncertainty-aware analyses of spatial transcriptomics data.

1 Introduction

Spatial transcriptomics technologies extend high-throughput characterization of gene expression to the spatial dimension and have been used to characterize the spatial distribution of cell types and transcripts across

multiple tissues and organisms [31, 7, 32, 18, 34, 46]. A major trade-off across all spatial transcriptomics technologies is between the number of genes profiled and the spatial resolution such that most spatial transcriptomics technologies with single-cell resolution are limited to the measurement of a few hundred genes but typically not the whole transcriptome [24]. Given the resource-intensive nature of single-cell spatial transcriptomics data acquisition, computational methods for upscaling the number of genes and/or predicting the expression of additional genes of interest are highly desirable.

There exist several methods for imputing or predicting spatial gene expression using a paired single-cell RNA-seq dataset. Generally, these methods proceed by joint embedding of the spatial and RNA-seq datasets and then predicting expression of new spatial genes by aggregating the nearest neighboring cells in the RNA-seq data [1, 39, 2, 47] or by joint probabilistic modeling, mapping, or transport [27, 8, 44, 46, 11, 33]. For example, SpaGE relies on joint embedding of spatial and RNAseq data using PRECISE domain adaptation followed by k-nearest neighbors regression [35, 1]; a method referred to as "Harmony" here relies on Harmony integration of the two data modalities and averaging of nearest cell expression profiles [2]; and Tangram uses an optimal transport framework with deep learning to devise a mapping between single-cell and spatial transcriptomics data [8]. Applications of these methods have been used in the characterization of spatial differences in aging of mouse neural and glial cell populations [2], recovery of immune signatures in primary tumor samples [44], and identification of spatial patterns in gene regulation [8].

Since the relative performance of these models varies significantly depending on the application area and underlying datasets, there is no best model across all use cases [24]. Moreover, variability in model performance may adversely affect downstream analysis, particularly in promoting false discoveries due to prediction errors. At the same time, few existing gene expression prediction methods provide uncertainty measures for the predicted expression profiles and there are no approaches for utilizing uncertainty in downstream analyses like differential gene expression analysis, hypothesis testing, clustering, visualization, or predictive modeling. As a result, it is often difficult to rely on predicted spatial gene expression profiles without significant external validation or understanding of their context-specific uncertainties.

Here, we present TISSUE (Transcript Imputation with Spatial Single-cell Uncertainty Estimation) as a general wrapper framework around any spatial gene expression imputation or prediction model that produces well-calibrated uncertainty measures that are tailored to the context of each individual model and its use case. We show that TISSUE can be leveraged for improvements in various uncertainty-aware data analysis tasks including the calculation of prediction intervals, hypothesis testing using a multiple imputation approach, clustering and visualization using uncertainty-weighted principal component analysis, and predictive modeling (e.g. cell type classification, anatomic region classification) by uncertainty-aware filtering of cells before training and prediction.

2 Results

2.1 TISSUE: Cell-centric variability and calibration scores for prediction errors

Spatial gene expression prediction generally relies on leveraging spatial transcriptomics and RNAseq data from similar cell types. The RNAseq data is used to impute the expression of genes not measured in the limited spatial transcriptomics panel and can recover up to whole-transcriptome coverage of genes (Fig. 1A). To motivate the need for uncertainty quantification, we benchmarked three popular spatial gene expression prediction methods (Harmony [2], SpaGE [1], and Tangram [8]) on eight publicly available spatial transcriptomics datasets (spanning seqFISH [29], MERFISH [12], STARmap [45], ISS [21], FISH [23], osmFISH [13], and ExSeq [4] technologies; spatial data visualized in Fig. S1A) paired with single-cell or single-nuclei RNAseq datasets (spanning Smart-seq, Drop-seq, and 10X Chromium technologies) from the same organism and tissue regions [24, 26, 15, 19, 9, 45, 16, 20, 36, 13, 4, 17, 48, 42]. No method consistently outperformed other methods across all spatial transcriptomics datasets. Similarly, methods that performed the best under one metric (e.g. Spearman rank correlation between measured and predicted gene expression, see Fig. 1B) did not necessarily perform the best under an orthogonal evaluation metric (e.g. mean absolute error in predicted expression, see Fig. S1B). For a given method, there is also substantial heterogeneity in the relative performance of the model between genes and cells (Fig. 1A, Fig. S1BC), suggesting that accurate estimation of uncertainty in spatial gene expression predictions may improve confidence in interpretations and downstream analyses.

Here we leverage conformal inference as a natural framework for estimating well-calibrated uncertainties. First, we establish an initial measure of prediction uncertainty that is scalable to unseen observations and agnostic to the prediction error. In order to calibrate these uncertainties to the prediction error, we build distributions of calibration scores by linking these initial measures of uncertainty to the observed prediction errors on existing genes in the spatial transcriptomics data. Finally, these calibration score distributions are used for computing well-calibrated prediction intervals and improving downstream spatial transcriptomics data analysis.

To construct an initial measure of uncertainty that can be universally applied to all existing spatial gene expression prediction methods, we posit that, on average, spatially proximate cells with similar measured gene expression profiles will also have similar expression of genes that are not measured in the spatial transcriptomics gene panel. As a result, large differences in predicted gene expression between neighboring cells of the same cell type would indicate low predictive performance and highly similar predicted gene expressions between neighboring cells would signify high predictive performance for the spatial gene expression prediction method. To quantify this intuition, we introduce the cell-centric variability measure, U_{ij}, which,

given a gene, computes for each cell a weighted measure of deviation between the predicted expression of that cell and the cells within a spatial neighborhood of it (Eq. (1) and Eq. (2)).

$$U_{ij} = 1 + \sqrt{\frac{\sum_{k \in N_i} W_{ik}(\hat{X}_{kj} - \hat{X}_{ij})^2}{\sum_{k \in N_i} W_{ik}}} \tag{1}$$

$$W_{ik} = \exp\left(\frac{\hat{X}_{i,:} \cdot \hat{X}_{k,:}}{||\hat{X}_{i,:}||||\hat{X}_{k,:}||}\right) \tag{2}$$

Here, $\hat{X}_{ij}$ is the predicted gene expression of cell i and gene j. For a given cell i, its spatial neighborhood N_i corresponds to the K closest cells in the spatial transcriptomics data according to Euclidean distance. For each neighboring cell k, we compute a weight W_{ik} equal to the exponential cosine similarity in predicted gene expression profiles between the central cell i and its neighbor. These weights prioritize variability in predicted gene expression for similar cells (e.g. cells of the same cell type) and downplays expected variability in gene expression from dissimilar cells without the need to explicitly define cell types or states.

The cell-centric variability is generally positively correlated with the absolute prediction error for spatial gene expression (Fig. 1C). However, the cell-centric variability does not provide an exact estimate of the magnitude of these errors and the relationship between these two quantities is highly context-dependent (Fig. S2A). To explicitly link the cell-centric variability to the prediction error, we adapt a conformal inference framework by computing the calibration score, which is defined as the ratio between the absolute prediction error and the cell-centric variability (Eq. (3), Fig. 1A):

$$s_{ij} = \frac{|X_{ij} - \hat{X}_{ij}|}{U_{ij}}, \tag{3}$$

where X_{ij} denotes the measured gene expression for cell i and gene j.

The distribution of s_{ij} can subsequently be used to calibrate uncertainties for new expression predictions by multiplying the cell-centric variability of those predictions by specific quantiles of the s_{ij} calibration score set, which returns values on the scale of prediction errors (see next section for details). To permit flexible calibration schemes within the same spatial transcriptomics dataset, TISSUE allocates calibration scores to disjoint groups of genes and cells, referred to as the stratified calibration sets or groups, which are determined by k-means clustering of genes and then k-means clustering of cells by predicted gene expression (Fig. 1A, see Methods for further details). The number of gene and cell subsets (k_g, k_c) are user-specified, but downstream results are generally robust to exact specifications of these stratified groupings. Empirically, the distribution of calibration scores can vary significantly across different identified subsets, suggesting the identification of heterogeneous calibration score sets (Fig. 1D and Fig. S2B with $((k_g, k_c) = (4, 1))$. Due to

the asymmetric distribution of transcript counts, the calibration scores are further separated by the sign of the prediction error into a lower set for $X_{ij} - \hat{X}_{ij} < 0$ and upper set for $X_{ij} - \hat{X}_{ij} > 0$ (Fig. 1A).

2.2 TISSUE provides prediction intervals for predicted gene expression

We leverage a conformal inference framework to convert cell-centric variability of new spatial gene expression predictions into well-calibrated prediction intervals using the calibration scores derived from the measured gene panel. Given a gene expression prediction for cell a and gene b and confidence level α, we compute the cell-centric variability U_{ab} and multiply it by the $(\lceil (n+1)(1-\alpha) \rceil / n)$-th quantile of the upper and lower calibration sets of s_{ij} corresponding to all predicted expression values from the same stratified group as the predicted expression of cell a and gene b, which yields the asymmetric TISSUE prediction interval with approximate $1-\alpha$ coverage (Fig. 2A). Using this approach, TISSUE prediction intervals can be obtained for every predicted gene expression and every cell in the spatial transcriptomics data (see Methods for details and mathematical guarantees). The TISSUE prediction interval width is positively correlated with the absolute prediction error of measured genes under cross-validation (Fig. 2B). For individual genes of interest, the TISSUE prediction interval width generally reflects the spatial distributions of absolute prediction errors such as for *Plp1* in osmFISH profiling of mouse somatosensory cortex (Fig. 2C), *Neta* in a virtual spatial transcriptomics profile of *Drosophila* embryo (Fig. 2D), and *Tnfaip6* in MERFISH profiling of mouse primary visual cortex (Fig. 2E). Averaged across all genes and cells, the TISSUE prediction interval provides well-calibrated coverage of prediction errors on unseen genes for a broad range of confidence levels and across all prediction methods and spatial trancriptomics datasets (Fig. 2F). For individual genes, there is a general tendency towards well-calibrated prediction intervals (Fig. S3A). Similar calibration quality for prediction intervals was also observed under other choices of k_g and k_c (Fig. S3BC).

2.3 Uncertainty-aware hypothesis testing and differential gene expression analysis with TISSUE

Hypothesis testing of differences in gene expression between experimental conditions, cell types, or other groupings is an important tool in understanding biological heterogeneity and perturbation effects using spatial transcriptomics. We extend TISSUE calibration scores for more robust hypothesis testing of differential predicted gene expression across conditions. Specifically, TISSUE hypothesis testing involves sampling multiple imputations for the predicted gene expression values by first sampling calibration scores $\hat{s}_{ij}$'s from corresponding calibration sets and then perturbing the original predicted expression values by $U_{ij} \cdot \hat{s}_{ij}$ with the direction of the perturbation dependent on whether the sampled score was in the upper or lower set

(Fig. 3A). Repeating this process D times yields D possible imputations for each cell and gene. Using multiple imputation theory, TISSUE then derives corrected measures of statistical significance using a modified independent two-sample t-test (Fig. 3A, see Methods for further details). These corrected statistics account for the uncertainty in prediction as encoded by the sampling of scores for generating new imputations. This multiple imputation framework can easily be extended to other statistics of interest [37, 3, 25].

To compare TISSUE hypothesis testing to traditional hypothesis testing using only the predicted gene expression values, we generated synthetic data in which there are two groups of cells with the same ground truth gene expression. The measured gene expression is generated by adding standard Gaussian noise to the ground truth gene expression. Then, predicted gene expression is generated by adding biased Gaussian noise with mean μ to a portion of the genes in one of the cell groups but not the other, and standard Gaussian noise to all other measured gene expression values. Under this context, TISSUE hypothesis testing exhibited near zero error rate in calling differentially expressed genes between the two cell groups ($(k_g, k_c) = (2, 2)$, Benjamini-Hochberg corrected p-value cutoff for FDR=5%) across different levels of prediction bias while the error rate quickly saturated under traditional hypothesis testing even for low values of prediction bias (Fig. 3B).

In order to evaluate TISSUE hypothesis testing on real-world data, we compared it to the traditional hypothesis testing approach on three spatial transcriptomics datasets with associated cell type or anatomic region labels (see Methods for details on data labeling). For each label, we computed statistical significance of gene expression differences within that label as compared to all cells with different labels (i.e. one-versus-all approach). Statistical significance was assessed for all genes in the measured gene expression with traditional hypothesis testing and in the predicted gene expression with both TISSUE and traditional hypothesis testing. Using the differentially expressed genes detected using measured gene expression values as the ground truth, we observed lower false discovery rate of differentially expressed genes using TISSUE hypothesis testing as compared to traditional hypothesis testing across all prediction methods and datasets. The lower FDR was observed across for different numbers of differentially expressed gene detections (Fig. 3C) and for different p-value cutoffs (Fig. S4A). This decrease in FDR was also observed across different choices of k_g and k_c (Fig. S4BC). As such, application of TISSUE hypothesis robustly guards against false discoveries when performing differential gene expression analysis with predicted spatial gene expression profiles.

To illustrate a specific use case of TISSUE hypothesis testing, we applied the method to a ISS mouse primary visual cortex dataset. Using unbiased Leiden clustering, we identified several broad neuronal cell type clusters along with specific non-neuronal cell type clusters. We were unable to further resolve the neuronal cell clusters and used spatial gene prediction with SpaGE to predict the expression of additional neuronal subtype markers that were not in the original ISS gene panel. For one neuronal cluster, which

localized to the dentate gyrus of the hippocampus (Fig. 3D) differential expression was detected across most neuronal subtype markers, but under TISSUE hypothesis testing, we observed that this cluster had selective differential expression of predicted marker genes associated with granule cells (*Pdzd2*, *Gsg1l*, *Grp*), which are concentrated in the dentate gyrus (DG) of the hippocampus [6], and no significant expression of other cell type markers, including those for mossy cells (*Calb2*, *Tm4sf1*) and neural stem cells (*Prom1*, *Sox9*), which are other cell types found in the DG [6], and those for other neuronal subtypes (*Sst*, *Vip*, *Hcrt*, *Agrp*, *Pomc*, *Drd1*, *Tph2*) (Fig. 3E). The identification of this neuronal cell cluster as a granule cell cluster was confirmed by spatial localization of these cells to the hippocampal DG and by further confirmation with measured gene marker *Lrtm4*, which has been previously implicated with granule cell processes [40]. Under traditional differential expression testing with the predicted spatial gene expression values, we were unable to recover the same specificity for granule cell markers. We observed similar but reduced trends for Tangram prediction and lack of significance for any markers under Harmony prediction, likely due to the low performance of that model on this dataset (Fig. 1B).

2.4 Uncertainty-aware clustering and separation of labels with TISSUE

Cell clustering and visualization are commonly used to identify cell types in spatial transcriptomics data and intuitively understand high-dimensional differences between different metadata labels. TISSUE prediction intervals can be used to generate improved clustering and visualization of predicted gene expression profiles. To build an uncertainty-aware framework for these tasks, we leveraged weighted principal component analysis (WPCA) [14] with weights corresponding to a transformation of the inverse TISSUE prediction interval width for each cell and gene expression prediction (Fig. 4A, see Methods for further details). After applying this TISSUE-WPCA approach to the predicted gene expression data, we then extracted a subset of the top principal components for subsequent clustering and visualization of cells based on their gene expression.

To compare the performance of TISSUE-WPCA to traditional approaches, we generated synthetic spatial transcriptomics data with two distinct cell type clusters where half of the profiled genes are higher in expression in one cell type. Measured data is simulated by adding standard Gaussian noise to the ground truth expression (Fig. 4A). Then, predicted gene expression is simulated by selectively adding biased Gaussian error to a random proportion of cells in one cell type such that the expression profiles of those cells resemble the ground truth of the other cell type (i.e. mix-in bias, Fig. 4A). Using TISSUE, we computed 67% prediction interval widths (analogous to the asymmetric standard error), which were selectively higher in cells with mix-in bias in predicted gene expression and correlated with the absolute prediction error (Fig. 4B). We measured the separability of the two cell type clusters after applying either TISSUE-WPCA or normal PCA to the

predicted spatial gene expression profiles by fitting a support vector classifier with a linear kernel to the top 15 principal components and reporting the classification accuracy. TISSUE-WPCA robustly separated the two cell type clusters for all mixing proportions, including up to complete mix-in of the two cell type clusters, while normal PCA observed comparatively lower linear separability for all mixing proportions (Fig. 4C). Using stacked visualization of 20 independent synthetic datasets with DynamicViz [41], we observed that TISSUE-WPCA was able to separate the two cell groups along the first two principal components while the normal PCA was unable to do so under complete mix-in of the two cell groups (Fig. 4D). The DynamicViz variance score was also lower for the TISSUE-WPCA visualization than for the normal PCA visualization (median variance score of 0.223 compared to 0.252), indicating more stable visualization quality in the former, likely due to the improved representation of differences between the two cell type clusters.

Typically, Leiden clustering is used to cluster single-cell and spatial transcriptomics data, often using a subset of the top principal components as input [43]. On all real-world spatial transcriptomics datasets, we observed modest but significant improvement in Leiden clustering quality under TISSUE-WPCA as compared to normal PCA as measured by the adjusted rand index (ARI) with the same Leiden clustering applied to PCA of the measured gene expression serving as the reference clustering (P=0.0164 under paired two-sample t-test for positive mean improvement in ARI using TISSUE-WPCA). The TISSUE-WPCA clustering also reduced the effect of over-clustering that was present for normal PCA on the predicted spatial gene expression profiles with a median number of 17.0 clusters compared to 19.5 clusters respectively where the median number of clusters on the measured gene expression values was 11.0.

2.5 Uncertainty-aware predictive modeling with TISSUE

Predictive modeling is a common practice with single-cell and spatial transcriptomic data and can lead to useful models for predicting quantities of interest such as biological age [10], cell cycle state [38], and perturbational responses [28]. Substantial errors in spatial gene expression prediction may adversely affect the generalization performance of predictive models trained using the predicted gene expression profiles. We illustrate the utility of TISSUE on improving predictive models in a simple binary classification task using logistic regression. We performed stratified ranking and filtering of cells from the training and evaluation data with respect to the magnitude of uncertainty in their respective classes (Fig. 5A, see Methods for details). On the same synthetic data setup as in Fig. 4A, the classifiers trained and evaluated on TISSUE-filtered data generally outperformed the classifiers trained and evaluated on unfiltered spatial transcriptomics data in separating the two cell groups across different levels of mix-in prediction bias and under cross-validation (Fig. 5B).

To assess TISSUE-filtered training and evaluation on predictive modeling of real-world spatial transcriptomics datasets, we designed two simple classification tasks by retaining cells belonging to the three most prevalent classes of either cell type or anatomic region labels in the osmFISH mouse somatosensory cortex dataset. We then compared the cross-validated performance of logistic regression models trained and evaluated on the TISSUE-filtered predicted spatial gene expression to the performance of logistic regression models trained and evaluated on the unfiltered predicted gene expression profiles. Across three different performance metrics (accuracy, area under the receiver-operator characteristic curve, and F1 score), the TISSUE-filtered classifiers consistently outperformed the unfiltered classifiers on both prediction tasks (Fig. 5C). These trends were generally consistent when aggregating performances across additional spatial transcriptomics datasets for which we derived metadata labels for cell type and anatomic region (Fig. S5, see Methods for details on data labeling).

3 Discussion

Spatially resolved transcriptomics at single-cell resolution have paved the way to greater understanding of spatial patterns of gene expression and their underlying biological processes, but are currently limited to a relatively small number of detectable genes. Methods for predicting additional spatial gene expression profiles (e.g. from paired single-cell RNA-seq data) have been proposed to address this limitation. However, these methods can exhibit significant heterogeneity in prediction quality for different cells and different genes, and this heterogeneity manifests differently depending on the selected method and application. The lack of confidence measures for these predictions likely limits the adoption of these methods in spatial transcriptomics data analysis. Here, we developed a computational framework, TISSUE, to compute well-calibrated and context-specific measures of uncertainty for predicted spatial gene expression profiles.

In addition to well-calibrated prediction intervals, TISSUE provides general frameworks for leveraging uncertainty in downstream analysis such as differential gene expression testing, clustering and visualization, and predictive modeling. These frameworks are flexible and can be easily adapted into existing spatial transcriptomics data analysis workflows in place of traditional analysis methods that do not account for uncertainty. For example, the differential gene expression analysis testing approach can be adapted to other hypothesis tests (outside of the two-sample t-test showcased here) by modifying the within-imputation statistics measured for each imputation. Likewise, the weighted principal components obtained from TISSUE-WPCA can be used for any downstream algorithms that utilize the reduced dimensionality representation of PCA as input. Finally, the TISSUE-based filtering of training and evaluation data for predictive modeling strictly modifies the data in a model-agnostic manner and therefore can be extended to both training and deployment

of any predictive model across both regression and classification tasks.

TISSUE may have limited performance in contexts where spatial gene prediction patterns are not represented in the calibration set and for genes with extremely sparse expression patterns. Due to the assumptions underlying TISSUE, there may also be reduced performance on rare cell types that are not spatially co-localized. Since TISSUE performance is dependent on the size and diversity of the calibration sets, the method will generally scale better to spatial transcriptomics datasets with a large number of cells, a large number of genes, or more uniform representation of cell types. Although, specification of k_g and k_c to the natural dimensions of the data based on domain knowledge may provide optimal TISSUE performance, this performance is generally robust to most reasonable choices of k_g and k_c.

The main computational burden imposed by TISSUE is the cross-validated prediction of gene expression on the calibration set, which is necessary for building context-specific uncertainties. For k-fold cross-validation with a given prediction method, TISSUE prediction would take k times longer than a single prediction. Alternative approaches for measuring uncertainty such as variance in predicted expression values over multiple predictions would require significantly more computation than cell-centric variablity to generate predictions for context-specific calibration (i.e. $k \times l$ times longer than a single prediction for the variance over l predictions). Here we used 10-fold cross-validation in our experiments, but in practice, the number of cross-validation folds can be specified by the user to achieve a desired runtime. The computational burden for computing cell-centric variability, calibration scores, and prediction intervals is comparatively less than that for generating the initial predictions (Fig. S6).

Although TISSUE has thus far been tested in the spatial transcriptomics setting, the underlying assumptions can generalize to other spatial data modalities, such as spatial proteomics. As whole-transcriptome spatial gene expression profiling becomes possible in the future and other spatial -omics technologies mature, we anticipate that TISSUE will find additional use in the prediction and quantification of uncertainty for enhanced spatial data analysis across multiple modalities.

4 Main Figures

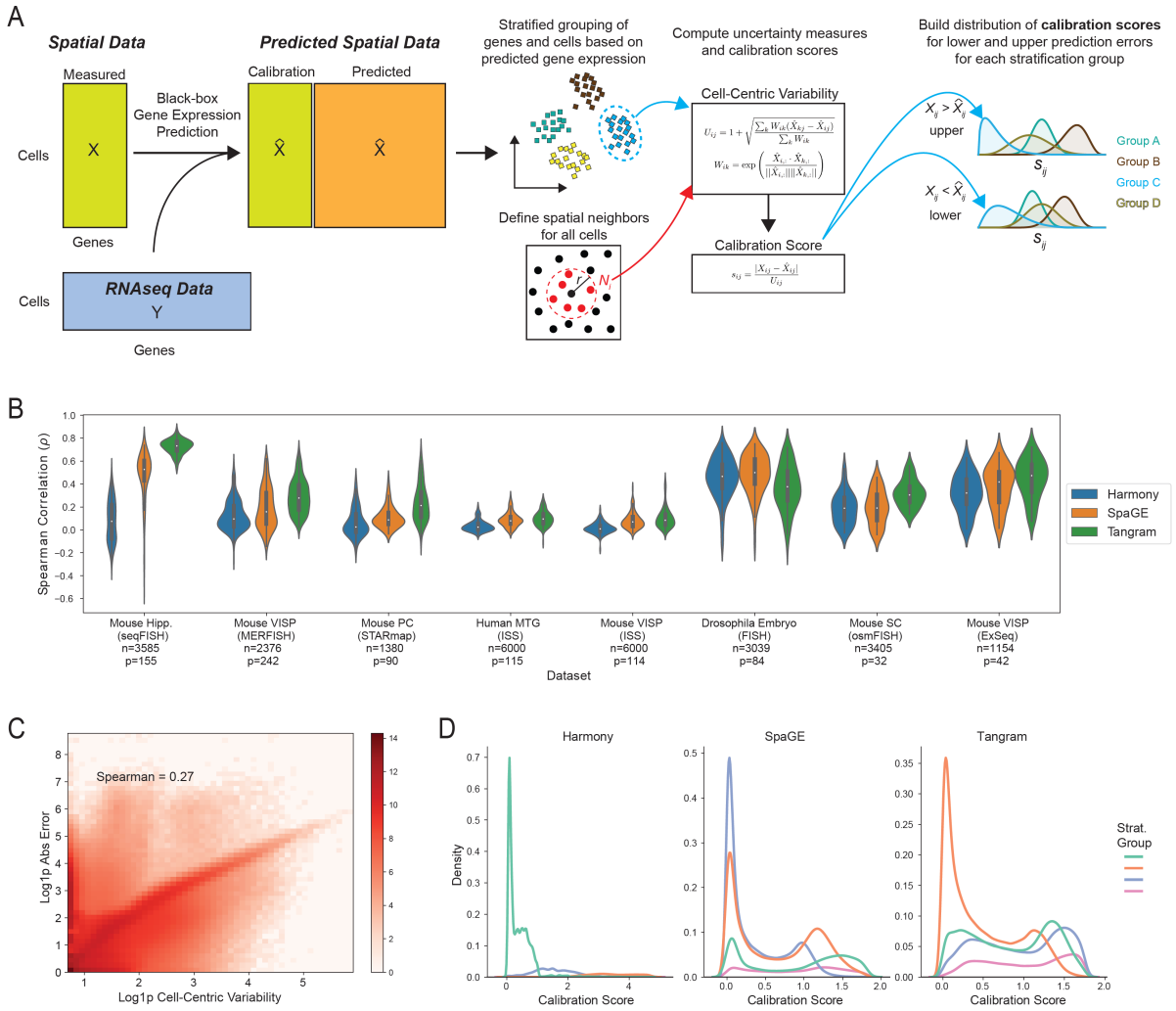


Figure 1: Cell-centric variability and calibration scores for conformal inference. (A) Schematic illustration of the TISSUE calibration score generation pipeline with black-box gene prediction (shown is an example method using paired spatial and RNAseq datasets), stratified grouping of genes and cells, calculation of cell-centric variability measure, and computation and allocation of the calibration score to different stratification groups. (B) Performance of three popular gene prediction methods (Harmony, SpaGE, Tangram) on eight benchmark datasets as measured by the Spearman correlation between predicted and actual gene expression over 10-fold cross-validation. Also shown are the number of cells (n) and number of genes (p) in the spatial transcriptomics datasets. (C) Correlation of TISSUE cell-centric variability and absolute prediction error across all dataset and prediction method combinations computed over 10-fold cross-validation. Log density with added pseudocount (Log1p) is shown by color. (D) Distribution of TISSUE calibration scores on mouse hippocampus ISS dataset and all three prediction method combinations ($(k_g, k_c) = (4, 1)$). Details of each dataset and prediction method can be found in Methods.

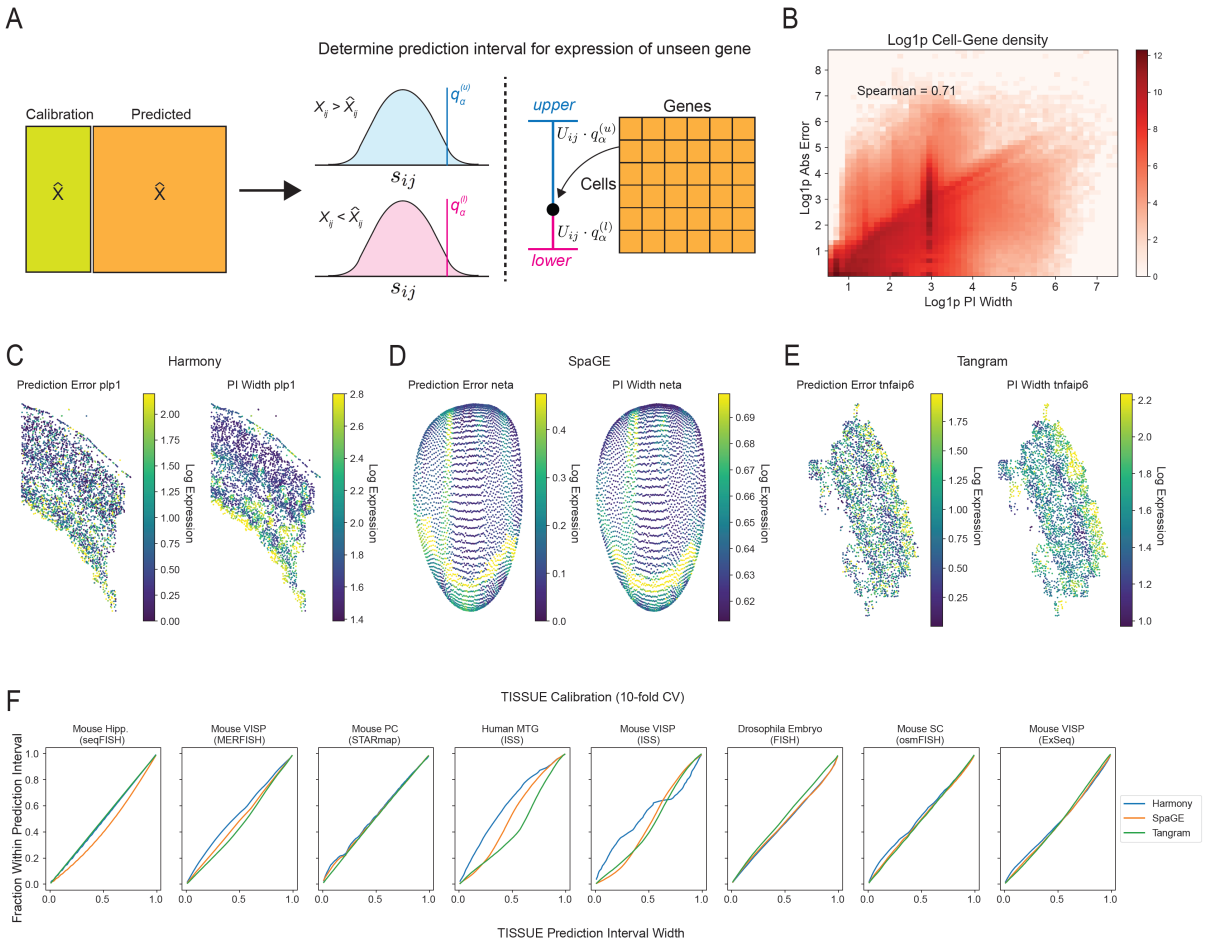
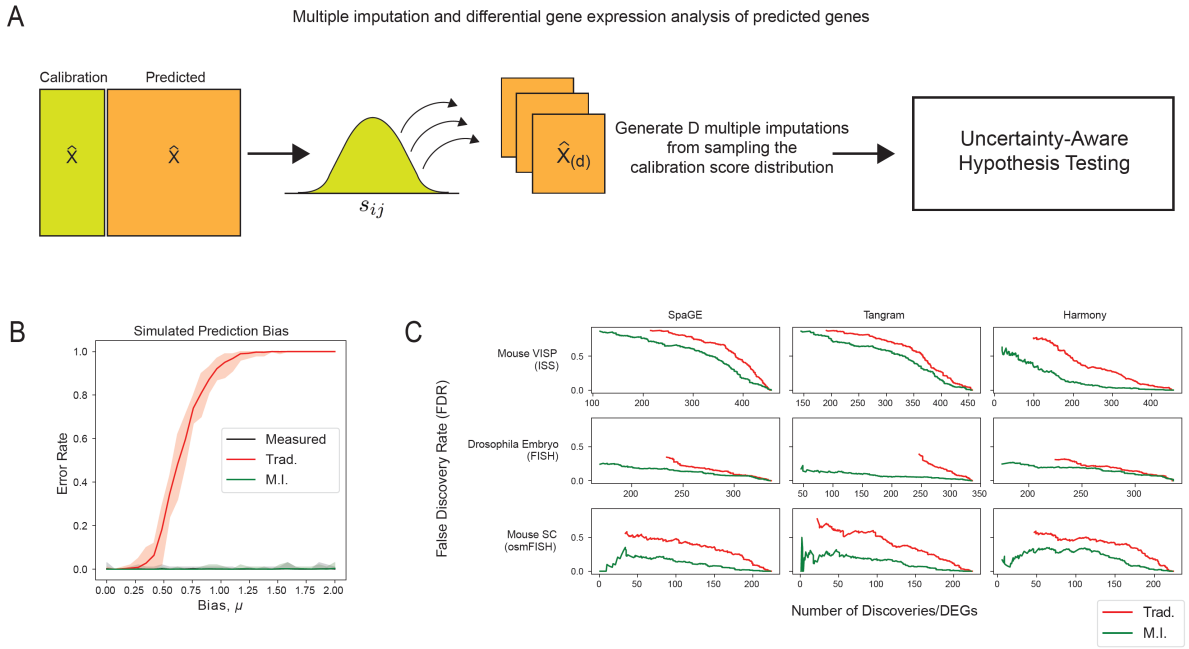
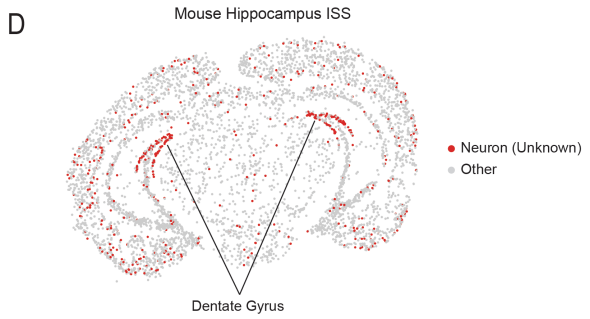


Figure 2: Prediction intervals for spatial gene expression. (A) Schematic illustration of the TISSUE prediction interval retrieval process from the calibration scores for a given confidence level. (B) Correlation of the 67% prediction interval width and the absolute prediction error across all dataset and prediction method combinations computed over 10-fold cross-validation. Log density with added pseudocount (Log1p) is shown by color. (C) Comparison of absolute prediction error (left) and the 67% prediction interval width (right) for a representative gene in the mouse somatosensory cortex osmFISH dataset; (D) in the virtual Drosophila embryo spatial transcriptomics dataset; (E) and in the mouse primary visual cortex MERFISH dataset. (F) Calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation. All prediction intervals were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping.





E

	Gene	SpaGE Trad.	SpaGE M.I.
Interneuron	*Sst*	n.s.	n.s.
	Vip	n.s.	n.s.
Peptidergic	*Hcrt*	**P < 1 E-20**	n.s.
	Agrp	n.s.	n.s.
	Pomc	**P = 4.3 E-5**	n.s.
MSN	*Drd1*	**P = 4.0 E-4**	n.s.
Serotonergic	*Tph2*	**P = 5.2 E-4**	n.s.
Mossy cell	*Calb2*	n.s.	n.s.
	Tm4sf1	n.s.	n.s.
NSC	*Prom1*	n.s.	n.s.
	Sox9	n.s.	n.s.
Granule cell	*Pdzd2*	**P < 1 E-20**	**P = 1.7 E-10**
	Gsg1l	**P < 1 E-20**	**P = 1.6 E-11**
	Grp	**P < 1 E-20**	**P = 7.2 E-10**

Figure 3: Uncertainty-aware differential gene expression analysis with TISSUE. (A) Schematic illustration of the TISSUE multiple imputation pipeline for hypothesis testing. Calibration scores are randomly sampled and used to compute new predicted gene expression profiles and statistics are compiled across all imputations using a modified two-sample t-test. (B) Error rate of significance testing of gene expression between a homogeneous group of cells ($n = 1000$, $p = 100$) as a function of the bias in prediction error for half of the genes in half of the cells using a Benjamini-Hochberg correction for 5% false discovery rate. Shown are error rates for two-sample t-test on the measured gene expression profiles (black), two-sample t-test on predicted gene expression (Trad., red), and modified two-sample t-test using the multiple imputation approach on predicted gene expression (M.I., green). Results were obtained using stratified grouping settings of $(k_g, k_c) = (2, 2)$. (C) False discovery rate of differentially expressed genes between cell type or anatomic region labels (one versus all approach) using the differentially expressed genes on the measured gene expression profiles as the ground truth across different p-value cutoffs. Shown are results for all three prediction methods and all spatial transcriptomics datasets with cell type or region labels available. All calibration scores were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping. (D) Mouse primary visual cortex ISS dataset with the unknown neuronal cluster colored in red and spatially localized to the dentate gyrus of the hippocampus. (E) Marker genes for the unknown neuronal cell cluster are differentially expressed for multiple neuronal cell type gene sets using traditional hypothesis testing with two-sample t-test on the predicted gene expression (SpaGE Trad.), but are selectively differentially expressed for granule cell marker genes using the modified two-sample t-test with multiple imputation from TISSUE calibration scores (SpaGE M.I.). P-values are shown for all predicted neuronal marker genes with significance threshold of Bonferroni-adjusted $p < 0.05$. All calibration scores were generated with $(k_g, k_c) = (4, 1)$ settings for stratified grouping.

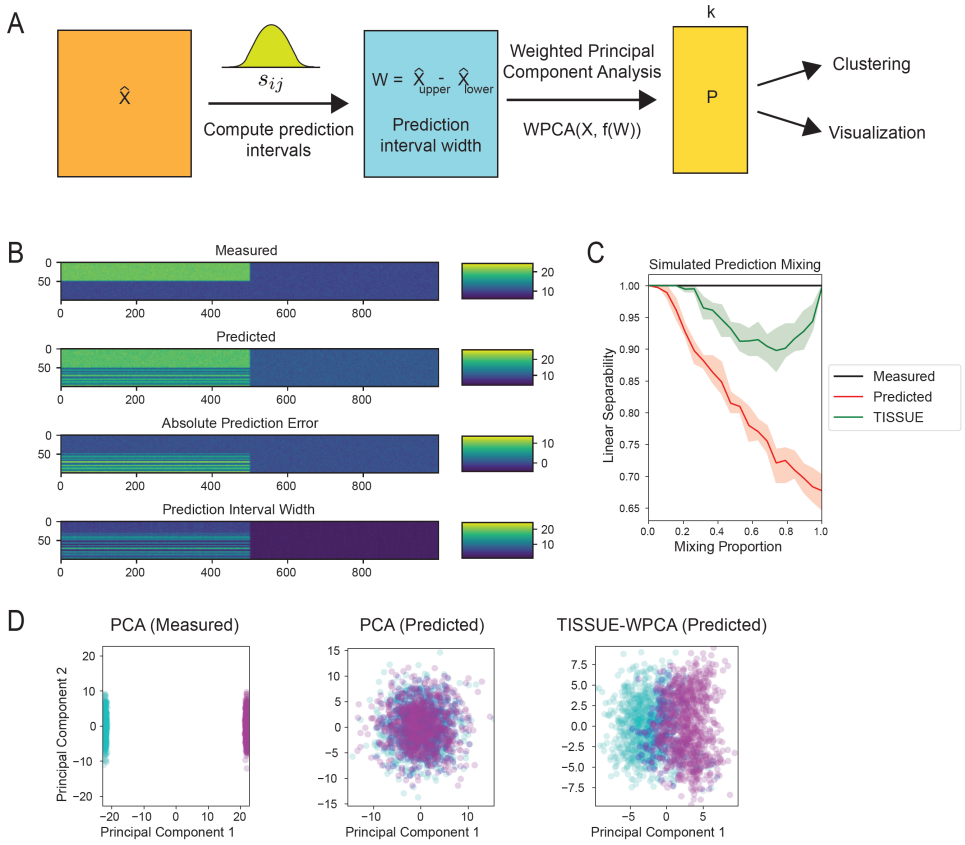


Figure 4: Uncertainty-aware clustering and label separation with TISSUE. (A) Schematic illustration of the weighted principal component analysis (WPCA) pipeline where the inverse TISSUE prediction interval width is used to obtain principal components from WPCA, which are then used for downstream tasks of clustering and label separation. (B) Gene expression profiles for the simulated data visualized by magnitude of expression with rows corresponding to cells and columns corresponding to genes. Shown are the measured gene expression profile simulated by Gaussian noise on top of the true gene expression signal from two distinct cell groups, the predicted gene expression profile with additional Gaussian noise and biased noise added to simulated mixing-in of the top cell cluster into the bottom cell cluster, and the absolute prediction error and 67% prediction interval width obtained through TISSUE. (C) Linear separability measured as the binary classification accuracy of a linear kernel support vector classifier on the two cell clusters as a function of the simulated mix-in proportion. The classifier was trained on the top 15 principal components obtained from the measured gene expression profiles with PCA, predicted gene expression profiles with PCA, and predicted gene expression profiles with TISSUE-WPCA. Results were obtained using stratified grouping settings of $(k_g, k_c) = (2, 2)$. (D) DynamicViz scatter plots of all cells by the first and second principal components from the measured gene expression profiles with PCA, predicted gene expression profiles with PCA, and predicted gene expression profiles with TISSUE-WPCA with complete mixing of the two synthetic cell clusters. DynamicViz visualization was generated by aligning the PCA embeddings from 20 synthetic datasets. Cells are colored to represent the two simulated clusters. Results were obtained using stratified grouping settings of $(k_g, k_c) = (2, 2)$.

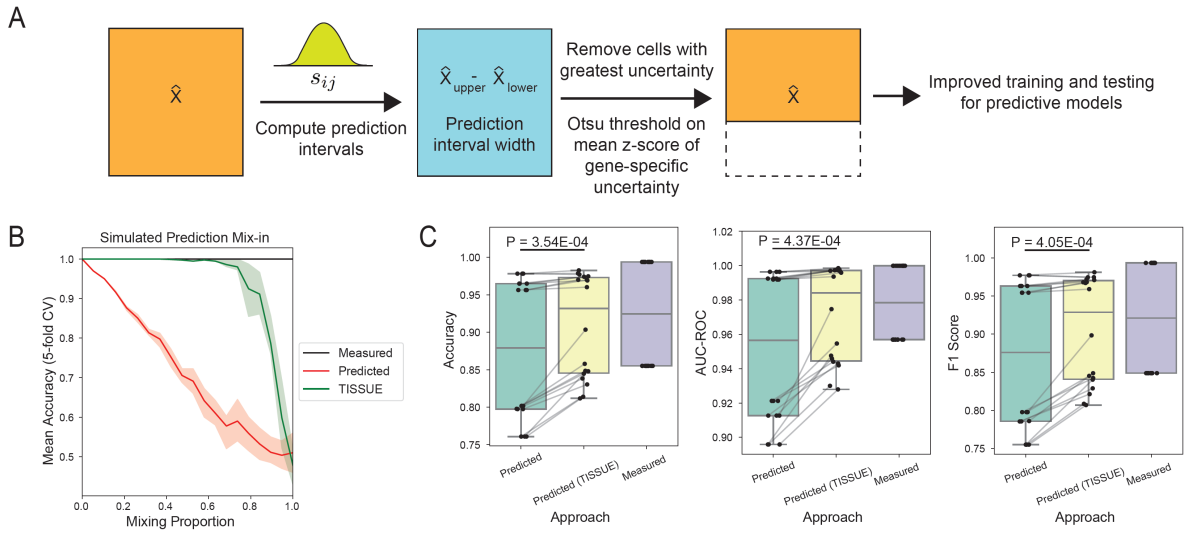


Figure 5: Uncertainty-aware predictive modeling with TISSUE. (A) Schematic illustration of the uncertainty-aware predictive modeling pipeline where the TISSUE prediction interval width is used to filter out cells with poor gene expression predictions from the dataset before training predictive models. (B) Mean accuracy for logistic regression models trained to classify the two cell clusters from the simulated dataset as a function of the mix-in bias. Accuracy was measured across 5-fold cross-validation for models trained and evaluated on the measured gene expression values (black), predicted gene expression values (red), and TISSUE-filtered predicted gene expression values (green). Results were obtained using stratified grouping settings of $(k_g, k_c) = (2, 2)$. (C) Accuracy, F1 score, and AUROC (area under the receiver-operator characteristic curve) metrics for logistic regression models trained on the predicted gene expression, TISSUE-filtered predicted gene expression, or measured gene expression for classification of the three most prominent class labels for the mouse somatosensory osmFISH dataset. Shown are metrics for all three prediction methods, three different stratified grouping settings ($(k_g, k_c) = (4, 1)$, $(k_g, k_c) = (4, 2)$, $(k_g, k_c) = (3, 3)$), and two sets of class labels (cell type and anatomic region). P-value was computed using a paired two-sample t-test.

5 Methods

5.1 Preprocessing of datasets

We followed data preprocessing approaches from prior benchmark comparisons of spatial gene prediction methods [24], which found highest predictive performance for these methods when non-normalized single-cell spatial transcriptomics data was paired with normalized single-cell RNAseq data. The RNAseq data was normalized using the Scanpy function pp.normalize_total() with its default settings followed by log transformation with an added pseudocount. We selected only genes expressed in at least one percent of cells in the RNAseq data.

5.2 Prediction of spatial gene expression

5.2.1 General framework for spatial gene expression prediction

The spatial gene prediction problem involves paired data from spatial transcriptomics and RNA-seq that are approximately from the same tissue and organism. We denote the spatial transcriptomics data as $X \in \mathbb{R}^{n \times p}$ and the RNA-seq data as $Y \in \mathbb{R}^{m \times q}$, where rows are cells and columns are genes. Generally, spatial gene prediction considers the case where $q >> p$ and the genes present in X are a subset of those in Y. A spatial gene prediction method predicts the expression of a gene that is present in Y but not in X for each cell in X using information from both X and Y.

5.2.2 Harmony

Harmony (as referred to in this work) involves joint embedding of the spatial and RNAseq data using the Harmony algorithm [22] followed by k-nearest-neighbor averaging to calculate predicted expression values for each spatial cell based on its nearest neighbors in the RNAseq data. We implemented the Harmony algorithm following the description outlined in previous applications [2]. We used default Harmony settings in the Scanpy external.pp.harmony_integrate() implementation. We averaged across the 10 nearest RNAseq neighbors for each spatial cell using the first $\min\{30, p\}$ Harmony principal components.

5.2.3 SpaGE

SpaGE performs spatial gene prediction using a two-step approach consisting of alignment using the domain adaptation algorithm PRECISE [35] and then performing k-nearest-neighbor regression [1]. We used a local download of the SpaGE algorithm available at `https://github.com/tabdelaal/SpaGE` with version corresponding to a download date of July 19, 2022. We set the number of principal vectors in SpaGE equal

to 20 if $p < 40$ and to $\min\{n, p\}/2$ rounded to the nearest integer if $p \geq 40$ and otherwise used the default settings.

5.2.4 Tangram

Tangram uses a deep learning framework to create a mapping for projecting RNAseq gene expression onto space [8]. We followed preprocessing details for Tangram according to previous benchmarks [24], which consisted of Leiden clustering on the scaled highly variable genes in the spatial data using Scanpy methods with default settings unless otherwise specified: pp.highly_variable_genes(), sc.pp.scale() with max_value=10, tl.pca(), pp.neighbors(), and tl.leiden() with resolution=0.5. After preprocessing, the identified clusters were used by Tangram to project the RNAseq cells onto space using map_cells_to_space() with mode='clusters' and density_prior='rna_count_based' and project_genes() with default settings.

5.3 Calibration scores for spatial gene expression prediction

5.3.1 Cross-validated spatial gene expression prediction

In order to compute calibration scores, we obtain estimated gene expression predictions on genes that are already measured in the spatial transcriptomics data. This is achieved through a cross-validation approach where a subset of the genes in the spatial transcriptomics data are left out and the gene prediction method is fitted to the remaining genes to make predictions on the left-out subset of genes. In practice, we use 10-fold cross-validation to obtain predictions for all genes in the spatial transcriptomics data but the TISSUE implementation provides options to customize the cross-validation procedure according to user specifications.

5.3.2 Cell-centric variability

We outline three desiderata to guide the development of a scalar uncertainty measure for spatial gene prediction:

1. To ensure computational scalability, the measure can be calculated on a single set of predicted gene expression values.
2. To accurately measure heterogeneity in prediction performance, the measure provides specific values for any cell and gene.
3. The measure ideally leverages spatial and gene expression similarity information.

We introduce the cell-centric variability to satisfy these desiderata. Specifically, for a cell i and gene j, the cell-centric variability U_{ij} is computed according to Eq. (1) and Eq. (2) using cells in its local neighborhood

N_i. We defined cell neighborhoods as the 15 nearest cells by Euclidean distance for each cell, and removed outliers by subsequently excluding neighbors with distance greater than $Q3 + 1.5 \times (Q3 - Q1)$, where $Q3$ and $Q1$ are the third and first quartiles of neighbor distances across all cell neighborhoods. The intercept term in Eq. (1) is included to ensure well-defined calibration scores and non-zero prediction interval widths for cells with no differences in gene expression across its neighborhood, which can result from the high sparsity of single-cell transcriptomics data. The weights W_{ik} in Eq. (2) are used to impose a soft weighting of the cell-centric variability for similar neighboring cells (i.e. of the same cell type) over dissimilar neighboring cells (i.e. of different cell types) without the need for user specification of discrete cell type clusters. The cell-centric variability can be computed for all cells and genes in both the calibration set (genes in the spatial transcriptomics data) and evaluation and test sets (genes to be predicted that are not in the spatial transcriptomics data).

5.3.3 Calculation of calibration scores from variability measure

To link the cell-centric variability to the prediction error, we compute the calibration score as shown in Eq. (3). We compute calibration scores for all pairs of cells and genes in the calibration set (i.e. present in spatial transcriptomics data) and allocate them to their corresponding stratification groups (see following section for details). Calibration scores are further separated by the sign of $X_{ij} - \hat{X}_{ij}$ to construct non-symmetric uncertainty bounds around the predicted expression value with $X_{ij} \geq \hat{X}_{ij}$ designating inclusion in the calibration scores set for the upper interval and $X_{ij} \leq \hat{X}_{ij}$ designating inclusion in the calibration scores set for the lower interval.

5.3.4 Stratified cell and gene grouping for calibration scores

In addition to context-specific construction of calibration scores, TISSUE can also provide finer groupings of genes and cells, each with their own calibration score sets. These stratified groupings are specified by the number of gene groups k_g and the number of cell groups k_c for a total for $k_g \times k_c$ groups. Stratified grouping is performed for genes first through k-means clustering with $k = k_g$ of the genes on the first 15 principal components representing the transposed predicted gene expression matrix. Then, within each of the identified gene strata, we perform further k-means clustering with $k = k_c$ of the cells on the first 15 principal components representing the predicted gene expression matrix restricted to genes present in that strata.

Since there is no guarantee that all stratified groups will contain genes in the calibration set or that all stratified groups will have adequate number of scores for calibration, for stratified groups with less than 100 scores in either the upper interval or lower interval calibration score sets, we defaulted the calibration score

set to the union of all calibration score sets across all stratified groups.

5.4 Conformal prediction intervals

5.4.1 Retrieval of prediction intervals from calibration scores

For a given confidence level α, we construct the prediction interval with approximate probability coverage $(1-\alpha)$ by retrieving the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantiles of the upper interval calibration scores and lower interval calibration scores. Referring to these quantiles as $q_\alpha^{(u)}$ and $q_\alpha^{(l)}$ respectively, the non-symmetric conformal prediction interval for the predicted gene expression of cell i and gene j can be computed as:

$$I_{ij} = \left[\hat{X}_{ij} - U_{ij} \cdot q_\alpha^{(l)}, \hat{X}_{ij} + U_{ij} \cdot q_\alpha^{(u)}\right] \tag{4}$$

Since this prediction interval does not explicitly depend on the measured prediction error, it can be calculated for all predicted gene expression values, even if the gene was not originally present in the calibration set.

5.4.2 Coverage guarantee for TISSUE prediction intervals

Under regularity conditions, the conformal inference framework provides consistent symmetric prediction intervals when applied to a scalar uncertainty measures such as cell-centric variability [5]. Building on that result, we show that this consistency is still valid with the non-symmetric prediction intervals that we compute using TISSUE.

Proposition 1 *Let $\left\{X_{[:,j]}\right\}_{j\in[1,\ldots,p,test]}$ be i.i.d. from some distribution, then $P\left(X_{ij} \in [l_{ij}, u_{ij}]\right) \geq 1-\alpha$ for any confidence level $0 \leq \alpha \leq 1$, where l_{ij} and u_{ij} are the quantiles of the lower and upper calibration score sets corresponding to X_{ij}.*

Here, 'test' refers to the index or set of indices for predicted genes that are not in the measured spatial transcriptomics data. Using the notation $l_{ij} = \hat{X}_{ij} - U_{ij} \cdot q_\alpha^{(l)}$ and $u_{ij} = \hat{X}_{ij} + U_{ij} \cdot q_\alpha^{(u)}$, the coverage of the TISSUE prediction interval for some confidence level α can be represented as follows:

$$\begin{aligned}
P\left(X_{ij} \in [l_{ij}, u_{ij}]\right) &= P\left(X_{ij} \in [l_{ij}, \hat{X}_{ij}]\right) + P\left(X_{ij} \in [\hat{X}_{ij}, u_{ij}]\right) \\
&= P\left(X_{ij} \in [l_{ij}, \hat{X}_{ij}] \mid X_{ij} < \hat{X}_{ij}\right) P\left(X_{ij} < \hat{X}_{ij}\right) + P\left(X_{ij} \in [\hat{X}_{ij}, u_{ij}] \mid X_{ij} > \hat{X}_{ij}\right) P\left(X_{ij} > \hat{X}_{ij}\right) \\
&\geq (1-\alpha)\left(P\left(X_{ij} < \hat{X}_{ij}\right) + P\left(X_{ij} > \hat{X}_{ij}\right)\right) \\
&\geq 1-\alpha,
\end{aligned} \tag{5}$$

with the first inequality following from Theorem 1 of [5]. And thus, given that symmetric intervals provide proper coverage [5], then we are also guaranteed proper coverage with the asymmetric prediction intervals produced by TISSUE, which is further evident through the empirical coverage assessment for TISSUE (Fig. 2F). This guarantee extends to the stratified group setting for $k_g > 1$ and/or $k_c > 1$ (see Proposition 2 in [5]).

5.4.3 Evaluation of prediction intervals

We evaluate the empirical coverage of the prediction intervals using 10-fold cross-validation where the evaluation set of genes in the spatial transcriptomics data is left out of the calibration set and using the calibration scores, we compute conformal prediction intervals for all cells and genes in the evaluation set. We then compute empirical coverage of the conformal prediction intervals as the fraction of measured gene expression values in the evaluation set that fall within the conformal prediction interval across all cells and genes and alternatively across all cells for each gene with nonzero predicted and measured expression. The empirical coverage is computed separately for the calibration set as a standard of comparison.

5.5 Uncertainty-aware hypothesis testing for predicted spatial gene expression

5.5.1 Generating multiple imputations using calibration scores

We introduce a multiple imputation procedure for generating uncertainty-aware hypothesis testing for predicted spatial gene expression. Multiple imputations are generated by uniformly sampling calibration scores from the corresponding union of upper and lower interval calibration score sets for each predicted spatial gene expression value. Given a uniform random sample of such calibration scores, $S_{(d)} \in \mathbb{R}^{n \times p}$, we compute an alternative imputation as:

$$\hat{X}_{(d)} = \hat{X} + \delta_{u/l,(d)} * S_{(d)} * U \tag{6}$$

where $*$ denotes element-wise multiplication, $U \in \mathbb{R}^{n \times p}$ is equal to the cell-centric variability measures computed on $\hat{X}$, and $\delta_{u/l} = 1$ if the score was sampled from the upper interval set and $\delta_{u/l} = -1$ if the score was sampled from the lower interval set. This sampling is repeated $D-1$ times to generate D multiple imputations including the original imputation $\hat{X}$. We tempered the multiple imputations against outliers by restricting the sampling to the scores within the set corresponding to the 80% conformal prediction interval.

5.5.2 Modified two-sample t-test for multiple imputation

After generating D multiple imputations from the calibration scores, we perform hypothesis testing using a modified two-sample t-test. Consider two groups with sets of sample/cell indices A and B. Then, the mean

difference and variance under normal two-sample t-test for a single imputation are:

$$\mu_d = \frac{1}{n_A} \sum_{i \in A} \hat{X}_{i,:}^{(d)} - \frac{1}{n_B} \sum_{i \in B} \hat{X}_{i,:}^{(d)} \tag{7}$$

$$s_d^2 = \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \left(\frac{(n_A - 1)\mathrm{Var}\{\hat{X}_{i,:}^{(d)} : i \in A\} + (n_B - 1)\mathrm{Var}\{\hat{X}_{i,:}^{(d)} : i \in B\}}{n_A + n_B - 2} \right), \tag{8}$$

where $\hat{X}^{(d)}$ denotes the d-th imputation among the D multiple imputations. Extending these statistical measures to the multiple imputation case, we use the standard modification for multiple imputation [37, 30, 3, 25], which results in the following mean and variance:

$$\mu_{MI} = \frac{1}{D} \sum_{d=1}^{D} \mu_d \tag{9}$$

$$s_{MI}^2 = s_W^2 + \left(1 + \frac{1}{D} \right) s_B^2 \tag{10}$$

where s_W^2 is the within-imputation variance and s_B^2 is the between-imputation variance, computed as:

$$s_W^2 = \frac{1}{D} \sum_{d=1}^{D} s_d^2 \tag{11}$$

$$s_B^2 = \frac{1}{D-1} \sum_{d=1}^{D} (\mu_d - \mu_{MI})^2. \tag{12}$$

Then, the modified test statistics for the two-sample t-test is:

$$\tilde{t}_{MI} = \frac{\mu_{MI}}{\sqrt{s_{MI}^2}}, \tag{13}$$

which is t-distributed with degrees of freedom $(D-1)\left(1 + \frac{D s_W^2}{(D+1) s_B^2}\right)^2$ and the resulting probability can be interpreted as the posterior probability of a significant difference in means between the two groups (i.e. $\beta \neq 0$) accounting for both evidence of this effect and the reliability of the imputations by inflating the standard error for this effect [37]. Under some regularity assumptions (approximate normality of imputed values and missing at random), the multiple imputation approach produces consistent estimates [3, 25].

5.6 Simulated prediction bias for differential gene expression analysis

We generated synthetic data for comparing the uncertainty-aware hypothesis testing approach against traditional hypothesis testing. First, we created a ground truth spatial gene expression matrix consisting of

$n = 100$ cells and $p = 1000$ genes with zero expression values for each instance. The first 50 cells were identified as under the first condition A and the remaining 50 cells were identified as under the second condition B. The spatial locations of the cells spanned a 10×10 grid where each cell was placed in ascending index across rows (left to right) and moving upwards (bottom to top). To simulate measurement error, we added standard Gaussian noise (mean equal to zero, variance equal to one) to the ground truth expression. Then, to simulate prediction bias for cells under condition B, we added shifted Gaussian noise (mean equal to $\mu \geq 0$, variance equal to one) to the measured values of the last 500 genes for all cells in condition B. Standard Gaussian noise was added to the measured values for the other cells. Under this simulation, there are no true differences between the cells in condition A as compared to condition B in the ground truth and virtually no differences detected in the measured gene expression value but differences are observed after prediction due to added bias. TISSUE, which is calibrated on prediction error of the calibration genes, is able to group and estimate more appropriate prediction intervals based on the spatial localization of these biased prediction errors.

5.7 Cell type annotation procedure for mouse VISP MERFISH dataset

We preprocessed the data using a standard Scanpy pipeline. Starting with the counts matrix, we normalized the data using pp.normalize_total() with default settings, log-transformed the data using pp.log1p(), and scaled the data with pp.scale(). We computed principal components and a neighbors graph using tl.pca() followed by pp.neighbors() with 20 principal components. Finally, we performed Leiden clustering using tl.leiden() with resolution of 0.1, which yielded 11 cell clusters. We used tl.rank_genes_groups() with the Wilcoxon method to identify the top five marker genes for each cell cluster and manually identified the clusters using those markers. In total, we identified endothelial cells, oligodendrocytes, astrocytes, and eight neuron-like cell clusters.

5.8 Anatomic region annotation procedure for Drosophila embryo dataset

We used the same preprocessing procedure as for the mouse VISP MERFISH dataset. We identified seven Leiden clusters and grouped them into four region labels based on their spatial localization with two "posterior" clusters, one "anterior" cluster, one "bottom" cluster, and three "middle" clusters.

5.9 Weighted PCA for uncertainty-aware clustering

We implemented a weighted version of principal component analysis where each value in the gene expression matrix is assigned a scalar weight. We compute the weights according to the following steps. First, we

compute the inverse of the TISSUE prediction interval width (i.e. 67% prediction interval upper bound minus lower bound). Then, we normalize these values for each gene by the mean value across that gene to correct for expression level differences between genes. Finally, we binarize these normalized values so that the top 80% of normalized values will have ten-fold higher relative weight than the bottom 20% of normalized values. These binary values are used as weights for Weighted Principal Component Analysis (WPCA). WPCA directly decomposes the weighted covariance matrix to obtain principal vectors, and then applies weighted least squares optimization to retrieve the principal components [14]. We used the implementation of WPCA in the wpca (v.0.1) Python package with default settings and weights set according to our specification.

To compute separability or clustering results from the WPCA principal components, we used the first 15 principal components as input either in the Leiden clustering algorithm (using Scanpy's pp.neighbors() followed by tl.leiden() with default settings) or in fitting a support vector classifier with linear kernel (using Scikit-Learn's svm.SVC() with kernel='linear'). To obtain ground truth labels for the Leiden clustering on the measured gene expression, we normalized and log-transformed the gene expression values and then performed PCA and Leiden clustering using the same specifications as before.

5.10 Dynamic visualization of first two principal components

We generated dynamic visualizations of the first two principal components for visual comparison of PCA on the measured spatial gene expression, PCA on the predicted spatial gene expression, and WPCA on the predicted spatial gene expression. We used DynamicViz (v.0.0.3) to center and rigidly align the cells across 20 synthetic datasets, visualized the resulting alignments using viz.stacked(), and scored their variability by computing variance scores for each cell using score.variance() with method='global'.

5.11 Simulated prediction bias for clustering and predictive modeling

We generated synthetic data for benchmarking the uncertainty-aware clustering/separation and predictive modeling approaches. First, we created a ground truth spatial gene expression matrix consisting of $n = 100$ cells and $p = 1000$ genes with zero expression values for each instance. The first 50 cells were identified as cell type A and the remaining 50 cells were identified as cell type B. All cells of cell type A have ground truth expression of 20 for the first 500 genes and expression of 10 for the other 500 genes; cells of cell type B have ground truth expression of 10 for all 1000 genes. The spatial locations of the cells spanned a 10×10 grid where each cell was placed in ascending index across rows (left to right) and moving upwards (bottom to top). To simulate measurement error, we added standard Gaussian noise (mean equal to zero, variance equal to one) to the ground truth expression. Then, to simulate prediction bias for cells in cell type B, we

added shifted Gaussian noise (mean equal 10, variance equal to one) to the first 500 genes of a proportion of the cells in cell type B sampled randomly and determined by the specified 'mix-in' proportion. For the sampled cells, this simulated prediction bias shifts their predicted gene expression profiles to be more similar to that of cell type A rather than cell type B. Standard Gaussian noise was added to all other measured values for both cell types.

5.12 Uncertainty-aware cell filtering for predictive modeling

Using the TISSUE prediction interval, we perform filtering of cells in training and/or evaluation data for predictive modeling. First, we convert all gene expression values to z-scores using the mean and standard deviation of expression for each gene in the data. Then, for each cell, we assign a score based on its average z-score across all genes. The cells with the highest scores are removed from the filtered data. The threshold for removal is automatically determined using Otsu's method, which finds a threshold that maximizes the variance between the filtered and unfiltered score sets. In the context of classification, we avoid inter-class differences in prediction uncertainty by performing this filtering procedure independently within each class.

Data Availability

Raw data were accessed from existing benchmark datasets [24] and are available from the following studies:

Mouse Hippocampus: Spatial transcriptomics (seqFISH) at `https://content.cruk.cam.ac.uk/jmlab/SpatialMouseAtlas2020/`; RNAseq (10X Chromium) at GSE158450 in GEO for 'HIPP_sc_Rep1_10X sample'.

Mouse VISP: Spatial transcriptomics (MERFISH) at `https://github.com/spacetx-spacejam/data/`; RNAseq (Smart-seq) at `https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq` for mouse primary visual cortex (VISp).

Mouse Prefrontal Cortex (PC): Spatial transcriptomics (STARmap) at '20180419_BZ9_control' in `https://www.starmapresources.com/data`; RNAseq (10X Chromium) at GSE158450 in GEO for 'PFC_sc_Rep2_10X'.

Human Middle Temporal Gyrus (MTG): Spatial transcriptomics (ISS) at `https://github.com/spacetx-spacejam/data`; RNAseq (Smart-seq) at `https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq`.

Mouse VISP: Spatial transcriptomics (ISS) at https://github.com/spacetx-spacejam/data; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex (VISp).

Drosophila Embryo: Spatial transcriptomics (FISH) at https://github.com/rajewsky-lab/distmap; RNAseq (Drop-seq) at GSE95025 in GEO.

Mouse Somatosensory Cortex (SC): Spatial transcriptomics (osmFISH) at http://linnarssonlab.org/osmFISH/ for cortical region subset; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-smart-seq for mouse somatosensory cortex (SSp).

Mouse VISP: Spatial transcriptomics (ExSeq) at https://github.com/spacetx-spacejam/data; RNAseq (Smart-seq) at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq for mouse primary visual cortex (VISp).

Code Availability

The TISSUE Python package and associated code and documentation are available at https://github.com/sunericd/TISSUE, and all code for generating figures and analyses are separately available at https://github.com/sunericd/tissue-figures-and-analyses.

Acknowledgements

Funding support was provided by Knight-Hennessy Scholars program (E.D.S.), Paul and Daisy Soros Fellowship for New Americans (E.D.S.), the National Science Foundation Graduate Research Fellowship Program (E.D.S.), Professor David Donoho at Stanford University (R.M.), NSF CAREER 1942926 (J.Z.), NIH P30AG059307 (J.Z.), 5RM1HG010023 (J.Z.) and grants from the Silicon Valley Foundation (J.Z.) and the Chan-Zuckerberg Initiative (J.Z.). We would like to thank M. Yuksekgonul for helpful discussions.

Author's contributions

E.D.S. and J.Z. conceived of the study. E.D.S. designed and implemented the method and ran all associated analyses with input from J.Z. and R.M. E.D.S. prepared a draft of the manuscript. J.Z. and R.M. edited the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Tamim Abdelaal, Soufiane Mourragui, Ahmed Mahfouz, and Marcel J T Reinders. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Research*, 48(18):e107, October 2020.

[2] William E. Allen, Timothy R. Blosser, Zuri A. Sullivan, Catherine Dulac, and Xiaowei Zhuang. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell*, 186(1):194–208.e18, January 2023. Publisher: Elsevier.

[3] Paul D. Allison. *Missing Data*. SAGE Publications, Inc., 2002.

[4] Shahar Alon, Daniel R. Goodwin, Anubhav Sinha, Asmamaw T. Wassie, Fei Chen, Evan R. Daugharthy, Yosuke Bando, Atsushi Kajita, Andrew G. Xue, Karl Marrett, Robert Prior, Yi Cui, Andrew C. Payne, Chun-Chen Yao, Ho-Jun Suk, Ru Wang, Chih-Chieh (Jay) Yu, Paul Tillberg, Paul Reginato, Nikita Pak, Songlei Liu, Sukanya Punthambaker, Eswar P. R. Iyer, Richie E. Kohman, Jeremy A. Miller, Ed S. Lein, Ana Lako, Nicole Cullen, Scott Rodig, Karla Helvie, Daniel L. Abravanel, Nikhil Wagle, Bruce E. Johnson, Johanna Klughammer, Michal Slyper, Julia Waldman, Judit Jané-Valbuena, Orit Rozenblatt-Rosen, Aviv Regev, IMAXT Consortium19, George M. Church, Adam H. Marblestone, and Edward S. Boyden. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, 371(6528):eaax2656, January 2021. Publisher: American Association for the Advancement of Science.

[5] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, September 2022. arXiv:2107.07511 [cs, math, stat].

[6] Benedetta Artegiani, Anna Lyubimova, Mauro Muraro, Johan H. van Es, Alexander van Oudenaarden, and Hans Clevers. A Single-Cell RNA Sequencing Study Reveals Cellular and Molecular Dynamics of the Hippocampal Neurogenic Niche. *Cell Reports*, 21(11):3271–3284, December 2017.

[7] Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegård, Fredrik Salmén, Cecilia Österholm, Patrik L. Ståhl, Erik Sundström, Elisabet Åkesson, Olaf Bergmann, Magda Bienko, Agneta Månsson-Broberg, Mats Nilsson, Christer Sylvén, and Joakim Lundeberg. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell*, 179(7):1647–1660.e19, December 2019.

[8] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11):1352–1362, November 2021. Number: 11 Publisher: Nature Publishing Group.

[9] A. Sina Booeshaghi, Zizhen Yao, Cindy van Velthoven, Kimberly Smith, Bosiljka Tasic, Hongkui Zeng, and Lior Pachter. Isoform cell-type specificity in the mouse primary motor cortex. *Nature*, 598(7879):195–199, October 2021. Number: 7879 Publisher: Nature Publishing Group.

[10] Matthew T. Buckley, Eric D. Sun, Benson M. George, Ling Liu, Nicholas Schaum, Lucy Xu, Jaime M. Reyes, Margaret A. Goodell, Irving L. Weissman, Tony Wyss-Coray, Thomas A. Rando, and Anne Brunet. Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. *Nature Aging*, 3(1):121–137, January 2023. Number: 1 Publisher: Nature Publishing Group.

[11] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1):2084, April 2020. Number: 1 Publisher: Nature Publishing Group.

[12] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, April 2015. Publisher: American Association for the Advancement of Science.

[13] Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, November 2018. Number: 11 Publisher: Nature Publishing Group.

[14] L. Delchambre. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society*, 446(4):3545–3555, February 2015.

[15] Daniel Edsgärd, Per Johnsson, and Rickard Sandberg. Identification of spatial expression trends in single-cell gene expression data. *Nature methods*, 15(5):339–342, May 2018.

[16] Daniel Gyllborg, Christoffer Mattsson Langseth, Xiaoyan Qian, Eunkyoung Choi, Sergio Marco Salas, Markus M Hilscher, Ed S Lein, and Mats Nilsson. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Research*, 48(19):e112, November 2020.

[17] Rebecca D. Hodge, Trygve E. Bakken, Jeremy A. Miller, Kimberly A. Smith, Eliza R. Barkan, Lucas T. Graybuck, Jennie L. Close, Brian Long, Nelson Johansen, Osnat Penn, Zizhen Yao, Jeroen Eggermont, Thomas Höllt, Boaz P. Levi, Soraya I. Shehata, Brian Aevermann, Allison Beller, Darren Bertagnolli, Krissy Brouner, Tamara Casper, Charles Cobbs, Rachel Dalley, Nick Dee, Song-Lin Ding, Richard G. Ellenbogen, Olivia Fong, Emma Garren, Jeff Goldy, Ryder P. Gwinn, Daniel Hirschstein, C. Dirk Keene, Mohamed Keshk, Andrew L. Ko, Kanan Lathia, Ahmed Mahfouz, Zoe Maltzer, Medea McGraw, Thuc Nghi Nguyen, Julie Nyhus, Jeffrey G. Ojemann, Aaron Oldre, Sheana Parry, Shannon Reynolds, Christine Rimorin, Nadiya V. Shapovalova, Saroja Somasundaram, Aaron Szafer, Elliot R. Thomsen, Michael Tieu, Gerald Quon, Richard H. Scheuermann, Rafael Yuste, Susan M. Sunkin, Boudewijn Lelieveldt, David Feng, Lydia Ng, Amy Bernard, Michael Hawrylycz, John W. Phillips, Bosiljka Tasic, Hongkui Zeng, Allan R. Jones, Christof Koch, and Ed S. Lein. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68, September 2019. Number: 7772 Publisher: Nature Publishing Group.

[18] Andrew L. Ji, Adam J. Rubin, Kim Thrane, Sizun Jiang, David L. Reynolds, Robin M. Meyers, Margaret G. Guo, Benson M. George, Annelie Mollbrink, Joseph Bergenstråhle, Ludvig Larsson, Yunhao Bai, Bokai Zhu, Aparna Bhaduri, Jordan M. Meyers, Xavier Rovira-Clavé, S. Tyler Hollmig, Sumaira Z. Aasi, Garry P. Nolan, Joakim Lundeberg, and Paul A. Khavari. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell*, 182(2):497–514.e22, July 2020.

[19] Anoushka Joglekar, Andrey Prjibelski, Ahmed Mahfouz, Paul Collier, Susan Lin, Anna Katharina Schlusche, Jordan Marrocco, Stephen R. Williams, Bettina Haase, Ashley Hayes, Jennifer G. Chew, Neil I. Weisenfeld, Man Ying Wong, Alexander N. Stein, Simon A. Hardwick, Toby Hunt, Qi Wang, Christoph Dieterich, Zachary Bent, Olivier Fedrigo, Steven A. Sloan, Davide Risso, Erich D. Jarvis, Paul Flicek, Wenjie Luo, Geoffrey S. Pitt, Adam Frankish, August B. Smit, M. Elizabeth Ross, and

Hagen U. Tilgner. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nature Communications*, 12(1):463, January 2021.

[20] Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P. Zinzen. The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, October 2017. Publisher: American Association for the Advancement of Science.

[21] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, September 2013. Number: 9 Publisher: Nature Publishing Group.

[22] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019. Number: 12 Publisher: Nature Publishing Group.

[23] P R Langer-Safer, M Levine, and D C Ward. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79(14):4381–4385, July 1982.

[24] Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, Ke Liu, Xuetong Zhao, Jun Lin, Linzhao Cheng, Falai Chen, Tian Xue, and Kun Qu. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19(6):662–670, June 2022. Number: 6 Publisher: Nature Publishing Group.

[25] Roderick J. A. Little and Donald B. Rubin. Bayes and Multiple Imputation. In *Statistical Analysis with Missing Data*, pages 200–220. John Wiley & Sons, Ltd, 2002. Section: 10 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119013563.ch10.

[26] Brian Long, Jeremy Miller, and The SpaceTx Consortium. SpaceTx: A Roadmap for Benchmarking Spatial Transcriptomics Exploration of the Brain, January 2023. arXiv:2301.08436 [q-bio].

[27] Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements, May 2019. arXiv:1905.02269 [cs, q-bio, stat].

[28] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, August 2019. Number: 8 Publisher: Nature Publishing Group.

[29] Eric Lubeck, Ahmet F. Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4):360–361, April 2014. Number: 4 Publisher: Nature Publishing Group.

[30] Andrea Marshall, Douglas G Altman, Roger L Holder, and Patrick Royston. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology*, 9:57, July 2009.

[31] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, March 2021. Number: 3 Publisher: Nature Publishing Group.

[32] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C. Devlin, Maayan Baron, Cristina H. Hajdu, Diane M. Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, March 2020.

[33] Noa Moriel, Enes Senel, Nir Friedman, Nikolaus Rajewsky, Nikos Karaiskos, and Mor Nitzan. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9):4177–4200, September 2021. Number: 9 Publisher: Nature Publishing Group.

[34] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, pages 1–13, March 2022. Publisher: Nature Publishing Group.

[35] Soufiane Mourragui, Marco Loog, Mark A van de Wiel, Marcel J T Reinders, and Lodewyk F A Wessels. PRECISE: a domain adaptation approach to transfer predictors of drug response from preclinical models to tumors. *Bioinformatics*, 35(14):i510–i519, July 2019.

[36] Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, December 2019. Number: 7785 Publisher: Nature Publishing Group.

[37] Cameron Palmer and Itsik Pe'er. Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. *PLOS Genetics*, 12(6):e1006091, June 2016. Publisher: Public Library of Science.

[38] Antonio Scialdone, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, September 2015.

[39] Chen Shengquan, Zhang Boheng, Chen Xiaoyang, Zhang Xuegong, and Jiang Rui. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics*, 37(Supplement_1):i299–i307, July 2021.

[40] Tabrez J. Siddiqui, Parisa Karimi Tari, Steven A. Connor, Peng Zhang, Frederick A. Dobie, Kevin She, Hiroshi Kawabe, Yu Tian Wang, Nils Brose, and Ann Marie Craig. An LRRTM4-HSPG Complex Mediates Excitatory Synapse Development on Dentate Gyrus Granule Cells. *Neuron*, 79(4):680–695, August 2013.

[41] Eric D. Sun, Rong Ma, and James Zou. Dynamic visualization of high-dimensional data. *Nature Computational Science*, 3(1):86–100, January 2023. Number: 1 Publisher: Nature Publishing Group.

[42] Bosiljka Tasic, Zizhen Yao, Lucas T. Graybuck, Kimberly A. Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N. Economo, Sarada Viswanathan, Osnat Penn, Trygve Bakken, Vilas Menon, Jeremy Miller, Olivia Fong, Karla E. Hirokawa, Kanan Lathia, Christine Rimorin, Michael Tieu, Rachael Larsen, Tamara Casper, Eliza Barkan, Matthew Kroll, Sheana Parry, Nadiya V. Shapovalova, Daniel Hirschstein, Julie Pendergraft, Heather A. Sullivan, Tae Kyung Kim, Aaron Szafer, Nick Dee, Peter Groblewski, Ian Wickersham, Ali Cetin, Julie A. Harris, Boaz P. Levi, Susan M. Sunkin, Linda Madisen, Tanya L. Daigle, Loren Looger, Amy Bernard, John Phillips, Ed Lein, Michael Hawrylycz, Karel Svoboda, Allan R. Jones, Christof Koch, and Hongkui Zeng. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, November 2018. Number: 7729 Publisher: Nature Publishing Group.

[43] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, March 2019. Number: 1 Publisher: Nature Publishing Group.

[44] Milad R. Vahid, Erin L. Brown, Chloé B. Steen, Wubing Zhang, Hyun Soo Jeon, Minji Kang, Andrew J. Gentles, and Aaron M. Newman. High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. *Nature Biotechnology*, pages 1–6, March 2023. Publisher: Nature Publishing Group.

[45] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, July 2018. Publisher: American Association for the Advancement of Science.

[46] Runmin Wei, Siyuan He, Shanshan Bai, Emi Sei, Min Hu, Alastair Thompson, Ken Chen, Savitri Krishnamurthy, and Nicholas E. Navin. Spatial charting of single-cell transcriptomes in tissues. *Nature Biotechnology*, 40(8):1190–1199, August 2022. Number: 8 Publisher: Nature Publishing Group.

[47] Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, June 2019.

[48] Zizhen Yao, Cindy T. J. van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E. Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, Song-Lin Ding, Olivia Fong, Emma Garren, Alexandra Glandon, Nathan W. Gouwens, James Gray, Lucas T. Graybuck, Michael J. Hawrylycz, Daniel Hirschstein, Matthew Kroll, Kanan Lathia, Changkyu Lee, Boaz Levi, Delissa McMillen, Stephanie Mok, Thanh Pham, Qingzhong Ren, Christine Rimorin, Nadiya Shapovalova, Josef Sulc, Susan M. Sunkin, Michael Tieu, Amy Torkelson, Herman Tung, Katelyn Ward, Nick Dee, Kimberly A. Smith, Bosiljka Tasic, and Hongkui Zeng. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.e26, June 2021.

6 Extended Data Figures

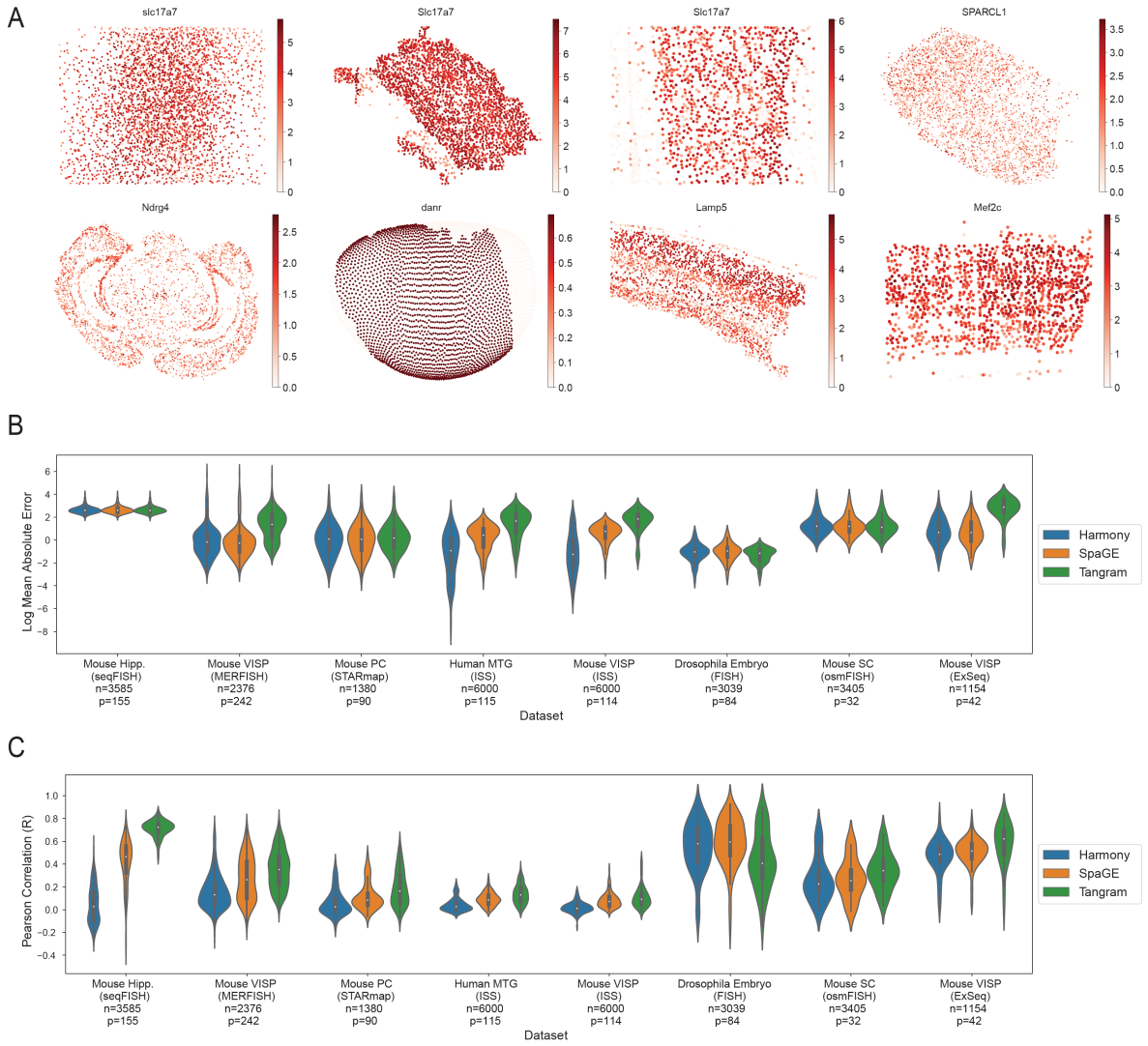


Figure S1: Overview of datasets and prediction performance. (A) Visualization of cells in the eight spatial transcriptomics datasets colored by the expression of the highest-expressed gene in each respective dataset. (B-C) Performance of all three gene prediction methods (Harmony, SpaGE, Tangram) on all eight datasets as measured by: (B) mean absolute error between predicted and actual gene expression over 10-fold cross-validation, and (C) Pearson correlation between predicted and actual gene expression over 10-fold cross-validation. Shown also are the number of cells (n) and number of genes (p) in the spatial transcriptomics datasets.

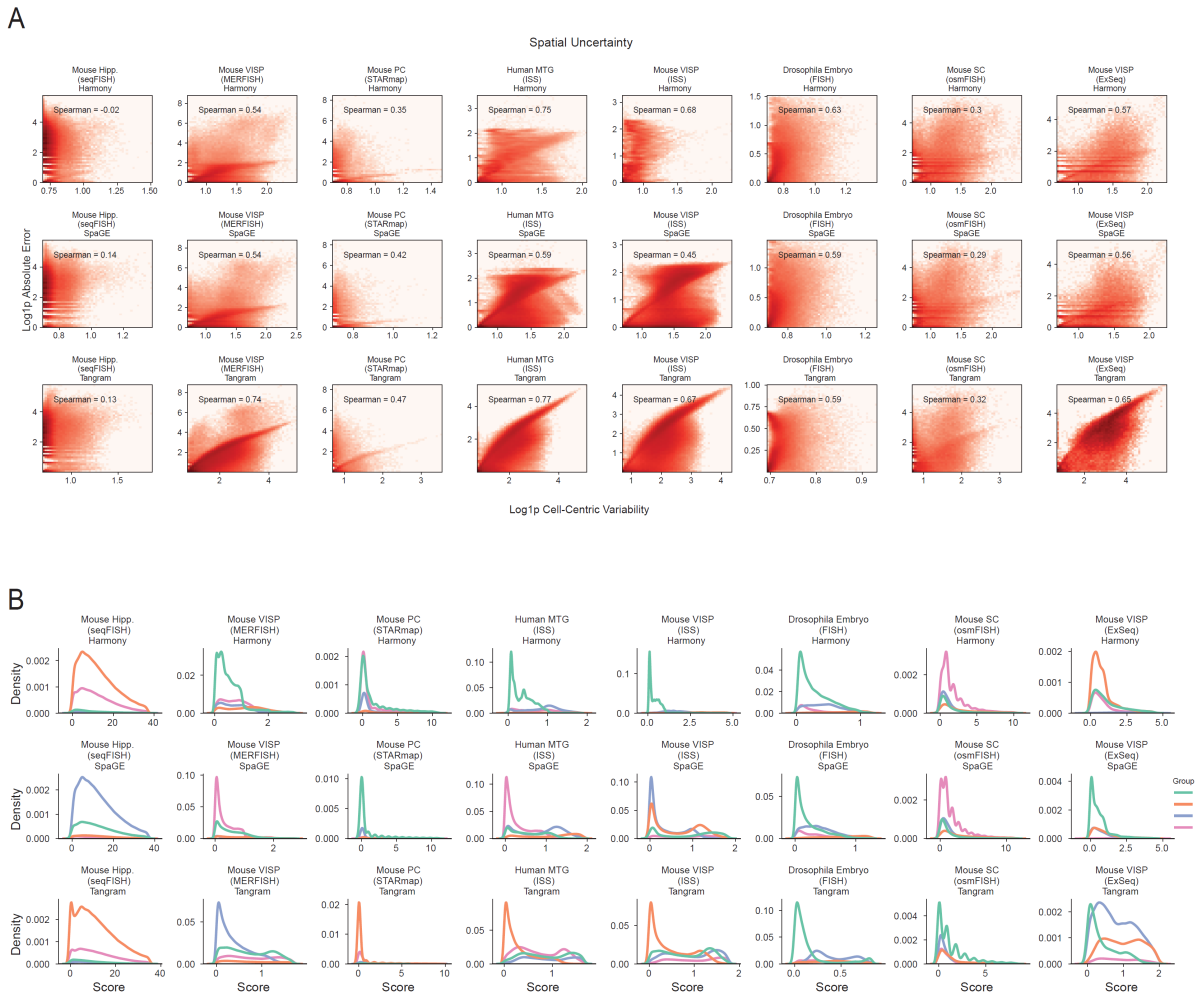


Figure S2: Individual cell-centric variability and calibration score distributions for all datasets and prediction methods. (A) Correlation of cell-centric variability and absolute prediction error shown individually for each dataset and prediction method combination computed over 10-fold cross-validation. Log density with added pseudocount (Log1p) is shown by color. (D) Distribution of TISSUE calibration scores shown individually for each dataset and prediction method combination ($(k_g, k_c) = (4, 1)$). Details on each dataset and prediction method can be found in Methods.

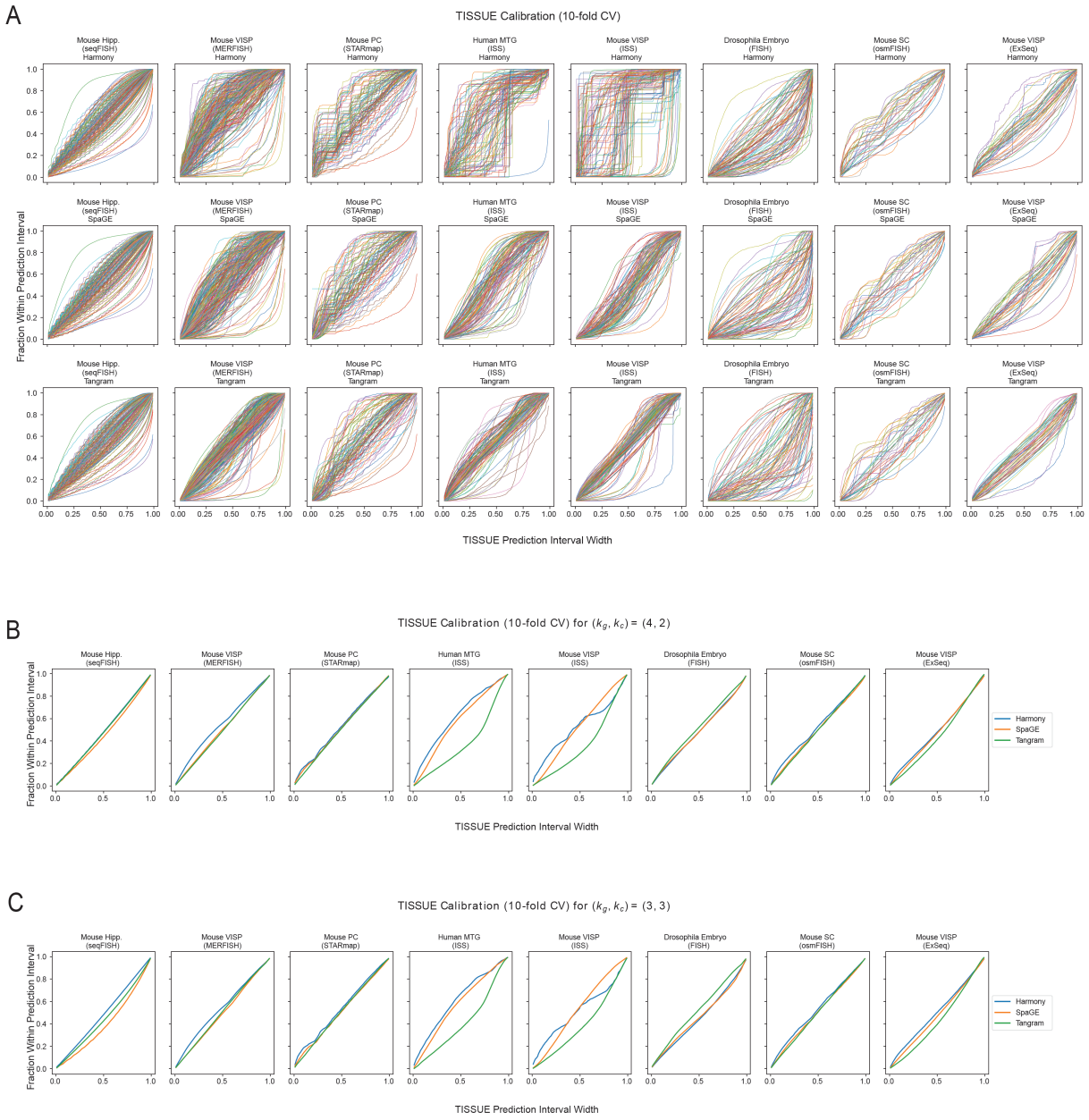


Figure S3: Additional prediction intervals for spatial gene expression. (A) Gene-level calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation. Each line corresponds to an independent gene in the spatial transcriptomics dataset. (B-C) Calibration curves for TISSUE prediction intervals showing empirical coverage as a function of the specified confidence level across 10-fold cross-validation for (B) $(k_g, k_c) = (4, 2)$ settings for stratified grouping, and (C) $(k_g, k_c) = (3, 3)$ settings for stratified grouping.

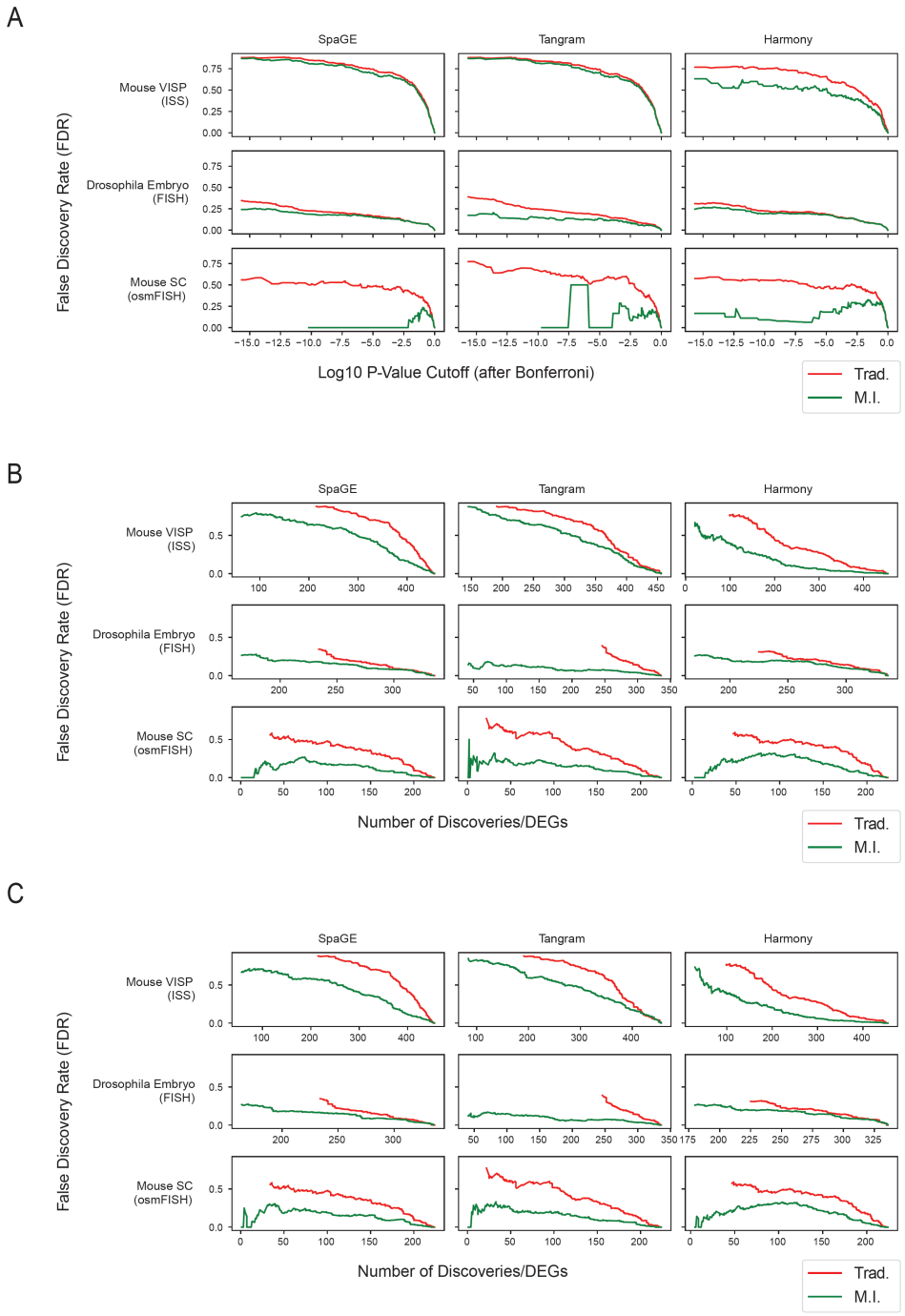


Figure S4: Additional differential gene expression analysis with TISSUE. (A) False discovery rate of differentially expressed genes between cell type or anatomic region labels (one versus all approach) as a function of the Bonferroni-corrected p-value threshold for significance and with $(k_g, k_c) = (4, 1)$ settings for stratified grouping. (B-C) False discovery rate of differentially expressed genes between cell type or anatomic region labels (one versus all approach) as a function of the number of discoveries and with (B) $(k_g, k_c) = (4, 2)$ settings for stratified grouping, and (C) $(k_g, k_c) = (3, 3)$ settings for stratified grouping.

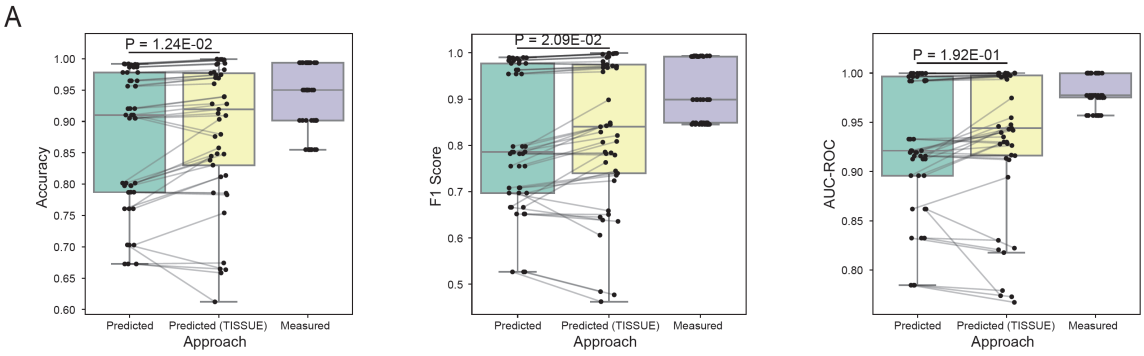


Figure S5: Aggregate performance of TISSUE-filtering for predictive modeling. (A) Accuracy, F1 score, and AUROC (area under the receiver-operator characteristic curve) metrics for logistic regression models trained on the predicted gene expression, TISSUE-filtered predicted gene expression, or measured gene expression for classification of the three most prominent class labels for three datasets. Shown are metrics for all three prediction methods, three different stratified grouping settings ($(k_g, k_c) = (4, 1)$, $(k_g, k_c) = (4, 2)$, $(k_g, k_c) = (3, 3)$), and two sets of class labels (cell type and anatomic region). P-value was computed using a paired two-sample t-test.

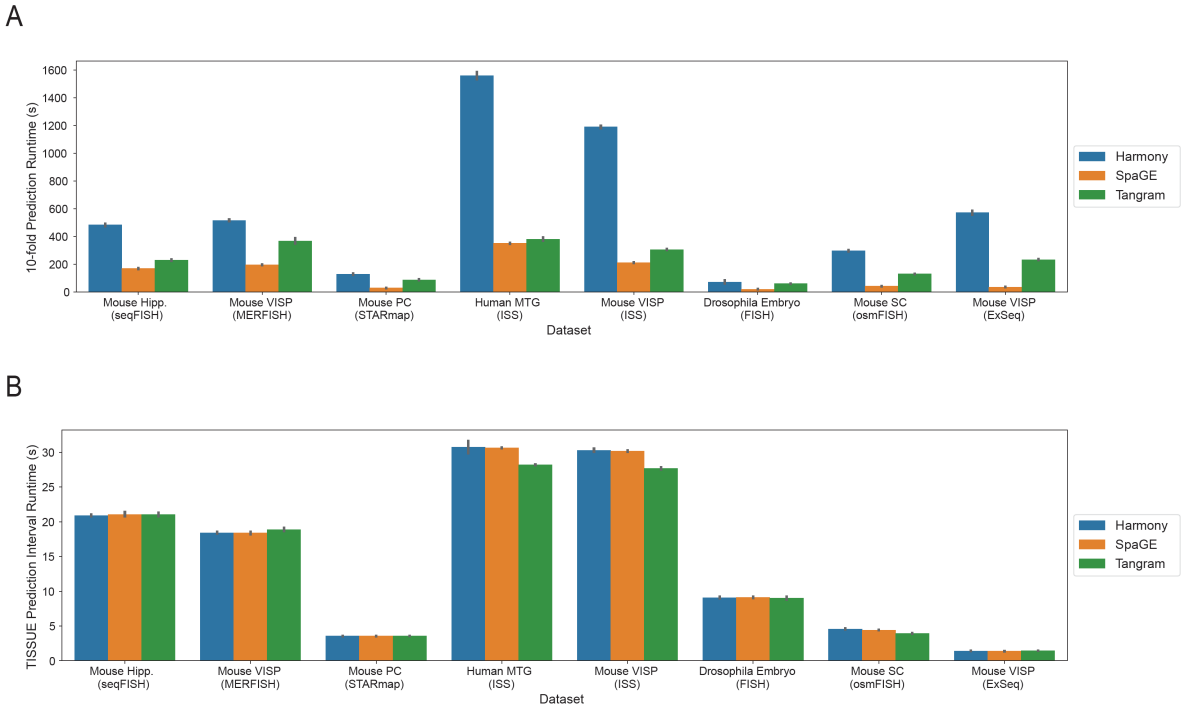


Figure S6: Computational runtime for TISSUE. (A) Bar plots of total runtimes for spatial gene expression prediction computations over 10 predictions to generate estimated predictions on all calibration genes. Error bars correspond to 95% confidence interval over an outer 10-fold cross-validation. (B) Bar plots of total runtimes for TISSUE prediction interval calculation including computation of cell-centric variability and calibration score sets. Error bars correspond to 95% confidence interval over an outer 10-fold cross-validation.