Tissue Reassembly with Generative AI

Tingyang Yu[1], Chanakya Ekbote[1], Nikita Morozov[1], Antonio Herrera[4], Jiashuo Fan[1], Albert Dominguez Mantes[4], Salvatore Novello[4], Hilal Lashuel[4], Pascal Frossard[2], Stéphane d'Ascoli[1,3], Gioele La Manno[4] and Maria Brbić[1,4,5†]

[1] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

[2] School of Engineering, EPFL, Lausanne, Switzerland

[3] Meta AI, Paris, France

[4] School of Life Sciences, EPFL, Lausanne, Switzerland

[5] Swiss Institute of Bioinformatics, Switzerland

[†]Corresponding author. Email: mbrbic@epfl.ch

The spatial arrangement of cells is fundamental to their function, but single-cell RNA sequencing loses spatial context by dissociating cells from tissues. We present LUNA, a generative AI model that reassembles dissociated cells into tissue structures solely from gene expression by learning spatial priors from existing spatially resolved datasets. We apply and validate LUNA across multiple technologies including reconstructing the MERFISH whole mouse brain atlas with over 1.2 million cells, de novo reassembly of mouse central nervous system scRNA-seq atlas and inference of the spatial locations of nuclei lost during Slide-tags profiling. Furthermore, we show that LUNA generalizes to unseen cell types and to pathological samples in the Parkinson's disease mouse model, identifying regions of change due to pathology. We envision that AI-driven tissue reassembly can help to overcome current technological limitations and advance our understanding of tissue organization and function.

Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have enabled the high-throughput profiling of cellular transcriptomes, making it possible to identify and explore unique transcriptional profiles within individual cells [1, 2]. These technologies have revolutionized our understanding of cellular diversity, providing unprecedented insights into the complexity of cells and tissues [3–8]. However, a major limitation of scRNA-seq technologies is the loss of spatial context, as cells are dissociated from their native microenvironments during the sequencing process. The spatial context is crucial for understanding how cells interact with their neighbors and the surrounding microenvironment, influencing their roles in healthy tissue function and disease development. Although spatially resolved sequencing technologies aim to address this limitation [9–11], they are restricted in the number of genes that can be measured and are not always an option.

To address the limitations of individual technologies, computational methods have been developed to infer spatial context from gene expression of single-cells. Early approaches relied on in situ hybridization (ISH)-based reference atlases of landmark gene expression and leveraged them to infer the spatial location of cells in scRNA-seq datasets [12, 13]. More recently, reference mapping methods such as Tangram [14], CytoSPACE [15], STEM [16], scSpace [17], STALocator [18] and CeLEry [19], proposed mapping dissociated tissue data to reference spatial datasets. Although these methods have shown promising results, they rely on an exact spatial reference for mapping scRNA-seq datasets, thus primarily functioning as data integration tools. On the other hand, novoSpaRc [20, 21] enabled a *de novo* spatial reconstruction of single-cell gene expression. However, novoSpaRc assumes that cells in close physical proximity exhibit similar gene expression profiles — an assumption that is often violated in complex tissue architectures.

Here, we present LUNA (Location reconstrUction using geNerative Ai), a generative AI model that reassembles complex tissue structures from gene expression of cells by learning spatial priors over spatial transcriptomics datasets. LUNA learns cell representations that capture cellular interactions globally and locally across the entire tissue slice, enabled by an attention mechanism [22, 23] that takes into account interactions across all cells. LUNA operates as a diffusion model [24–26]: during training it learns to denoise corrupted cell coordinates, while during inference it starts from random noise and reconstructs the physical locations of cells *de novo* solely from their gene expression.

We apply LUNA to reassemble the MERFISH whole mouse brain atlas, which consists of 1.23 million cells spanning all brain regions [27], as well as the scRNA-seq mouse central

nervous system atlas of 1.08 million cells [28]. LUNA effectively reconstructs complex mouse brain tissue across a wide range of brain regions and captures spatial gene expression patterns, including cells from previously unseen cell types without explicit training on those specific cell types. Furthermore, LUNA enables the enrichment of Slide-tags datasets [29] by inferring the tissue locations of nuclei lost during cell profiling. LUNA correctly placed different subtypes of neuronal cells in mouse embryonic tissue as well as tumor cells belonging to spatially distinct compartments in a human metastatic melanoma sample, enriching the sparse Slide-tags data by up to 95.4%. Finally, we generated a Xenium spatial transcriptomics dataset from the brains of wild-type control mice and α-synuclein-induced Parkinson's disease (PD) model mice. LUNA successfully identified clusters of cells with pathological α-synuclein, even though it was trained only on control samples, as confirmed by immunostaining. LUNA is highly scalable, with linear time and memory complexity relative to the number of input cells, and it can infer the spatial coordinates of tens of thousands of cells in less than a few minutes on a single GPU.

Results

Overview of LUNA. Given gene expression of dissociated cells, we aim to address the challenge of reassembling cells into compact tissue structures (Fig. 1a). To achieve that, we designed a diffusion model [24–26], named LUNA, that generates complex tissue structures by predicting the physical locations of individual cells based on their gene expression. The key challenge in generating tissue structures *de novo* arises from the inherent complexity of tissues, which are composed of a vast array of heterogeneous cells engaged in intricate cell-cell communication. LUNA overcomes these challenges by enabling each cell to "communicate" with all other cells guided by the learned weights, *i.e.*, attention scores (Methods). In this way, LUNA learns cell embeddings that capture cellular interactions on both the global and local levels, allowing LUNA to directly map cell embeddings to spatial locations. During training, LUNA learns spatial priors on existing spatial transcriptomics data. In the inference stage, LUNA generates complex tissue structures solely from gene expression of cells.

The neural network in LUNA consists of a multi-head self-attention mechanism [22, 23] that allows each cell to focus on gene expression of specific cells when making predictions (Fig. 1b). The cell embeddings are then mapped from the latent space to the physical space via a fully connected layer. To ensure robustness to rotations and translations of input slices, the loss function

in LUNA is specially designed to be invariant to these transformations and preserve the relative spatial relationships of the cells (Methods).

The training stage of LUNA comprises two processes: corruption and denoising, both of which are repeated over a range of diffusion times (Methods). During the corruption process, a controlled amount of Gaussian noise is added to the ground truth cell coordinates. Using the noised cell coordinates generated from this process and conditioned on the gene expression data, LUNA then learns to recover the original cell coordinates through the denoising process. During the inference stage, LUNA only takes gene expression profiles as model input and reconstructs tissue structures by denoising cell coordinates in an autoregressive manner. Initially, cell coordinates are generated from random Gaussian noise, and LUNA iteratively removes noise from these coordinates, predicting spatial locations of cells *de novo* (Fig. 1c, Supplementary Fig. 1). LUNA is highly scalable, with linear time and memory complexity relative to the number of cells. It infers physical locations of tens of thousands of cells in just a few minutes on a single GPU. For example, LUNA inferred locations of $40,000$ cells from the Allen Brain Cell (ABC) MERFISH mouse brain atlas [27] in approximately four minutes using around 6 GB of GPU memory (Fig. 1d).

LUNA accurately reconstructs the whole mouse brain atlas. LUNA effectively scales to atlas-scale large datasets and accurately reconstructed a whole mouse brain of the Allen Brain Cell (ABC) MERFISH mouse brain atlas [27]. We trained LUNA on 2.85 million cells across 147 slices from one mouse (Supplementary Note 1). We then applied it to reassemble cells from the whole brain of another mouse consisting of 1.23 million cells and 338 identified subclasses in 66 slices (Fig. 2a).

We evaluated LUNA's predictions using the ground-truth cell locations. We found that LUNA's predictions agree well with the ground-truth cell locations across 11 major regions identified in the ABC atlas, despite their distinct structural characteristics (Supplementary Figs. 2-4). For example, LUNA accurately captured the circular structure of the olfactory bulb region, the layered organization of the isocortex region, and the anatomical separation between the brain stem and cerebrum (Fig. 2b). In the slice from the olfactory bulb region affected by sequencing artifacts, LUNA inferred the spatial distribution of cells in areas lacking ground truth spatial information. To further evaluate the generation results, we compared LUNA's predictions across various levels of categorization, ranging from neurotransmitter level (4 types) to subclass level (338 types), and cell supertype level ($1,201$ types). LUNA accurately positioned cells at different levels of

categorization, from coarser to finer (Fig. 2c and Supplementary Fig. 5).

We next assessed LUNA's ability to infer the spatial distribution of gene expression and correctly position cells within a cell type based on their gene expression profiles. We identified spatially variable genes by computing Moran's I statistic [30] and visualized the expression values of genes with the highest Moran's I values (Supplementary Note 3). LUNA successfully captured the spatial expression patterns of genes within individual cell types, demonstrating its ability to preserve biological information. For example, LUNA accurately positioned cells expressing *Rxfp1* and *Rorb* in extratelencephalic and intratelencephalic neurons containing the neurotransmitter Glut (Fig. 2d). We also observed a high reconstruction quality across non-neuronal cells (Supplementary Fig. 6). For example, LUNA correctly positions the physical locations of a small group of 85 cells out of $3,725$ astrocyte-ependymal cells expressing the *Cfap206* gene (Supplementary Fig. 7).

To quantitatively evaluate the conservation of spatial gene expression patterns, we computed spatial autocorrelation using the Moran's I statistic for all $1,122$ genes across the 66 slices. We compared Moran's I values derived from LUNA's predicted locations with those from the ground-truth cell locations. Our analysis revealed a strong correlation between the Moran's I values based on LUNA's predictions and those based on the true locations, achieving an average Pearson correlation coefficient of 0.95 across all genes (Supplementary Fig. 8). For instance, the spatial autocorrelation for the spatially variable *Rorb* gene was consistently well-preserved across all slices, with a near perfect Pearson's correlation coefficient (Fig. 2e). These results demonstrate that LUNA effectively preserves spatial gene expression patterns throughout the mouse brain.

We further validated LUNA's capability to predict spatial information of missing slices. We trained LUNA on a subset of spatially measured slices from the ABC dataset and evaluated on 41 randomly held-out slices. LUNA's predictions agree well with ground-truth cell locations on this cross-slice generalization task (Supplementary Fig. 9 and 10).

LUNA generalizes to unseen cell classes. We next evaluated the zero-shot capability of LUNA to predict the spatial locations of cells from classes unseen during training, without any additional training or fine-tuning of the model (Fig. 2f). We excluded each of the 31 broad cell class categories that have sufficient number of cells (more than 1000 cells across the entire atlas) identified in [27] and then trained LUNA on slices from the Animal 1 in the ABC atlas without the held-out cell class (Supplementary Note 1). We then applied LUNA to predict tissue structures for all cell classes in Animal 2, including the held-out cell class. Across different cell classes, LUNA correctly

positioned cells within the spatial architecture of the tissue (Supplementary Figs. 11-13). For example, LUNA correctly positioned NP-CT-L6b Glut cells in the isocortex, cortical subplate and hippocampal formation regions (Fig. 2g, h). We additionally examined the spatial expression patterns of spatially variable genes in NP-CT-L6b Glut cells. We identified genes *Tshz2* and *Syt6* as genes with the highest Moran's I values among $1,122$ genes based on the ground-truth cell locations. We found that within the unseen NP-CT-L6b Glut class, LUNA correctly positioned cells according to the *Tshz2* and *Syt6* expression patterns (Fig. 2g,h). To analyze the similarity of the unseen NP-CT-L6b Glut cells cell class with seen classes, we examined the gene expression similarity between this class and other seen classes. NP-CT-L6b Glut cells partially share gene expression patterns with several seen classes (Supplementary Fig. 14). This suggests that LUNA captures shared spatial gene expression patterns and models continuous spatial variation in gene expression, enabling it to infer spatial coordinates for previously unseen cell types. We further analyzed other randomly excluded cell classes, such as the vascular cell class that is spread across the entire slice (Fig. 2i, Supplementary Fig. 15) and the HY GABA cell class (Supplementary Fig. 16), confirming that LUNA preserves key spatial gene expression patterns of unseen cell types, demonstrating strong zero-shot capabilities.

LUNA outperforms other methods by a large margin. LUNA is a unique method in its capability to reconstruct complex tissue structures. Existing methods are reference mapping methods and require the pre-selection of a single reference slice to which then new cells are mapped based on the similarity of the gene expression between the cells in the reference slice and the input gene expression profiles (Fig. 3a). This mapping is performed either by explicitly learning a mapping matrix, as in novoSpaRc [20], Tangram [14], STEM [16], and CytoSPACE [15], or by implicitly regressing from gene expression space to the physical space, as in scSpace [17], STALocator [18], CeLEry [19] (Supplementary Note 2).

We compared the performance of LUNA to alternative methods on the spatially resolved cell atlas of the mouse primary motor cortex sequenced by MERFISH [31] (Supplementary Note 1). We used 31 slices ($118,036$ cells) from one animal as the test set and trained LUNA on 33 slices ($158,379$ cells) from another mouse to learn spatial priors across multiple slices. During inference, all methods were conditioned only on the gene expression. Since existing methods can not accept multiple slices as input, we randomly selected a single slice from the training mouse to serve as the reference, repeating the procedure for each slice of the testing mouse.

We evaluated the results using the Spearman's rank correlation between the predicted and ground truth locations, averaging the performance across all slices from the unseen mouse (Supplementary Note 3). LUNA achieved an average correlation coefficient of 44.8%, outperforming the best alternative baseline scSpace by 63% (Fig. 3b). Performance gains of LUNA were retained using other evaluation metrics, such as precision and RSSD (Supplementary Note 3, Supplementary Fig. 17). CeLEry is the only other method besides LUNA capable of potentially leveraging information from multiple slices. We reran CeLEry using all slices from the training mouse as the training dataset. However, since CeLEry is a regression model that does not account for geometric constraints, it is not robust to variations in slice orientation or cutting shapes within the training set. Consequently, we observed a significant performance decline when when CeLEry was trained with multiple slices compared to using a single slice (Supplementary Fig. 17). We visualized the spatial locations predicted by different methods on a sample slice from the unseen test mouse (Fig. 3c, Supplementary Fig. 18). LUNA accurately predicted the layer-wise structure of mouse primary motor cortex tissues and correctly inferred the spatial priors of the tissue, such as the contour of the tissue. We analyzed the spatial gene expression patterns of cell type specific marker genes. LUNA distinguished cells from different cell classes, accurately reflecting the spatial gene expression patterns (Fig. 3d-f). In contrast, other methods struggled due to the dissimilarity between the reference and test slices (Supplementary Fig. 19).

The limitations of other methods stem from their reliance on pre-selected reference slices, which constrains their generalization capabilities. To systematically evaluate the sensitivity of existing methods on the reference slice, we selected different slices as a reference training slice and evaluated performance on a fixed test slice. We computed the average performance of these methods across different reference slices. LUNA achieved a 121% higher Spearman's rank correlation coefficient compared to the average performance of the best alternative method, scSpace, where the average is calculated over different reference slices (Supplementary Fig. 20). Given the sensitivity of reference mapping methods to the choice of reference slices, we further explored alternative methods by selecting the most similar slice from another animal as the reference by using cell class information as a prior (Supplementary Note 2). We then reassessed the performance of alternative methods using this prior information. This prior information improved the performance of alternative baselines (Fig. 3g); however, they still lagged significantly behind LUNA's performance. Even without using any prior information, LUNA outperformed all baselines, achieving 37.5% better performance on SRC than the best alternative method, CeLEry (Supplementary Fig.

21). We further compared LUNA to other alternative methods on the MERFISH ABC Atlas in the setting that selects the best reference slice for alternative methods as a prior (Supplementary Note 2). LUNA again significantly outperformed all alternative methods in all evaluation metrics even without using any prior information (Supplementary Fig. 22).

De novo reconstruction of scRNA-seq data. We next applied LUNA for *de novo* generation of tissue structures of 1.08 million dissociated single cells across 13 coronal slices from the mouse central nervous system (CNS) scRNA-seq atlas [28] (Fig. 4a). We used the model trained on the ABC MERFISH mouse brain atlas to predict spatial locations for cells in the scRNA-seq mouse CNS atlas integrated with the MERFISH atlas using Harmony [32]. Since the scRNA-seq atlas lacks spatial information, we validated LUNA's performance using estimated cell locations derived from the integration of the scRNA-seq atlas with the STARmap PLUS CNS spatial atlas [33] (Supplementary Note 1). In [33], the mouse CNS spatial atlas was created using the in situ sequencing STARmap PLUS, which was then integrated with the scRNA-seq data using Harmony [32]. The integrated dataset underwent careful quality control by human experts, ultimately providing estimated cell locations of the scRNA-seq atlas [33] (Fig. 4b). To validate LUNA's predictions, we used these estimated cell locations obtained after the integration. Notably, LUNA never used spatial information from the STARmap PLUS atlas.

Applying LUNA to the scRNA-seq CNS atlas, its predictions aligned closely with cell locations estimated through the integration data with the STARmap PLUS atlas at both the coarse cell class level (27 types; Fig. 4c; Supplementary Fig. 23) and the finer submolecular class level, which includes 216 distinct types (Supplementary Fig. 24). The agreement is particularly pronounced in the isocortex, hippocampal formation, thalamus, and hypothalamus regions. We further evaluated LUNA's performance in reconstructing mouse CNS tissue architecture across various cell classes (Supplementary Fig. 25). Additionally, we examined LUNA's predictions for specific neuronal subtypes including di- and mesencephalon excitatory neurons (Fig. 4d), and telencephalon projecting excitatory neurons (Fig. 4e), along with their respective sub-molecular classes. The results indicate that LUNA not only accurately placed cells across major cell classes but also precisely predicted the spatial relationships of complex sub-molecular classes within these neuronal cells. This underscores LUNA's capability to capture both broad and intricate cellular architectures within the CNS.

To evaluate the conservation of spatial gene expression patterns, we analyzed the spatial

distribution of three highly variable genes—*LPL*, *IQGAP2*, and *MEF2C*, within telencephalon-projecting excitatory neurons (Fig. 4f). These genes were selected based on their high Moran's I values, calculated from a dataset of $10,844$ genes using LUNA's predictions. This analysis revealed that the spatial gene expression patterns predicted by LUNA closely align with the estimated ground truth, demonstrating LUNA's effectiveness in capturing gene expression gradients at the cellular level. Specifically, *IQGAP2* gene shows a gradual downregulation across the Ammon's horn region (CA2) of the hippocampus, while *LPL* displays an opposite trend with gradual upregulation in this region. For *MEF2C* gene, expression is minimal in telencephalon-projecting excitatory neurons within the CA2 region but substiantially upregulated in the isocortex and olfactory areas. Overall, these results confirm that LUNA effectively reconstructs the complex spatial architecture of the mouse CNS scRNA-seq atlas and captures gene expression spatial changes. This underscores LUNA's applicability in reconstructing cell locations for scRNA-seq atlases. We evaluate these predictions quantitatively by computing spatial autocorrelation using Moran's I statistics across all $11,844$ genes from the scRNA-seq CNS atlas and observe high agreement between LUNA's predictions and estimated groundtruth locations (Supplementary Fig. 26). To further assess LUNA's performance in the presence of batch effects, we evaluated LUNA on both non-integrated and integrated versions of the scRNA-seq data. LUNA maintains good performance despite batch effects, demonstrating robustness to technological differences. However, it still benefits from batch correction to further enhance overall performance. (Supplementary Fig. 27).

By predicting the locations of dissociated scRNA-seq data, LUNA can be used to detect spatially variable genes that are absent from the spatial dataset used for model training. To demonstrate this, we detected cell-class specific spatially variable genes among genes that are unseen during training (*i.e.*, specific to scRNA-seq data) using LUNA's predicted cell locations (Supplementary Note 5). Using predicted locations, we identified genes with distinct spatial patterns in both neuronal and non-neuronal cells (Fig. 4g). Additionally, we report the top five (among the total of $11,844$ genes) unseen genes with the highest Moran's I value for different cell classes (Supplementary Table 1).

As a generative model, LUNA allows capturing multiple plausible spatial arrangements, enabling the estimation of uncertainty in predictions. To estimate uncertainty in locations of scRNA-seq data, we generated 150 samples from the LUNA model using different random seeds. We then applied kernel density estimation to the predicted locations of each cell to analyze spatial uncer-

tainty (Supplementary Note 6). For each individual cell, LUNA can estimate high density regions (HDR) and thus model uncertainty in predictions (Fig. 4h). Extending this modeling on the level of cell classes, LUNA can be used to model uncertainty of cell classes and identification of regions in which locations of cells are interchangeable. For example, we found that telencephalon projecting inhibitory neurons on average have smaller HDR regions and thus lower uncertainty in the spatial locations of cells belonging to these cell classes, while ependymal cells have the highest uncertainty (Fig. 4i).

LUNA infers locations of spatially unmapped nuclei in the Slide-tags data. LUNA can be used to predict the tissue locations of nuclei lost during cell profiling with Slide-tags technology [29]. Slide-tags enables profiling single-cell and spatially resolved transcriptome by tagging nuclei with spatial barcode oligonucleotides derived from DNA-barcoded beads with known positions and then using tagged nuclei as an input to single-nucleus profiling assays [29]. However, many nuclei are lost during the barcoding process due to a combination of dissociation and microfluidic losses. To compensate for the sparsity in Slide-tags data, we applied LUNA to assign spatial locations to the spatially unmapped nuclei in the Slide-tags data.

We first applied LUNA to the sagittal section of the embryonic mouse brain at embryonic day 14 (E14) dataset consisting of $4,614$ spatially mapped with Slide-tags (Fig. 5a) (Supplementary Note 1). To evaluate LUNA's performance using ground truth nuclei locations, we used $4,152$ of spatially mapped nuclei as the train set and evaluated LUNA's performance on remaining 462 cells. LUNA's predictions agree well with the ground truth locations of cells obtained with Slide-tags (Fig. 5b). Without using any cell class information, LUNA accurately predicted cell locations, even for different neuronal subclasses. We further validated these predictions by computing spatial autocorrelation using Moran's I index across all genes used for training and observe high agreement between LUNA's predictions and true nuclei locations (Fig. 5c).

Encouraged by these results, we applied LUNA to infer locations of spatially unmapped nuclei due to nuclei losses with Slide-tags. We trained LUNA on spatially mapped cells of mouse E14 tissue and applied the model to infer locations of $4,414$ spatially unmapped nuclei. To validate predictions of LUNA on the cell class level, we ran linear support vector classifier to infer cell classes of spatially unmapped nuclei since annotations for spatially unmapped nuclei were missing in the original dataset (Supplementary Note 1). We found that LUNA placed spatially unmapped nuclei in proximity of cells with the same class, successfully enriching sparse Slide-tags dataset.

We further examined the spatial distribution of specific cell types, additionally confirming that LUNA placed nuclei from both neuronal and non-neuronal cells to the correct spatial regions placed (Fig. 5e, Supplementary Fig. 28).

We next applied LUNA to a human metastatic melanoma sample obtained with Slide-tags technology [29]. This sample includes a diverse array of cells, including immune cells and tumor cells. We trained LUNA on $4,804$ spatially mapped nuclei and applied it to generate locations for the spatially unmapped nuclei. We again transferred cell type annotations of spatially unmapped nuclei with the linear support vector classifier to validate LUNA's predictions (Supplementary Note 1). We found that LUNA successfully enriched the dataset to the total of $6,466$ spatially mapped cells (Fig. 5f). LUNA correctly placed two tumor subpopulations into spatially distinct compartments according to predicted cell class annotations (Fig. 5g), as well as non-tumor cells (Supplementary Fig. 29). To further validate LUNA's predictions, we investigated the the heterogeneity of these two tumor populations by identifying highly differential genes between the two cell classes. Cells expressing *CLU* gene are placed by LUNA in different spatial compartments than cells *DISC2FP1* gene, which is also the case in the original dataset of spatially mapped nuclei of the melanoma sample (Fig. 5h, i). We applied LUNA to other Slide-tags datasets, including samples from human cortex (Supplementary Fig. 30) and human tonsil (Supplementary Fig. 31).

We investigated whether LUNA-enriched Slide-tags data can increase the power to detect spatially variable genes in the human metastatic melanoma sample. We detected SVGs using SpatialDE [34], focusing exclusively on tumor cells in the sample (Supplementary Note 5). Using the LUNA-enriched dataset, we identified additional 396 spatially variable genes (FDR-corrected [35] p-value < 0.01) that were not detectable in the original sample, increasing the number of spatially variable genes by 83% (Fig. 5j). We anayzed Gene Ontology (GO) enrichment on these 396 genes using KEGG [36] and MSigDB [37] databases (Supplementary Fig. 32). From the MSigDB database, we found strongest enrichment of "Myc Targets V1", which includes genes regulated by the MYC oncogene known to promote cell proliferation and closely linked to tumor aggressiveness and survival outcomes [38]. From the KEGG database, the most significantly enriched pathway was the Ribosome pathway with known causal associations with increasing cancer risk [39]. Among genes uniquely detected using LUNA-enriched sample, the genes with the lowest p-value include *U2SURP* (p-value $= 3.4e-8$), *POLR3D* (p-value $= 5.2e-8$), and *WWTR1* (p-value $= 7.6e-8$) (Fig. 5k, top-right). The abnormal activity of these genes is known to be associated with cancer [40–43]. We further use the Fraction of Spatial Variance (FSV) [34] to measure the

extent to which spatial patterns contribute to the total variance observed in gene expression data (Fig. 5k). Of all significant genes detected using the LUNA-enriched sample, the 50 genes demonstrate spatial dependencies that explain more variance than the average (FSV $= 7.4\%$, averaged over 833 identified SVGs), including *CD2AP* (FSV=12.2%), *RPL22* (FSV=11.7%) and *RPS4X* (FSV=11.4%) genes. *RPL22* is known to act as a tumor suppressor gene [44], while previous studies have suggested that *CD2AP* and *RPS4X* are associated with certain malignant tumors [45–48].

Interpreting gene contributions to spatial structure prediction. LUNA can enable identification of genes whose exclusion affects the prediction of spatial structure the most. We performed analysis to investigate how individual genes influence the predictions made by LUNA. Specifically, we measured the sensitivity of spatial predictions to exclusions of each individual gene by masking its expression to zero during inference, and quantified the resulting changes in predicted cell positions. We computed discrepancies between spatial predictions obtained with the full gene panel and those obtained with one gene excluded (Fig. 6a).

We investigated how individual genes influence the predicted spatial locations of major cell classes in the ABC atlas. We trained LUNA on all slices from the Animal 1 with the whole gene panel and during inference masked individual genes from Animal 2. To quantify the effect, we calculated the Spearman Rank Correlation (SRC) within each cell class and determined gene contributions in a cell-class-specific manner (Supplementary Fig. 34, Supplementary Table 2). Our analysis revealed that certain genes have a particularly strong effect on prediction accuracy, including *Igfbpl1*, *Cldn11*, *Vgf*, *Nefh*, and *Nr2f1* (Fig. 6b). These genes showed the highest impact within the olfactory bulb–intracerebral migrating stream GABA neurons (OB-IMN GABA), oligodendrocyte precursor cells and oligodendrocytes (OPC-Oligo), cortex–medial ganglionic eminence GABA neurons (CTX-MGE GABA), and cortex–caudal ganglionic eminence GABA neurons (CTX-CGE GABA), respectively. Functionally, these genes are closely tied to cell identity and function: *Igfbpl1* is involved in regulating cell growth and GABAergic transmission in mice [49, 50]; *Cldn11* encodes for oligodendrocyte-specific protein essential for myelin sheath formation [51]; *Vgf* and *Nefh* genes are biomarkers related to neurodegenerative diseases and neuronal damage [52, 53]; *Nr2f1* plays a crucial role in neurodevelopment, significantly influencing cortical development and network activity maturation [54, 55]. Interestingly, previous work has shown that myelin mutant mice lacking expression of the *Claudin11* gene in oligodendrocytes exhibit central auditory deficits, reduced anxiety-like behavior and neurotransmitter imbalances [56].

We further evaluated the performance variation with different initializations of the noise in the inference phase, confirming the robustness of our findings and systematic influence of *Igfbpl1*, *Cldn11*, *Vgf*, *Nefh* and *Nr2f1* genes on specific cell classes (Fig. 6c-e, Supplementary Fig. 35).

LUNA can be used to assess the impact of pairs of genes on spatial organization. Due to the combinatorial number of potential interactions, we narrowed our focus to ligand-receptor pairs curated in OmniPath [57], resulting in 332 ligand-receptor pairs that overlap with the ABC atlas gene library. We assessed the model's sensitivity to these interactions by jointly masking expression of each ligand-receptor pair during inference and measuring the impact on spatial predictions (Fig. 6f), focusing on the five cell classes that were most affected in the single-gene contribution analysis. We found that CTX-CGE GABA neurons are predominantly influenced by interactions involving the *Slc32a1* gene which functions as a GABA vesicular transporter (Fig. 6g). MH-LH Glut neurons are significantly affected by interactions with the *Slc17a7* gene, a vesicular glutamate (Glut) transporter (Fig. 6h). Notably, *Slc17a7* interactions also impact the structural integrity of CTX-CGE GABA neurons and oligodendrocytes, while interactions with *Thbs4* ligand have the effect on OB-IMN GABA neurons (Supplementary Fig. 36). The loss of *Thbs4* has been previously associated with a dystrophic phenotype in mice [58]. Furthermore, CTX-CGE GABA and MH-LH Glut neurons are influenced by the *Gad2-Gabbr2* ligand-receptor pair, with a pronounced effect observed in CTX-CGE GABA neurons.

LUNA generalizes to pathological conditions. To study the ability of LUNA to predict spatial locations of cells in tissues with undergoing pathological processes, we generated Xenium data from an α-synuclein-induced Parkinson's disease mouse model [59, 60] (Methods). Mice were unilaterally injected with $2\mu l$ of pre-formed fibrils (PFFs) into the dorsal striatum at eight weeks of age (Fig. 7a). In this model, misfolded proteins propagate to neighboring neurons, seeding further aggregation in anatomically connected regions. Three months post-injection (3MPI), brains were sectioned and profiled. The same sections were subsequently immunostained to detect misfolded α-synuclein, providing an independent experimental validation of the predicted pathological changes (Fig. 7b).

We used a LUNA model pretrained on the ABC MERFISH Atlas and fine-tuned it on generated Xenium wild-type sections. We then applied the model to infer the spatial locations of diseased 3MPI sections that the model has never seen (Fig. 7b, Supplementary Note 1). LUNA successfully reconstructed the spatial organization of the different coronal sections (Supplemen-

tary Fig. 37). We further visualized individual cell classes and observed strong alignment between LUNA 's predictions and the ground truth spatial distributions in the 3MPI section (Supplementary Fig. 38). Notably, in the amygdala region, LUNA 's predictions more closely resembled the spatial organization of the diseased samples than the wild-type samples on which the model was fine-tuned (Fig. 7c).

Encouraged by these results, we decided to test whether LUNA can be used to identify spatial gene expression deregulation, such as those we expect to be induced by α-synuclein pathology spread. Specifically, we reasoned that the physical space learned by LUNA can be used for cell alignment and the identification of regions of local gene-expression changes, without ever measuring spatial data. To achieve this, we first run LUNA on slices from both wild-type and 3MPI samples, resulting in spatial predictions in a shared coordinate framework (Fig. 7d, Supplementary Fig. 39). Then, we use these cell coordinates to quantify 3MPI-vs-wild-type local gene expression changes (Supplementary Note 7). We found differentially expressed genes like *Myl4*, *Cdh9* and *Bdnf* in the amygdala, and the hippocampus, some of which have been associated with neuronal stress or senescence [61–63]. We reasoned that the regions where cumulative changes across all genes were maximal were also the most likely to be the most affected by pathology. To test this, we computed a unified score for all cells based on their local gene expression differences between the 3MPI and wild-type samples. (Supplementary Fig. 40, Supplementary Note 7). We computed Jensen-Shannon Divergence across all genes to identify those with the most significant distributional shifts between wild-type and 3MPI samples. Local gene expression differences were then assessed by comparing spatially averaged expression across each cell's neighborhood. We found certain cluster of cells that are consistently flagged across more than half of the significantly divergent genes, revealing regions with potential pathological transcriptomic alterations (Fig. 7d, Supplementary Fig. 39).

To validate these regions of interest, we performed immunostaining against pathological α-synuclein in the 3MPI sample (Fig. 7e). We found that cells identified as locally significantly different by LUNA exhibited higher levels of α-synuclein aggregates (Fig. 7f, Supplementary Fig. 41). Interestingly, LUNA accurately predicted the locations of clusters of cells with pathological α-synuclein. These results showcase how LUNA can generalize beyond wild-type conditions, but can be applied to different biological conditions also providing an unique opportunity to identify spatial gene expression dysregulations without spatial data.

Discussion

LUNA is a generative AI model that reconstructs tissue architecture by reassembling dissociated cells based on their gene expression profiles, The model learns cell representations that allow for accurate positioning of cells within the tissue context, even in the absence of direct spatial information. LUNA operates in a continuous physical space, unrestricted by a finite number of physical spots, thus enabling the generation of tissue structures with arbitrary spatial density. Moreover, LUNA can infer spatial priors of the tissue and adapt to the context of the test sample without requiring explicit slice alignment. This is achieved by conditioning on the global gene expression profile of all cells, whereas alternative methods rely on selecting a specific reference slice for spatial reconstruction.

LUNA has broad implications in the field of spatial biology and addresses existing technological limitations. LUNA can infer 2D spatial locations for dissociated single cells from scRNA-seq datasets, enabling reconstruction of complex tissue architectures *de novo*. Reconstructing spatial structures from scRNA-seq data thus enables researchers to obtain both rich gene expression profiles and inferred spatial localization, offering the best of both modalities. In addition, LUNA can generate locations for cells or spots missing spatial information in the existing spatial technologies such as Slide-tags [29]. LUNA offers an effective solution for spatial transcriptomics by reducing the need for comprehensive imaging of all tissue slices. A smaller number of slices can be imaged and LUNA can be used to predict spatial information for the remaining slices, thus lowering experimental costs while maintaining high-resolution spatial insights. Finally, LUNA is a complementary tool for analyzing tissue spatial structures, gene expression gradients and cell-cell communication along with existing tools for spatial analysis such as Squidpy [64], Seurat V5 [65], STELLAR [66] and STalign [67], as well as with existing batch correction methods [32,68] whose embeddings can be used to condition LUNA.

Currently, LUNA operates within a two-dimensional (2D) spatial framework, which does not fully capture the inherently three-dimensional (3D) nature of tissue architectures. This limitation stems primarily from the constraints of existing sequencing technologies, which are not yet capable of measuring tissue structures in full 3D detail. LUNA framework can be readily extended beyond 2D spatial reconstruction and with ongoing advancements in technology [69], datasets that provide rich molecular organization of tissues in 3D are likely to become more widely available in the near future.

Our current analysis has mainly focused on mouse brain cell atlases due to the abundance of

available data. However, as our results on the Slide-tags human tissues show, LUNA is a versatile framework that can be readily applied to other organs and species. Future research could extend the use of LUNA to various tissues and model organisms, enabling to uncover organizational principles underlying tissue architectures across various biological systems.

We envision LUNA, as an AI-driven tissue reassembly model, holds potential for advancing our understanding of complex biological processes and generating the intricate architecture of tissues across diverse biological systems, in both healthy and diseased states. By training LUNA across a range of biological contexts, the model could lay the foundation for virtual tissue models: computational frameworks that capture the spatial and functional landscape of real tissues and simulate the effects of perturbations within the context of tissue architecture. LUNA supports this vision by its ability to generate tissue maps for unmeasured or perturbed samples, creating a virtual reconstruction of spatial tissue architecture.

Methods

Overview of LUNA. LUNA is a diffusion-based generative model designed to generate spatial locations of cells from their gene expression. Using an attention-based mechanism, LUNA learns how each individual cell should attend to other cells according to their relevance to a cell's own location and molecular features. During training, LUNA leverages spatial transcriptomics data with ground truth cell locations to learn to generate cell locations conditioned on gene expression. During inference, the input to LUNA is the gene expression of dissociated cells, and the model then generates cell locations from pure noise conditioned on gene expression, enabling *de novo* inference of cell locations.

In the training phase, the input to LUNA is spatial transcriptomics data from single or multiple slices. LUNA learns cell embeddings that capture both the local and global tissue structure using a multi-head self-attention mechanism [22,23] that enables cells to learn to attend to other cells. We designed the loss in LUNA as an SE(2)-invariant function [70–73], *i.e.*, it is robust to arbitrary rotations, translations and reflections of the predicted locations, only enforcing the preservation of relative distances between cells within each slice. We further conducted a comprehensive ablation study on LUNA's model components which confirms that LUNA's attention-based architecture is crucial for capturing complex spatial dependencies between cells, while the pairwise loss function plays a key role in maintaining relational consistency (Supplementary Note 8, Supplementary Fig. 42).

Training LUNA involves two main steps: *(i)* a corruption process that iteratively adds noise to the ground truth cell locations, and *(ii)* a denoising process that generates the ground truth locations from the noise-corrupted locations. During the denoising process, LUNA learns to reverse the corruption process, gradually transitioning from noisier to less noisy cell locations, ultimately reconstructing true cell locations from pure noise. LUNA as as generative model is more robust to distribution shifts, compared to a regression model which is prone to overfitting to specific tissue distributions (Supplementary Fig. 43). Another benefit of using generative modeling lies in estimating uncertainty in predictions (Fig. 4g,h, Fig. 6).

Learning cell representations. LUNA learns cell representations that capture the relationship between the molecular features of individual cells and their spatial context within tissue. These embeddings form the foundation for predicting the locations of cells.

The model is built on the premise that a cell's location is influenced not only by its own

molecular features but also by those of other cells in the tissue. To effectively model these interactions, LUNA employs a multi-head self-attention mechanism [22, 23], which dynamically integrates information from all cells in the slice. This mechanism allows the model to assign attention scores to different cells based on their relevance to the prediction of each cell's location, capturing both local and global cellular interactions. Once the cell embeddings are learned, LUNA projects each cell's embedding into a 2-dimensional space to predict its locations.

Formally, LUNA takes as input *(i)* molecular features of cells $\mathbf{X}_s \in \mathbb{R}^{m \times d}$ in a given slice s where m denotes the number of cells and d denotes the number of genes, *(ii)* time step $t \in \mathbb{N}$ uniformly sampled from 0 to the maximum diffusion time *T*, and *(iii)* corrupted cell locations $\tilde{\mathbf{R}}_{s,t} \in \mathbb{R}^{m \times 2}$ at time step t. The output of the model are the denoised cell locations $\hat{\mathbf{R}}_s \in \mathbb{R}^{m \times 2}$ for a given slice s. We omit slice index s for the ease of notation.

LUNA is composed of multiple transformer layers, each integrating fully connected layers followed by the self-attention block. At each layer l given a fixed diffusion time t, LUNA learns *(i)* cell embeddings $\mathbf{F}^{(l)} \in \mathbb{R}^{m \times d_x^{(l)}}$, *(ii)* a diffusion time embedding $\gamma_t^{(l)} \in \mathbb{R}$, and *(iii)* cell location embeddings $\mathbf{R}^{(l)} \in \mathbb{R}^{m \times d_r^{(l)}}$. The cell embeddings $\mathbf{F}^{(0)}$ in layer 0 are initialized via fully connected neural network layer that maps the molecular features of cells into the latent embedding space with dimensionality $d_x^{(0)}$:

$$\mathbf{F}^{(0)} = \mathit{FCN}(\mathbf{X}_s), \tag{1}$$

where *FCN* denotes a fully connected neural network layer. The diffusion time embedding $\gamma_t^{(0)}$ is initialized to the value of the diffusion time itself:

$$\gamma_t^{(0)} = t. \tag{2}$$

Finally, the cell location embeddings are initialized as the corrupted cell locations obtained from the corruption process at time step t:

$$\mathbf{R}^{(0)} = \tilde{\mathbf{R}}_t. \tag{3}$$

Then, at each layer $l \in \{1, \ldots, L\}$, LUNA first transforms cell embeddings, a diffusion time embedding and cell location embeddings using fully connected layers:

$$\mathbf{F}_f^{(l)} = \mathit{FCN}\left(\mathbf{F}^{(l-1)}\right), \tag{4}$$

$$\gamma_t^{(l)} = \mathit{FCN}\left(\gamma_t^{(l-1)}\right), \tag{5}$$

$$\mathbf{R}_r^{(l)} = \mathit{FCN}\left(\mathbf{R}^{(l-1)}\right). \tag{6}$$

The transformed cell embeddings, diffusion time embedding and cell location embeddings are concatenated to form a unified representation for each cell:

$$\mathbf{H}^{(l)} = \left[\mathbf{F}_f^{(l)}, \gamma_t^{(l)} \mathbf{1}_{m\times 1}, \mathbf{R}_r^{(l)}\right], \tag{7}$$

where $\mathbf{1}_{m\times 1}$ is a column vector of ones used to replicate the diffusion time embedding $\gamma_t^{(l)}$ across all m cells. To obtain final cell embeddings $\mathbf{F}^{(l)}$ in layer l, the unified representation $\mathbf{H}^{(l)}$ is passed through a self-attention block which computes attention weights between all pairs of cells [23]:

$$\mathbf{F}^{(l)} = \text{SelfAttention}^{(l)}\left(\mathbf{H}^{(l)}\right). \tag{8}$$

This self-attention operation enables the model to aggregate information from all cells in a slice, capturing how each cell's location is influenced by others. To ensure scalability, we implemented the efficient attention [23] – an approximation of attention computation with a linear complexity. This approach reduces memory usage and computational complexity by circumventing the need for explicit pairwise interactions between every query-key pair (Supplementary Note 9).

Finally, the cell embeddings $\mathbf{F}^{(l)}$ are used to predict the locations of each cell. The cell locations are generated from the embeddings using a fully connected layer that projects the location of each cell from the latent embedding space to the two-dimensional physical space:

$$\mathbf{R}^{(l)} = FCN\left(\mathbf{F}^{(l)}\right). \tag{9}$$

At each layer, the updated locations are centered by subtracting the mean, ensuring that the cell locations remain stable and consistent relative to the tissue structure. At the final layer, the output cell locations are the predicted clean cell locations $\hat{\mathbf{R}} = \mathbf{R}^{(L)}$. To better illustrate the scalability of our model components, especially the efficient attention mechanism and the FCN layer, we conducted kernel-level profiling of LUNA (Supplementary Note 10 and Supplementary Table 3).

Loss objective. To optimize LUNA, we introduce a specific loss objective that is SE(2)-invariant. This geometric invariance is crucial because the spatial arrangement of cells in the tissue may undergo transformations due to experimental artifacts, while the underlying gene expression profiles remain unchanged.

To ensure SE(2)-invariance, LUNA introduces a pairwise loss function which focuses on preserving the relative distances between all cells in the slice defined as follows:

$$\mathcal{L}_{\text{diff}} = \frac{1}{m^2}\|\Delta(\hat{\mathbf{R}}) - \Delta(\mathbf{R}_{gt})\|_F^2, \tag{10}$$

where $\Delta(\cdot) : \mathbb{R}^{m\times 2} \to \mathbb{R}^{m\times m}$ computes the pairwise squared Euclidean distances between the locations of all cells and $\mathbf{R}_{gt} \in \mathbb{R}^{m\times 2}$ represents the ground true cell locations.

The function $\Delta(\hat{\mathbf{R}})_{ij} = \|\hat{r}_i - \hat{r}_j\|_2^2$ computes the squared distance between the predicted locations of cells i and j, and similarly, $\Delta(\mathbf{R}_{gt})_{ij} = \|r_i - r_j\|_2^2$ computes the squared distance between the true locations. The pairwise loss then becomes:

$$\mathcal{L}_{\text{diff}} = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\|\hat{r}_i - \hat{r}_j\|_2^2 - \|r_i - r_j\|_2^2)^2, \tag{11}$$

which penalizes deviations between the predicted and true pairwise distances, ensuring that the spatial relationships between cells are preserved.

The final loss is computed over all slices $s \in \{1, ...S\}$ where S denotes the total number of slices:

$$\mathcal{L} = \frac{1}{S}\sum_{s=1}^{S}\mathcal{L}_{\text{diff}}^{(s)}. \tag{12}$$

Corruption process. The diffusion model training begins with a corruption process, where LUNA progressively introduces noise to the true cell locations in an autoregressive manner. This step gradually transitions the true cell locations into pure noise, with the resulting corrupted locations serving as inputs for the subsequent denoising process.

We denote the initial diffusion time step, where no noise has been introduced to the cell locations, as time step 0, where ground-truth spatial information is fully retained. Conversely, the maximum diffusion time step, denoted as T, represents the state where no ground-truth information is preserved, and the locations are sampled from pure Gaussian noise. The time step t is randomly selected from $\{0, 1, 2, \ldots, T\}$. The corrupted cell locations at time step t for the slice s are denoted as $\tilde{\mathbf{R}}_t \in \mathbb{R}^{m\times 2}$ where again we omit the slice index s for the ease of notation. From time step $t-1$ to t, Gaussian noise is added to the corrupted cell locations $\tilde{\mathbf{R}}_{t-1}$, yielding noisier locations $\tilde{\mathbf{R}}_t$ according to a predefined noise scheduler [24, 25, 74]. This noise scheduler is controled by the noise schedule exponent, which controls how much noise is injected or removed at each timestep of the reverse diffusion process and governs the randomness during sampling [75]. At time step T, the true cell locations $\mathbf{R}_{gt} = \tilde{\mathbf{R}}_0$ are transformed into corrupted cell locations $\tilde{\mathbf{R}}_T$ that have standard Gaussian distribution. The conditional distribution is defined as:

$$q_\phi(\tilde{\mathbf{R}}_t|\tilde{\mathbf{R}}_{t-1}) = \mathcal{N}(\tilde{\mathbf{R}}_t|\sqrt{\alpha_t}\tilde{\mathbf{R}}_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{13}$$

where $q_\phi(\cdot|\cdot)$ represents the corruption process with a noise scheduler ϕ that conditionally transitions the cell location distribution from less corrupted to more corrupted states. $\mathcal{N}(\cdot|\cdot,\cdot)$ denotes a Gaussian distribution, while $\mathbf{I}$ is the identity matrix. The noise scheduler ϕ, regulated by hyperparameters $\{\alpha_t\}$, controls the balance between retaining the original cell locations and adding noise and governs the randomness during sampling.

To mitigate computational complexity, we compute the conditional distribution of corrupted cell locations given the true cell locations $\mathbf{R}_{gt}$, rather than from the previous diffusion step [25,76]:

$$q_\phi(\tilde{\mathbf{R}}_t|\tilde{\mathbf{R}}_0) = \mathcal{N}(\tilde{\mathbf{R}}_t|\sqrt{\overline{\alpha}_t}\tilde{\mathbf{R}}_0, (1-\overline{\alpha}_t)\mathbf{I}). \tag{14}$$

Sampling from this conditional distribution is equivalent to adding Gaussian noise to the true cell locations. As shown in [76], the corrupted cell locations are computed as:

$$\tilde{\mathbf{R}}_t = \sqrt{\overline{\alpha}_t}\tilde{\mathbf{R}}_0 + \sqrt{(1-\overline{\alpha}_t)}\epsilon, \tag{15}$$

where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and ϵ is drawn from a Gaussian distribution with mean 0 and variance $\mathbf{I}$, *i.e.*, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. As t progresses from 0 to T, the cell locations become increasingly corrupted until they are indistinguishable from white noise. At $t = T$, the corrupted locations $\tilde{\mathbf{R}}_T$ are sampled purely from Gaussian noise, with $\overline{\alpha}_T = 0$. To ensure robustness to translations, we additionally subtract the center of mass from the noise ϵ and the data slices $\tilde{\mathbf{R}}_0$, thus working with distributions defined in the subspace where the center of mass is fixed to 0 [77,78].

Denoising process. Given the corrupted locations, we train the LUNA model to recover the true cell locations by reversing the corruption process using gene expression profiles. We denote the model as $\mu_\theta(\cdot,\cdot,\cdot)$ where θ represents all model parameters. LUNA progressively denoises the corrupted cell locations $\tilde{\mathbf{R}}_t$ producing less corrupted locations $\tilde{\mathbf{R}}_{t-1}$ given time step t, more corrupted locations $\tilde{\mathbf{R}}_t$ at the time step t, and the molecular features $\mathbf{X}$.

Eventually, LUNA learns to generate clean cell locations from pure white noise, transitioning from $t = T$ to $t = 0$. To achieve this, the model first predicts the clean cell locations $\hat{\mathbf{R}}$ using the learned network $\mu_\theta(\cdot,\cdot,\cdot)$. It then interpolates between the corrupted locations $\tilde{\mathbf{R}}_t$ and the predicted clean locations $\hat{\mathbf{R}}$, producing less corrupted locations for the next diffusion step:

$$\tilde{\mathbf{R}}_{t-1} = \frac{(1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\overline{\alpha}_t}\tilde{\mathbf{R}}_t + \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\hat{\mathbf{R}}. \tag{16}$$

Inference phase. Once the model $\mu_\theta(\cdot,\cdot,\cdot)$ is trained, LUNA generates cell locations based on the molecular features of the cells in the test set $\mathbf{X}_{ts} \in \mathbb{R}^{m_{ts}\times d}$. The inference process starts by sampling locations from a normal distribution $\epsilon \sim \mathcal{N}(0,\mathbf{I})$, where $\epsilon \in \mathbb{R}^{m_{ts}\times 2}$. The noisy locations, $\mathbf{R}_{ts,T} = \epsilon$ are then refined through a sequence of diffusion steps.

At each step t, the model takes the noisy locations $\mathbf{R}_{ts,t}$, the molecular features $\mathbf{X}_{ts}$, and the time step t as inputs to predict the denoised locations for the next time step $t-1$:

$$\mathbf{R}_{ts,t-1} = \frac{(1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\overline{\alpha}_t}\mathbf{R}_{ts,t} + \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\mu_\theta(\mathbf{R}_{ts,t},\mathbf{X}_{ts},t). \tag{17}$$

This process is repeated for each time step, iterating from $t = \{T,...,0\}$ until the model generates the final clean cell locations at $t = 0$.

Hyperparameters. We use a consistent learning rate of 5×10^{-4} across all experiments, with a maximum diffusion time of 1000 steps and a cosine noise schedule characterized by $\nu = 2$, which dictates the noise addition rate to the coordinates. A detailed analysis of hyperparameter selection and robustness to different values is provided in Supplementary Note 4 and Supplementary Fig. 44 and 45.

Xenium Data Collection. We generate a new spatial transcriptomics dataset using Xenium. Animal husbandry and experimental procedures adhered to the Swiss Federal Veterinary Office guidelines and were authorized by the Cantonal Veterinary Office (cantonal animal license no. VD3499).

Animal Procedures. Eight-week-old C57BL/6J mice were anesthetized with a Ketamine/Xylazine cocktail prepared in sterile saline. Once fully anesthetized, mice were placed in a stereotaxic frame. The head was shaved, Viscotears was applied to protect the eyes, and the scalp was disinfected with liquid betadine solution. After incision, connective tissue was removed to expose the skull and identify the bregma landmark for unilateral dorsal striatum injections.

Stereotaxic coordinates relative to bregma were: anterior-posterior $+0.6$ mm, medial-lateral $+2.0$ mm, and dorsal-ventral -2.6 mm. A small hole was drilled into the skull, preserving the dura mater, and $2\,\mu$l of mouse wild-type pre-formed fibrils (PFFs) at a concentration of $2.5\,\mu\text{g}/\mu\text{l}$ were injected into the right hemisphere at a flow rate of $0.4\,\mu\text{l}/\text{min}$ using a Hamilton syringe. The needle was left in place for 5 minutes post-injection before slow retraction to minimize backflow

and was cleaned from blood to avoid clogging. The incision was sutured, betadine gel was applied, and animals were monitored in a heated recovery cage until fully awake. Post-operative well-being was assessed daily in accordance with animal license scoring sheets.

Tissue Preparation for Xenium. Control and 3-months post-injected mice were euthanized via overdose of a Ketamine/Xylazine cocktail, and brains were rapidly removed. Brains were embedded in Tissue-Tek O.C.T. compound (catalog #4583) within $10 \times 10 \times 5$ mm cryomolds (Tissue-Tek #4565) and frozen by submersion in dry ice-chilled (approximately -70°C) isopentane (Sigma #M32631). Frozen embedded brains were maintained on dry ice until transfer to -80°C storage.

Coronal sections ($10\,\mu$m) were cut on a Leica cryostat and mounted onto Xenium capture slides (10x Genomics, #10006559). For each brain, one coronal section targeting the amygdala (-1.70 mm relative to bregma) and one targeting the striatum ($+0.6$ to -1 mm relative to bregma) were collected. Two sections were placed per capture slide, which were returned to -80°C until processing. The predesigned 248-gene Xenium Mouse Brain Gene Expression panel was profiled across 8 sections on the Xenium *in situ* platform (10x Genomics).

Post-Xenium Immunofluorescence. Following completion of the Xenium runs, slides were stored in PBS at 4°C. For immunofluorescence staining, tissues were permeabilized with PBS containing 0.5% bovine serum albumin (BSA) and 0.1% Triton X-100, blocked with 2% BSA in PBS, and then incubated overnight at 4°C with primary antibody against phosphorylated Ser129 α-synuclein (pSer129 α-synuclein, clone EP1536Y, Abcam ab51253).

Imaging was performed using a Leica DMi8 epifluorescence microscope equipped with a Lumencor SPECTRA X LED light source (nIR, 90-10172), a Leica DFC9000 GTC sCMOS camera (11547007), and a $20\times$ HC PC APO objective (NA 0.8, air), yielding a pixel size of $0.34\,\mu$m. Images were acquired with 10% tile overlap to cover the entire tissue section, collecting between 8 and 12 z-planes with $1\,\mu$m spacing.

Anatomical Registration. Anatomical registration was performed using Aligning Big Brains and Atlases (ABBA, https://go.epfl.ch/abba) [79], which combines affine transformations and nonlinear warping with anatomical border-guided feature registration. Images were automatically aligned to the Allen Brain Atlas CCFv3, and the alignment was subsequently manually refined to ensure precise anatomical correspondence.

Immunofluorescence data processing. 2D images were obtained from the acquired immunofluorescence stacks by performing maximum intensity projection channel-wise on the raw stacks.

The immunofluorescence images were then stitched into a mosaic image using the *Ashlar* [80]. Only the DAPI channel was used to compute the corrected stitched coordinates. The immunofluorescence images were then aligned to the Xenium images by running the software *wsireg* [81], a Python wrapper of *elastix* [82, 83]. Each image transformation was computed using only the DAPI channel of both acquisitions, and the same obtained transformation was applied to the other channels of the immunofluorescence images (*i.e.* no warping was applied to the Xenium data in the process). To calculate the α-synuclein intensity level per cell, we first normalized the aligned α-synuclein channel by applying percentile-based normalization, in which we transformed the intensity values of the image so that the $1st$ percentile is mapped to 0 and the $99.8th$ percentile is mapped to 1. The results from the normalization were then clipped to the $[0, 1]$ intensity range. The median normalized α-synuclein intensity per cell was then calculated by sequentially subsetting the normalized α-synuclein channel using the segmentation masks obtained from Xenium and aggregating accordingly.

Cell Annotation. To annotate the Xenium data at the single cell level, we used the scRNA-seq data published in [84] as a reference for label transfer. We downloaded all the given single cell references (v2 and v3) and then concatenated them to obtain a single *h5ad* data file. During the concatenation process, we randomly subsampled the data by a factor of 4 (*i.e.*, we keep 25% of the single cells) to alleviate posterior computational workload. We then ran *Tangram* [14] for each Xenium dataset using the pre-processed single cell data as reference. We ran *Tangram* in its default settings except for the run mode, for which we used *mode="clusters"*. This mode runs the mapping at the cluster level and substantially reduces computational complexity. We annotated the Xenium data using 34 clusters.

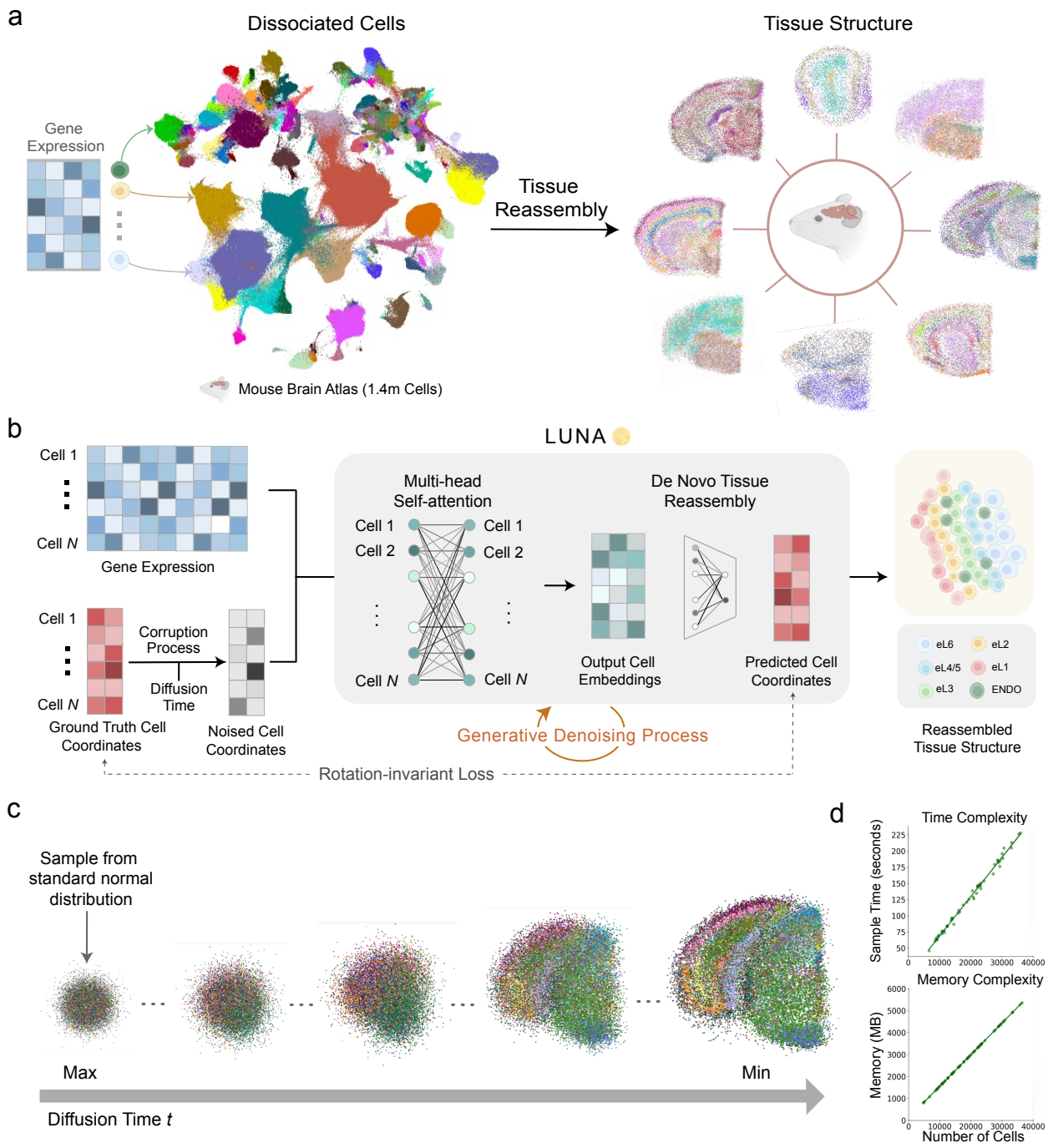

a
Dissociated Cells
Tissue Structure
Gene Expression
Tissue Reassembly
Mouse Brain Atlas (1.4m Cells)
b
LUNA
Cell 1
Cell N
Gene Expression
Cell 1
Cell N
Corruption Process
Diffusion Time
Ground Truth Cell Coordinates
Noised Cell Coordinates
Multi-head Self-attention
Cell 1
Cell 2
Cell N
Cell 1
Cell 2
Cell N
De Novo Tissue Reassembly
Output Cell Embeddings
Predicted Cell Coordinates
Generative Denoising Process
Rotation-invariant Loss
eL6
eL2
eL4/5
eL1
eL3
ENDO
Reassembled Tissue Structure
c
Sample from standard normal distribution
Max
Min
Diffusion Time t
d
Time Complexity
Sample Time (seconds)
Memory Complexity
Memory (MB)
Number of Cells

Figure 1 *(preceding page)*: LUNA is a generative AI model that reassembles cells into complex tissue structures from their gene expressions by learning spatial priors over existing spatially resolved datasets. (a) Given gene expressions of cells, the tissue reassembly process involves predicting the complex tissue architecture. We visualize the UMAP projection of cells based on their gene expression profiles ($1,122$ genes) from the MERFISH whole mouse Allen Brain Cell (ABC) Atlas [27] (left). Each cell is colored according to one of 338 distinct subclasses. LUNA's predictions of the spatial tissue structure is displayed on the right with eight slices randomly chosen from a total of 66 slices. LUNA was trained on spatial transcriptomics dataset of another mouse. **(b)** Overview of the training stage of LUNA. LUNA takes as input spatial transcriptomics data and corrupts the spatial locations of cells by adding noise. Using the attention mechanism, LUNA learns how to position each individual cell with respect to other cells based on their gene expressions. The loss function in LUNA ensures invariance to the rotations and reflections of the predicted slice. **(c)** Visualization of the denoising during the inference stage in LUNA across different diffusion steps on an example slice ($29,079$ cells) from the ABC atlas. The noise state is initialized from a standard normal distribution. Conditioned on the gene expressions of the cells, LUNA progressively removes noise at each diffusion step. **(d)** LUNA achieves linear time and memory complexity during inference relative to the number of input cells. Correlations between inference time and the number of input cells (top) and between reserved memory of the GPU and the number of input cells (bottom) across all 66 slices from the ABC atlas. Each point corresponds to a slice, correlating the cell count with either the time or memory used during the inference stage by LUNA.

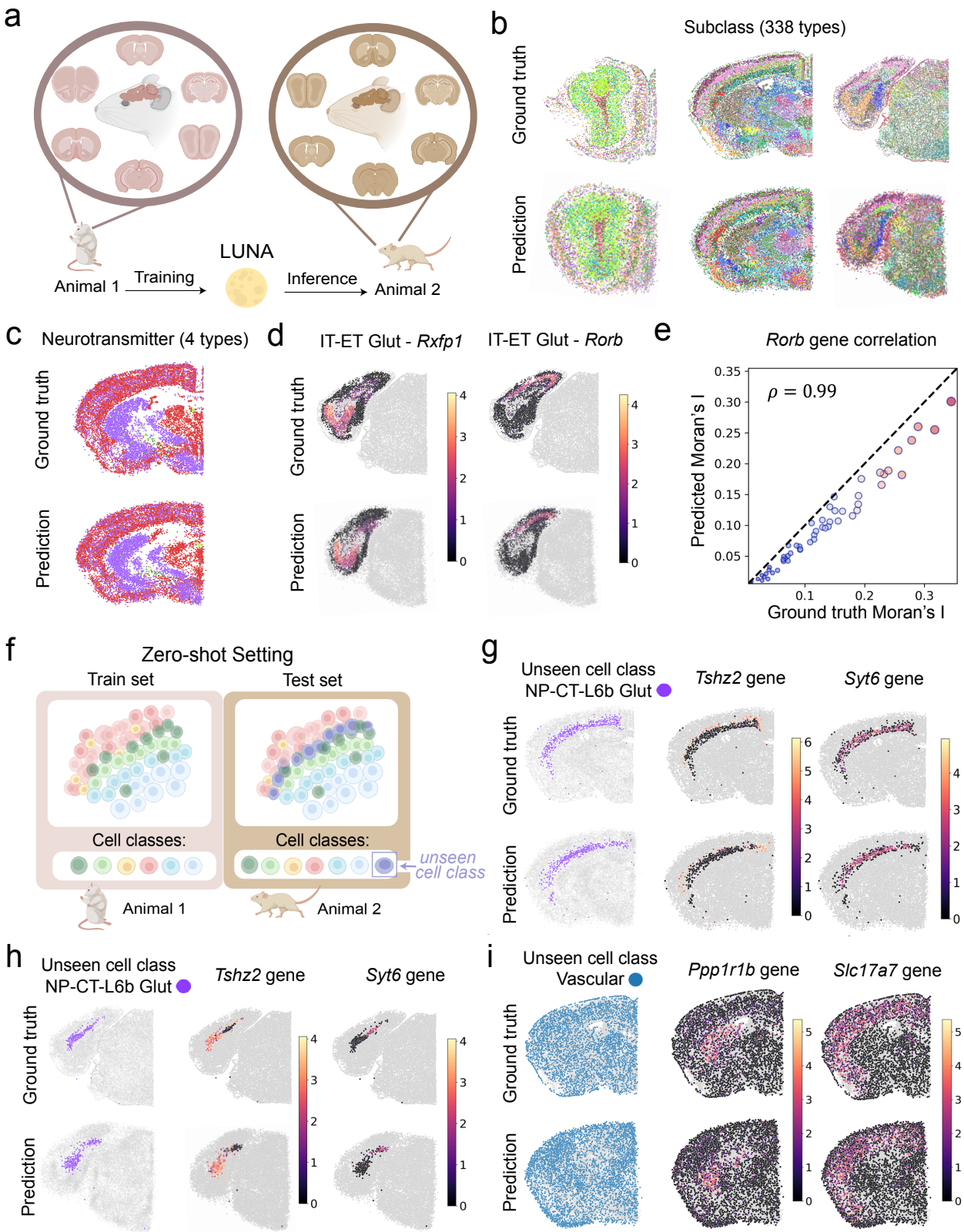

a
Animal 1
Training
LUNA
Inference
Animal 2
b
Subclass (338 types)
Ground truth
Prediction
c
Neurotransmitter (4 types)
Ground truth
Prediction
d
IT-ET Glut - Rxfp1
IT-ET Glut - Rorb
Ground truth
Prediction
e
Rorb gene correlation
ρ = 0.99
Predicted Moran's I
Ground truth Moran's I
f
Zero-shot Setting
Train set
Test set
Cell classes:
Cell classes:
unseen cell class
Animal 1
Animal 2
g
Unseen cell class NP-CT-L6b Glut
Tshz2 gene
Syt6 gene
Ground truth
Prediction
h
Unseen cell class NP-CT-L6b Glut
Tshz2 gene
Syt6 gene
Ground truth
Prediction
i
Unseen cell class Vascular
Ppp1r1b gene
Slc17a7 gene
Ground truth
Prediction

Figure 2 *(preceding page)*: LUNA accurately reconstructs the entire mouse brain atlas of 1.23 million cells. (a) LUNA was trained on all slices from the Animal 1 in the ABC atlas (2.85 million cells) and then applied to predict the tissue structure of cells from the Animal 2 unseen during training (1.23 million cells). **(b)** Examples of tissue reassembly results using LUNA on three representative slices from distinct major brain regions of the MERFISH whole mouse Allen Brain Cell (ABC) atlas [27], comprising $12,610$ cells, $30,931$ cells and $22,010$ cells, respectively. For visualization purposes, three slices were randomly selected from the ABC atlas. Cells are coloured according to their cell type annotations with 338 distinct cell subclasses, showing ground truth (top) and predicted (bottom) locations. **(c)** LUNA's predictions across 4 neurotransmitter types. Ground truth locations (top) and LUNA's predictions (bottom) for a slice with $35,738$ cells . **(d)** Spatial expression patterns of two spatially variable genes in the IT-ET Glut cell type. To select spatially variable genes, we calculated Moran's I values for all genes ($1,122$ in total) based on ground truth locations and identified those with the highest Moran's I value. Representative genes include *Rxfp1* and *Rorb* for IT-ET Glut. Cells from other cell types are shown in gray color. **(e)** Spatial autocorrelation for the *Rorb* gene computed using the Moran's I on the ground-truth locations and the locations predicted by LUNA. Each point represents a single slice, with a total of 66 slices. The slices are color-coded from blue (low) to red (high) based on the true Moran's I values for *Rorb*. Points near the diagonal indicate better spatial pattern preservation by LUNA. Pearson's correlation coefficient (ρ) is displayed in the top left. **(f)** Zero-shot setting evaluation of LUNA. We exclude all cells from a randomly chosen cell class (out of 34 total classes) during model training and evaluate LUNA's predictions on the cells from a held-out cell class, without any further training or refinement of the model. **(g, h)** LUNA effectively generalizes to unseen cells from NP-CT-L6b Glut cell class in the zero-shot setting. NP-CT-L6b Glut cells ($n = 69,641$ in Animal 1) were excluded during training. The model was evaluated on Animal 2, including NP-CT-L6b Glut cells ($n = 29,262$ in Animal 2). The spatial distribution of the excluded class (left), and the spatial expression patterns of the genes *Tshz2* and *Syt6* (middle and right) in the NP-CT-L6b Glut cell type across two representative slices (g, h). Genes *Tshz2* and *Syt6* were identified as the most spatially variable among $1,122$ genes, selected based on the highest Moran's I values calculated from ground-truth cell locations. **(i)** Zero-shot generalization to unseen cells from the vascular cell class. Vascular cells ($n = 447,941$ in Animal 1) were excluded during training. The model was evaluated on Animal 2, including vascular cells ($n = 198,550$ in Animal 2). The spatial distribution of the excluded class (left), and the spatial expression patterns of the genes *Ppp1r1b* and *Slc7a7* (middle and right) in vascular cells for a representative slice.

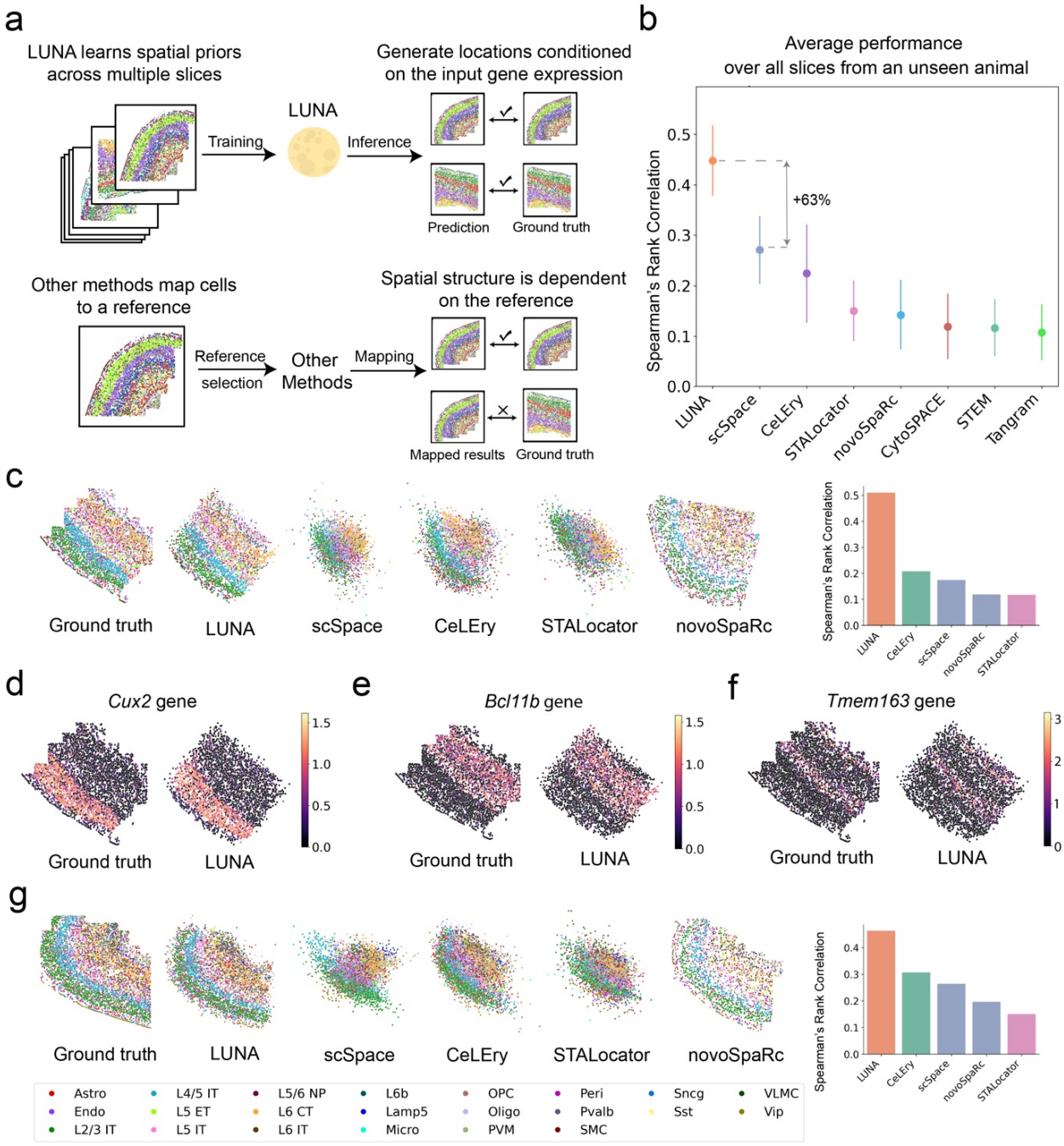

a
LUNA learns spatial priors across multiple slices
Training
LUNA
Inference
Generate locations conditioned on the input gene expression
Prediction
Ground truth
Other methods map cells to a reference
Reference selection
Other Methods
Mapping
Spatial structure is dependent on the reference
Mapped results
Ground truth
b
Average performance over all slices from an unseen animal
Spearman's Rank Correlation
+63%
LUNA
scSpace
CeLEry
STALocator
novoSpaRc
CytoSPACE
STEM
Tangram
c
Ground truth
LUNA
scSpace
CeLEry
STALocator
novoSpaRc
d
Cux2 gene
e
Bcl11b gene
f
Tmem163 gene
g
Astro
Endo
L2/3 IT
L4/5 IT
L5 ET
L5 IT
L5/6 NP
L6 CT
L6 IT
L6b
Lamp5
Micro
OPC
Oligo
PVM
Peri
Pvalb
SMC
Sncg
Sst
VLMC
Vip

Figure 3 *(preceding page)*: LUNA outperforms alternative baselines on the MERFISH mouse primary motor cortex atlas by a large margin. (a) LUNA fundamentally differs from the existing methods. LUNA takes as input different slices and learns spatial priors across all slices during training. In the inference stage, LUNA generates cell locations conditioned on their gene expressions by leveraging learnt spatial priors. In contrast, other methods require the pre-selection of a single reference slice and then map cells to the reference slice. Consequently, the spatial structure of their results heavily depends on the chosen reference slice. **(b)** Performance comparison on the cross-animal generalization task between LUNA and alternative baselines on the MERFISH mouse primary motor cortex atlas [31]. The performance is measured as the median Spearman's rank correlation between the ground truth spatial coordinates and the predicted spatial coordinates over all slices from an unseen mice (31 slices, $118,036$ cells). Higher correlation indicate better prediction performance. Error bars represent standard deviation across 31 slices. **(c)** Visualization of ground-truth locations and predictions using LUNA and alternative baselines for one example slice ($5,235$ cells) using randomly selected reference slice. Colors denote cell class labels. Performance comparison on that slice using Spearman's rank correlation between the ground truth spatial coordinates and the predicted spatial coordinates as a metric (far right). **(d, e, f)** Spatial expression patterns of **(d)** the *Cux2* gene, a marker for layer 2/3 (L2/3 IT), **(e)** the *Bcl11b* gene, a marker for layer 6 corticothalamic (L6 CT), and **(f)** the *Tmem163* gene, a marker for layer 6b (L6b). Ground-truth cell locations are shown in the left plot, while the predictions made by LUNA are displayed in the right plot. LUNA correctly captures spatial variability of gene expression. **(g)** Visualization of ground-truth locations and predictions using LUNA and alternative baselines for one example slice ($5,024$ cells) using the best matched slice as a reference for alternative baselines. The best matched slice was found using prior information about distribution of cell classes which LUNA does not use. Colors denote cell class labels. Performance comparison on that slice using Spearman's rank correlation between the ground truth spatial coordinates and the predicted spatial coordinates as a metric (far right).

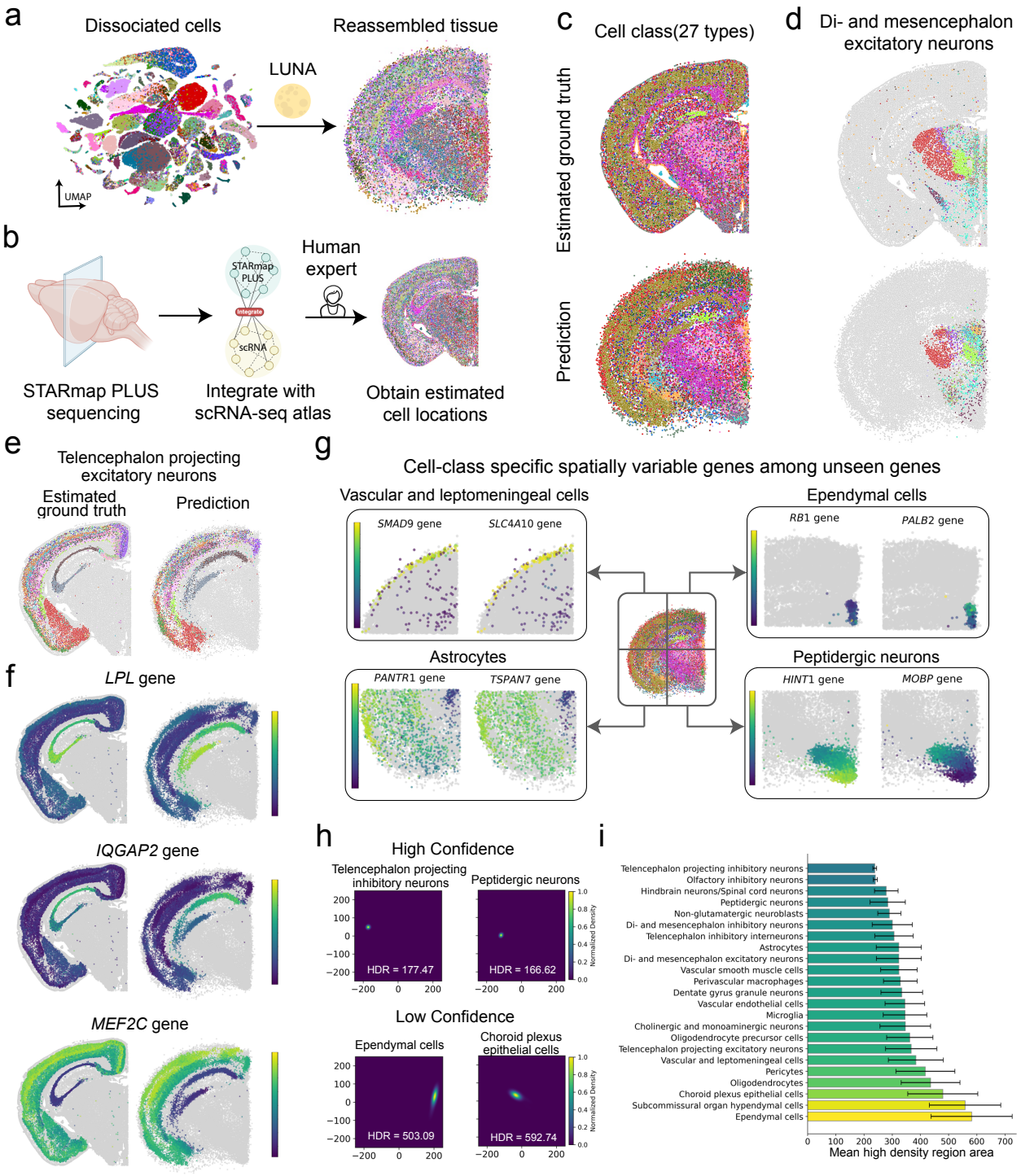

a
Dissociated cells
Reassembled tissue
LUNA
UMAP
b
STARmap PLUS
Integrate
scRNA
Human expert
STARmap PLUS sequencing
Integrate with scRNA-seq atlas
Obtain estimated cell locations
c
Cell class(27 types)
Estimated ground truth
Prediction
d
Di- and mesencephalon excitatory neurons
e
Telencephalon projecting excitatory neurons
Estimated ground truth
Prediction
f
LPL gene
IQGAP2 gene
MEF2C gene
g
Cell-class specific spatially variable genes among unseen genes
Vascular and leptomeningeal cells
SMAD9 gene
SLC4A10 gene
Ependymal cells
RB1 gene
PALB2 gene
Astrocytes
PANTR1 gene
TSPAN7 gene
Peptidergic neurons
HINT1 gene
MOBP gene
h
High Confidence
Telencephalon projecting inhibitory neurons
Peptidergic neurons
HDR = 177.47
HDR = 166.62
Low Confidence
Ependymal cells
Choroid plexus epithelial cells
HDR = 503.09
HDR = 592.74
Normalized Density
i
Telencephalon projecting inhibitory neurons
Olfactory inhibitory neurons
Hindbrain neurons/Spinal cord neurons
Peptidergic neurons
Non-glutamatergic neuroblasts
Di- and mesencephalon inhibitory neurons
Telencephalon inhibitory interneurons
Astrocytes
Di- and mesencephalon excitatory neurons
Vascular smooth muscle cells
Perivascular macrophages
Dentate gyrus granule neurons
Vascular endothelial cells
Microglia
Cholinergic and monoaminergic neurons
Oligodendrocyte precursor cells
Telencephalon projecting excitatory neurons
Vascular and leptomeningeal cells
Pericytes
Oligodendrocytes
Choroid plexus epithelial cells
Subcommissural organ hypendymal cells
Ependymal cells
Mean high density region area

Figure 4 *(preceding page)*: LUNA reassembles the tissue structure of 1.08 million cells from the scRNA-seq mouse central nervous atlas *de novo*. (a) Given the dissociated cells from the scRNA-seq atlas of the mouse central nervous system (CNS) atlas [28], LUNA reassembled the tissue architecture *de novo*. **(b)** Estimated ground truth spatial locations are obtained by sequencing mouse CNS using STARmap PLUS technology and integrating the CNS scRNA-seq dataset with the STARmap PLUS CNS dataset, followed by the manual evaluation of human experts. [33]. We use these estimated locations to validate LUNA's predictions. **(c)** Tissue reassembly of the scRNA-seq mouse central nervous system atlas using LUNA for one example slice ($35,738$ cells). Cells are colored based on the cell classes (27 types). LUNA was trained on all the cells from Animal 1 of the ABC atlas (2.85 million cells) and applied to generate cell locations for a scRNA-seq atlas from the mouse central nervous system (1.08 million cells, 13 coronal slices). **(d-e)** Tissue reassembly of the scRNA-seq mouse central nervous system atlas using LUNA for **(d)** di- and mesencephalon excitatory neurons and **(e)** telencephalon projecting excitatory neurons. The top plot shows estimated ground truth locations obtained by aligning scRNA-seq atlas with the STARmap atlas and the bottom plot shows LUNA's predictions. Cells from each cell type are colored by their sub-molecule class. Cells from other cell types are shown in gray color. LUNA accurately placed cells and preserves the spatial neighborhood of single cells at a very fine-grained level. **(f)** Spatial expression patterns of three spatially variable genes across telencephalon-projecting excitatory neurons. To identify these genes, we calculated Moran's I values for all $10,844$ genes based on LUNA's predictions, selecting three genes with distinct spatial patterns from the top 10 genes with the highest Moran's I values. The representative genes include *LPL*, *IQGAP2*, and *MEF2C*. Cells from other cell types are depicted in gray. **(g)** Cell-class specific spatially variable genes (SVGs) detected using LUNA's predictions among genes that are unseen during training and unique to the scRNA-seq data. Four example cell types are displayed, including both non-neuronal and neuronal cell types. Cells from other cell types are depicted in gray. Genes are randomly selected among their top five SVGs with the highest Moran's I values. **(h)** Kernel density estimation visualizations of example cells with high (top row) and low (bottom row) confidence predictions. We generate 150 samples with different random seeds and compute the area encompassing 80% of predictions, defining this as the high density region (HDR) value for each case cell. Smaller HDR region implies lower uncertainty of a cell's location. Cells with high confidence display sharply peaked distributions (low HDR values), while low-confidence cells exhibit more dispersed distributions. **(i)** Average area of high-density regions computed from kernel density estimation of predicted spatial coordinates for each cell type. Each bar represents a cell type, with its height corresponding to the mean HDR area for cells of that type, and error bars indicating standard deviations. Smaller HDR areas reflect smaller uncertainty in predicting locations of cells.

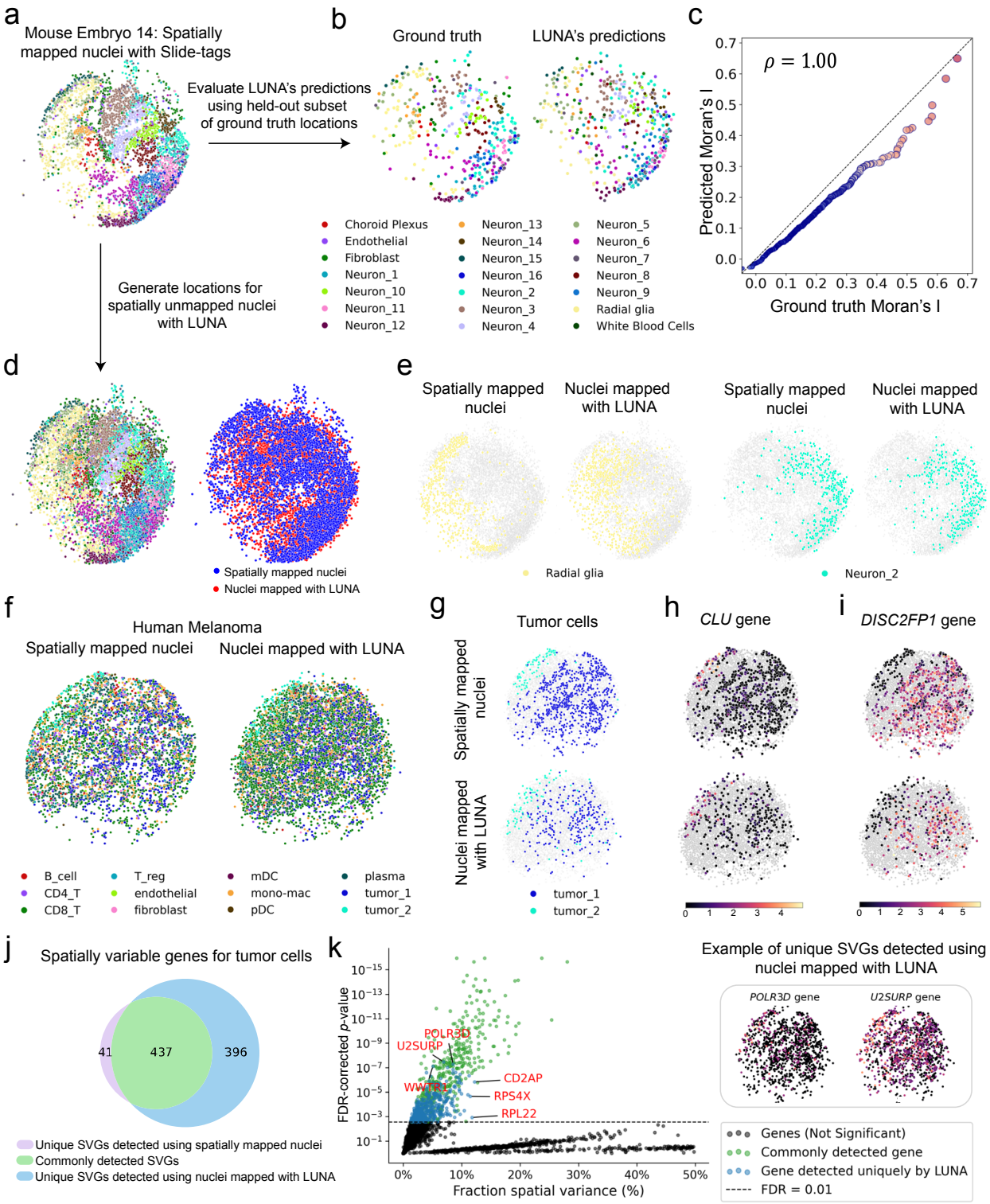
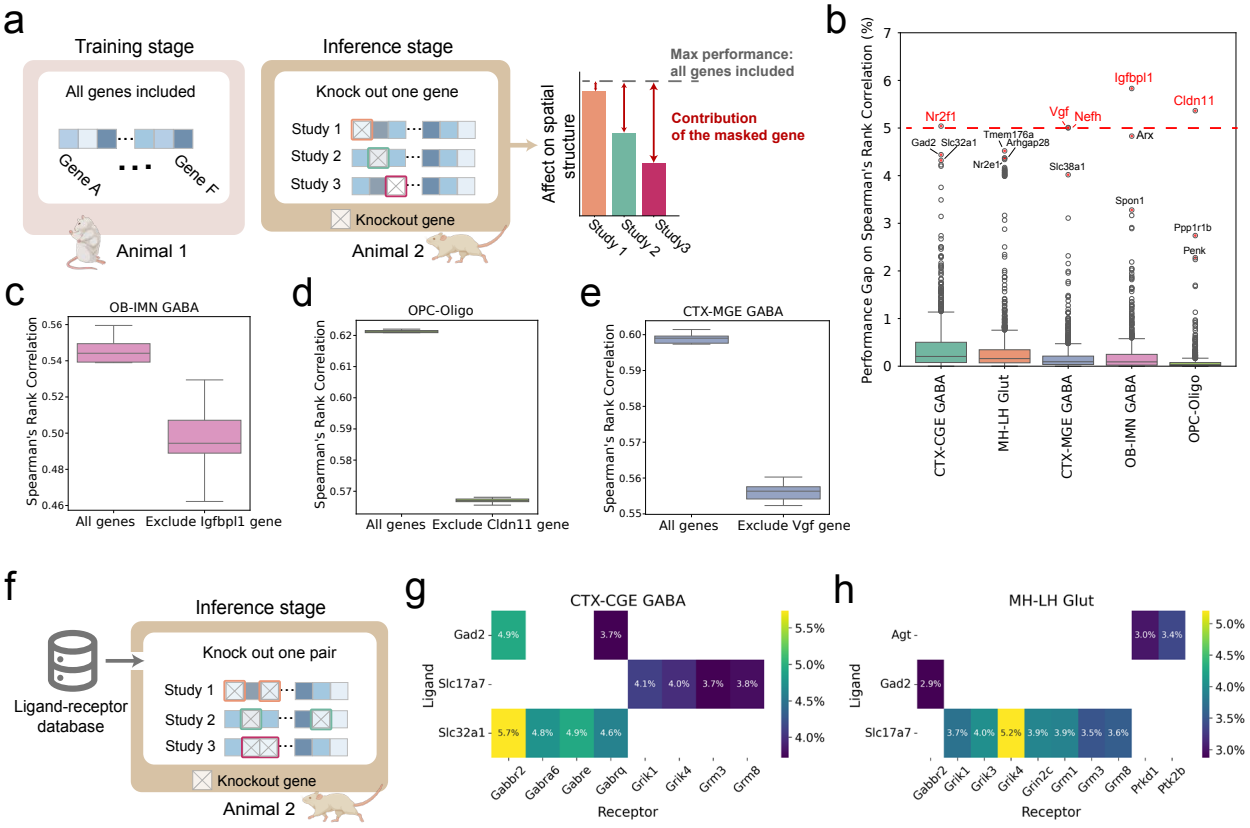

a
Mouse Embryo 14: Spatially mapped nuclei with Slide-tags
Evaluate LUNA's predictions using held-out subset of ground truth locations
Generate locations for spatially unmapped nuclei with LUNA
b
Ground truth
LUNA's predictions
Choroid Plexus
Endothelial
Fibroblast
Neuron_1
Neuron_10
Neuron_11
Neuron_12
Neuron_13
Neuron_14
Neuron_15
Neuron_16
Neuron_2
Neuron_3
Neuron_4
Neuron_5
Neuron_6
Neuron_7
Neuron_8
Neuron_9
Radial glia
White Blood Cells
c
$\rho = 1.00$
Predicted Moran's I
Ground truth Moran's I
d
Spatially mapped nuclei
Nuclei mapped with LUNA
e
Spatially mapped nuclei
Nuclei mapped with LUNA
Spatially mapped nuclei
Nuclei mapped with LUNA
Radial glia
Neuron_2
f
Human Melanoma
Spatially mapped nuclei
Nuclei mapped with LUNA
B_cell
CD4_T
CD8_T
T_reg
endothelial
fibroblast
mDC
mono-mac
pDC
plasma
tumor_1
tumor_2
g
Tumor cells
Spatially mapped nuclei
Nuclei mapped with LUNA
tumor_1
tumor_2
h
CLU gene
i
DISC2FP1 gene
j
Spatially variable genes for tumor cells
41
437
396
Unique SVGs detected using spatially mapped nuclei
Commonly detected SVGs
Unique SVGs detected using nuclei mapped with LUNA
k
FDR-corrected p-value
POLR3D
U2SURP
WWTR1
CD2AP
RPS4X
RPL22
Fraction spatial variance (%)
Example of unique SVGs detected using nuclei mapped with LUNA
POLR3D gene
U2SURP gene
Genes (Not Significant)
Commonly detected gene
Gene detected uniquely by LUNA
FDR = 0.01

Figure 5 *(preceding page)*: LUNA infers locations of spatially unmapped nuclei in the Slide-tags data. (a) Spatially mapped nuclei using Slide-tags from the embryonic mouse brain at embryonic day 14 (E14) containing $4,614$ cells [29]. Cells are coloured according to their cell type annotation. **(b)** LUNA's predictions of nuclei locations agree well with the ground truth locations from Slide-tags. Ground truth locations (left) and LUNA's predictions (right) on the subset of the held-out dataset of the mouse E14 embryonic brain. LUNA was trained on $4,152$ cells with ground truth locations from spatially mapped nuclei and we evaluated its predictions on the held-test set of 462 cells. **(c)** Spatial autocorrelation computed using the Moran's I on the ground truth locations and locations predicted by LUNA on the held-out test set of mouse E14 embryonic brain. Each point represents a gene and 700 highly differentially expressed genes are visualized. Genes are colored on a gradient from blue (low) to red (high) based on their ground truth Moran's I value. Points closer to the diagonal indicate better preservation of spatial patterns by LUNA. The Pearson's correlation coefficient (ρ) is reported on the top left. **(d)** LUNA generated locations of $4,414$ initially unmapped nuclei of the mouse E14 embryonic brain. LUNA was trained on spatially mapped nuclei and then applied to infer locations for the unmapped nuclei. The combined locations of nuclei mapped by LUNA and those spatially mapped by Slide-tags with cells coloured according to their cell type annotation (left). Cells colored based on the source of cell locations (right). **(e)** Nuclei mapped by Slide-tags (left) and nuclei mapped by LUNA for Neuron_2 (yellow color) and Radial glia (cyan color) cell types. All other cells are shown in gray color. **(f)** LUNA enriched the Slide-tags human metastatic melanoma sample from $4,804$ cells to $6,466$ cells by generating locations for spatially unmapped nuclei. The left panel shows the ground truth locations of spatially mapped nuclei used as the training dataset, while the right panel combines these with the nuclei locations predicted by LUNA. **(g)** LUNA correctly inferred the spatial compartment of two tumor clusters segregated into spatially distinct compartments. Spatially mapped nuclei by Slide-tags (top) and locations of unmapped nuclei generated by LUNA (bottom). Tumor cells are colored according to their type, while other cells are shown in gray color. **(h, i)** Spatial expression patterns of **(h)** the *CLU* gene and **(i)** the *DISC2FP1* gene across spatial compartments of tumor cells. We selected these genes as spatially variable genes (SVGs) for each tumor type. Spatially mapped nuclei by Slide-tags (top) and locations of unmapped nuclei generated by LUNA (bottom). Tumor cells are colored according to expression levels of these genes. Other cells are shown in gray color. **(j)** The number of SVGs identified with with SpatialDE [34] in tumor cells using spatially mapped nuclei and nuclei mapped with LUNA (FDR corrected [35] p-value < 0.01). **(k)** Fraction of variance explained by spatial variation (FSV) versus significance of spatial variation (FDR-corrected p-value) for SVGs. Each dot represents a gene ($n = 9,413$). The dashed horizontal line indicates the significance cutoff (FDR < 0.01). Genes are categorized as non-significant, significant genes commonly detected, and significant genes uniquely detected using nuclei mapped with LUNA. Among genes detected uniquely by LUNA, the top 3 genes that have the lowest FDR-corrected p-value (*POLR3D*, *U2SURP* and *WWTR1* genes) and the top 3 genes that have the highest FSV (*CD2AP*, *RPL22* and *RPS4X* genes) are highlighted in red. The spatial gene expression patterns of *POLR3D* and *U2SURP* genes are visualized (top right).


a
Training stage
All genes included
Gene A
Gene F
Animal 1
Inference stage
Knock out one gene
Study 1
Study 2
Study 3
Knockout gene
Animal 2
Affect on spatial structure
Max performance: all genes included
Contribution of the masked gene
Study 1
Study 2
Study3
b
Performance Gap on Spearman's Rank Correlation (%)
Nr2f1
Gad2
Slc32a1
Tmem176a
Arhgap28
Nr2e1
Vgf
Nefh
Slc38a1
Igfbpl1
Arx
Cldn11
Spon1
Ppp1r1b
Penk
CTX-CGE GABA
MH-LH Glut
CTX-MGE GABA
OB-IMN GABA
OPC-Oligo
c
OB-IMN GABA
Spearman's Rank Correlation
All genes
Exclude Igfbpl1 gene
d
OPC-Oligo
Spearman's Rank Correlation
All genes
Exclude Cldn11 gene
e
CTX-MGE GABA
Spearman's Rank Correlation
All genes
Exclude Vgf gene

f
Ligand-receptor database
Inference stage
Knock out one pair
Study 1
Study 2
Study 3
Knockout gene
Animal 2
g
CTX-CGE GABA
Ligand
Gad2
Slc17a7
Slc32a1
Receptor
Gabbr2
Gabra6
Gabre
Gabrq
Grik1
Grik4
Grm3
Grm8
h
MH-LH Glut
Ligand
Agt
Gad2
Slc17a7
Receptor
Gabbr2
Grik1
Grik3
Grik4
Grin2c
Grm1
Grm3
Grm8
Prkd1
Ptk2b

Figure 6 *(preceding page)*: Identification of genes that have a strong impact on predicting spatial organization. **(a)** We train LUNA on all slices from the Animal 1 in the ABC atlas (2.85 million cells) with the whole gene panel. During inference, we simulate a gene knock out by masking its expression to zero, repeating this process for each gene in the panel ($1,122$ genes). We then assess the contribution of individual genes for predicting spatial organization of Animal 2 with LUNA. To quantify the contribution of the masked gene, we compare the difference in the physical spaces between the full-gene model and the gene-excluded model. **(b)** Performance difference caused by excluding each gene for different cell classes. The performance is measured using the Spearman's Rank Correlation (SRC) coefficient computed within each cell class. The top and middle bars of each box represent the 25% quantile and the mean performance difference, respectively. The five cell classes were selected as the classes whose spatial structure is most affected by excluding specific genes across all cell classes. Genes causing more than a 5% performance change are marked in red. **(c-e)** The effect of different initializations of the initial noise during inference. We conduct 10 experiments with different random seeds, focusing on genes that significantly influence specific cell classes. We compare SRC when using all genes with scenarios excluding **(c)** *Igfbpl1* in OB-IMN GABA, **(d)** *Cldn11* in OPC-Oligo, and **(e)***Vgf* in CTX-MGE GABA. Box plots show the performance across seeds, with the top bar indicating the 25% quantile and the middle bar showing the average performance for each group. **(f)** We simulate knockout of known ligand-receptor gene pairs (332 pairs that overlap with the ABC gene panel) curated from the existing literature [57] and measure their effect on predicting spatial organization with LUNA. By masking one ligand-receptor pair at a time, we investigated how each pair influences the performance. **(g-h)** Influence of ten ligand-receptor pairs with the strongest effect on the spatial organization of **(g)** CTX-CGE GABA neurons and **(h)** MH-LH Glut neurons. The values in each entry represent the performance difference between the maximum performance with all genes and the results following the masking that specific ligand-receptor pair. Entries for gene pairs without documented interactions are blank.

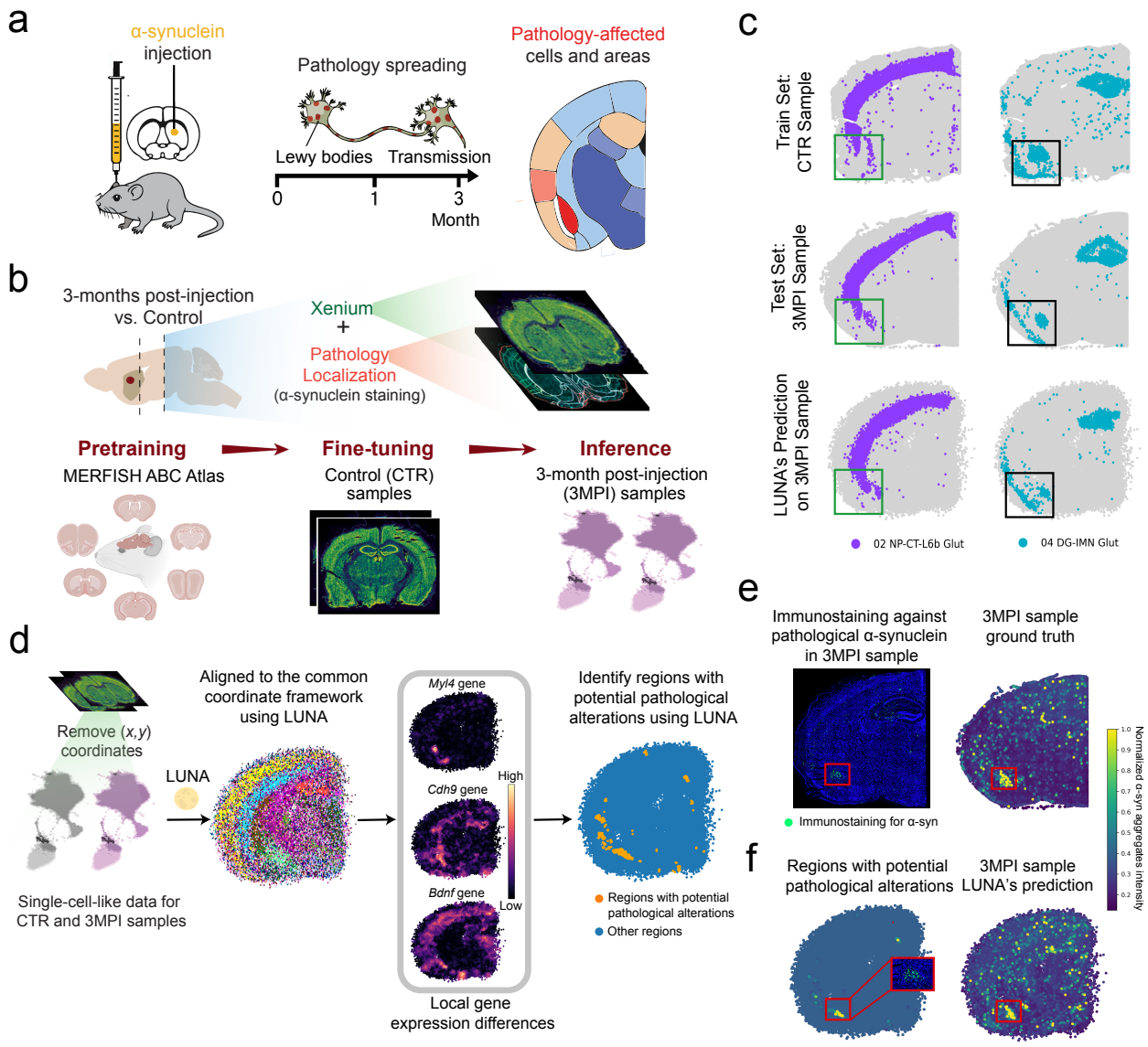

a
α-synuclein
injection
Pathology spreading
Lewy bodies
Transmission
0
1
3
Month
Pathology-affected
cells and areas
b
3-months post-injection
vs. Control
Xenium
+
Pathology
Localization
(α-synuclein staining)
Pretraining
MERFISH ABC Atlas
Fine-tuning
Control (CTR)
samples
Inference
3-month post-injection
(3MPI) samples
c
Train Set:
CTR Sample
Test Set:
3MPI Sample
LUNA's Prediction
on 3MPI Sample
02 NP-CT-L6b Glut
04 DG-IMN Glut
d
Remove (x,y)
coordinates
Single-cell-like data for
CTR and 3MPI samples
LUNA
Aligned to the common
coordinate framework
using LUNA
Myl4 gene
Cdh9 gene
Bdnf gene
High
Low
Local gene
expression differences
Identify regions with
potential pathological
alterations using LUNA
Regions with potential
pathological alterations
Other regions
e
Immunostaining against
pathological α-synuclein
in 3MPI sample
3MPI sample
ground truth
Immunostaining for α-syn
Normalized α-syn aggregates intensity
f
Regions with potential
pathological alterations
3MPI sample
LUNA's prediction

Figure 7 *(preceding page)*: LUNA enables robust spatial reconstruction of diseased samples using a model fine-tuned exclusively on wild-type control tissues. **(a)** Illustration of the α-synuclein (α-syn) pre-formed fibrils (PFF) injection model in the Parkinson's disease mouse model. Misfolded α-syn is transmitted to neighboring neurons, seeding further aggregation and pathology spread in connected regions. To model this pathology, wild-type (WT) mice were unilaterally injected with 2 μl of pre-formed fibrils into the dorsal striatum. **(b)** Schematic drawing of the experiment. Brains from control (CTR) mice and from mice at three months post-injection (3MPI) were extracted, sectioned, and processed using Xenium spatial transcriptomics. For each brain, two distinct anatomical positions were collected and analyzed. The same sections from 3MPI samples were subsequently subjected to immunostaining for misfolded α-syn, providing validation of pathology-affected regions. We use the model pre-trained with MERFISH ABC Atlas, and fine-tune it on CTR samples, and then apply the model to infer spatial structures of 3MPI samples. **(c)** Spatial distributions of neuronal subtypes for the posterior section. Shown are NP-CT-L6b glutamatergic neurons (02 NP-CT-L6b, purple) and DG-IMN glutamatergic neurons (04 DG-IMN Glut, blue) in a training sample (CTR, top), a 3MPI sample (middle), and the LUNA-predicted spatial coordinates for the 3MPI sample (bottom). Gray points denote all the other cells for reference. **(d)** Cells exhibiting differential transcriptomic profiles between CTR and 3MPI samples identified from LUNA's predictions. We removed spatial coordinates from the Xenium dataset and input single-cell-like gene expression data from the CTR and 3MPI sections into LUNA. During inference, LUNA aligns these samples to a common coordinate framework by jointly embedding them. We visualized the genes *Myl4*, *Cdh9*, and *Bdnf*, which were selected for their local spatial differential expression in the amygdala and hippocampus regions. Cells were colored by their local expression differences, with lighter shades indicating greater local shifts. To identify regions with potential pathological alterations, we further computed a unified score across genes with high distributional shifts between the 3MPI and CTR samples. Cells with significant local transcriptomic differences were labeled in yellow, while all others were shown in blue. These highlighted cells indicate regions with localized spatial transcriptomic alterations, potentially associated with pathological changes. **(e)** Immunostaining for α-syn pathology in the 3MPI sample (left) shows α-syn in green, with each dot representing an individual protein aggregate. Ground truth cell locations for the same sample are shown on the right, with each cell colored by the intensity of α-syn fibril staining. **(f)** LUNA enables spatial data-free identification of brain regions with spatial gene expression disregulations that correlate with α-syn pathology. Cells that are identified with potential pathological alternations between 3MPI and CTR (left) and all cells (right) are positioned according to LUNA's predicted locations and colored by α-syn fibril staining intensity. A close-up view highlights the immunostaining of the regions identified by LUNA that potentially associate with pathological alterations, overlaid above the left plot.

Acknowledgements

We are grateful to Ramon Viñas Torné, Shuyang Fan, Clément Vignac, Yiming Qin, Ed Lein, Philippe Schwaller, Evan Macosko, Mor Nitzan and Aviv Regev for valuable discussions. We gratefully acknowledge the support of the Swiss National Science Foundation (grant IC00I0-231922 and starting grant TMSGI2_226252) and Zeiss. Figure elements, including icons of species, were created with BioRender.com.

Author Contributions Statement

T.Y., C.E. and M.B. designed the study, performed research, contributed new analytical tools and analyzed data. T.Y. and C.E. performed experiments and developed the software. N.M., J.F. and S.A. contributed to the codebase. P.F. contributed to the algorithmic framework. A.H., S.N., H.L. and G.L.M. performed the mouse experiments and generated the Xenium data. T.Y., A.D.M., A.H., G.L.M. and M.B. analyzed the Xenium data. T.Y. and M.B. wrote the manuscript with the input from other authors. M.B. supervised the research.

Competing Interests Statement

We declare no competing interests.

Data availability

All analyzed datasets are publicly available. MERFISH Whole Mouse Brain Atlas (ABC Atlas) for Animal 1 is available at https://alleninstitute.github.io/abc_atlas_access/descriptions/Zhuang-ABCA-1.html and for Animal 2 is available at https://alleninstitute.github.io/abc_atlas_access/descriptions/Zhuang-ABCA-2.html. MERFISH Mouse Primary Motor Cortex Atlas is available at the Brain Image Library: https://doi.brainimagelibrary.org/doi/10.35077/g.21. scRNA-seq Mouse Central Nervous System Atlas is available at the Single Cell Portal: https://singlecell.broadinstitute.org/single_cell/study/SCP1830. Slide-tags Datasets for all the tissues are available at the Broad Institute Single Cell Portal under the following accession num-

bers: SCP2170 (mouse E14), SCP2171 (human melanoma), SCP2169 (human tonsil) and SCP2167 (human brain).

Code availability

LUNA was written in Python 3.9 using the PyTorch library. The source code is available on Github at `https://github.com/mlbio-epfl/LUNA`. The project website with links to data and code can be accessed at `https://brbiclab.epfl.ch/projects/LUNA/`.

References

1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

2. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

3. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

4. Consortium*, T. T. S. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

5. Lu, T.-C. *et al.* Aging Fly Cell Atlas identifies exhaustive aging features at cellular resolution. *Science* **380**, eadg0934 (2023).

6. Li, H. *et al.* Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).

7. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nature Medicine* **29**, 1563–1577 (2023).

8. Hickey, J. W. *et al.* Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 (2023).

9. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

10. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods* **15**, 932–935 (2018).

11. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).

12. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology* **33**, 503–509 (2015).

13. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).

14. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods* **18**, 1352–1362 (2021).

15. Vahid, A. *et al.* High-resolution mapping of the spatial organization of single cells in tissues using CytoSPACE. *Nature Methods* **20**, 362–373 (2023).

16. Hao, M. *et al.* STEM enables mapping of single-cell and spatial transcriptomics data with transfer learning. *Communications Biology* **7**, 56 (2024).

17. Qian, J. *et al.* Reconstruction of the cell pseudo-space from single-cell RNA sequencing data with scSpace. *Nature Communications* **14**, 2484 (2023).

18. Li, S., Shen, Q. & Zhang, S. Spatial transcriptomics-aided localization for single-cell transcriptomics with STALocator. *Cell Systems* **16** (2025).

19. Zhang, W. *et al.* Leveraging single-cell spatial transcriptomics with CeLEry for spatial reconstruction. *Nature Biotechnology* **41**, 245–256 (2023).

20. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).

21. Moriel, N. *et al.* NovoSpaRc: Flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Communications* **12**, 1–12 (2021).

22. Vaswani, A. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

23. Shen, Z., Zhang, M., Zhao, H., Yi, S. & Li, H. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021).

24. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning* (2015).

25. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* (2020).

26. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* (2019).

27. Zhang, M. *et al.* Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).

28. Zeisel, A. *et al.* Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).

29. Russell, A. J. *et al.* Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature* **625**, 101–109 (2024).

30. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).

31. Zhang, M. *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).

32. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods* **16**, 1289–1296 (2019).

33. Shi, H. *et al.* Spatial atlas of the mouse central nervous system at molecular resolution. *Nature* **622**, 552–561 (2023).

34. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nature Methods* **15**, 343–346 (2018).

35. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

36. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* **49**, D545–D551 (2021).

37. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell Systems* **1**, 417–425 (2015).

38. Schulze, A., Oshi, M., Endo, I. & Takabe, K. MYC targets scores are associated with cancer aggressiveness and poor survival in ER-positive primary and metastatic breast cancer. *International Journal of Molecular Sciences* **21**, 8127 (2020).

39. Pelletier, J., Thomas, G. & Volarević, S. Ribosome biogenesis in cancer: new players and therapeutic avenues. *Nature Reviews Cancer* **18**, 51–63 (2018).

40. Deng, L. *et al.* MYC-driven U2SURP regulates alternative splicing of SAT1 to promote triple-negative breast cancer progression. *Cancer Letters* **560**, 216124 (2023).

41. Zheng, X. *et al.* Integrative bioinformatics and experimental analyses identify U2SURP as a novel lactylation-related prognostic signature in esophageal carcinoma. *Immunologic Research* **73**, 1–18 (2025).

42. Hernández-Gómez, C., Hernández-Lemus, E. & Espinal-Enríquez, J. The role of copy number variants in gene co-expression patterns for luminal b breast tumors. *Frontiers in Genetics* **13**, 806607 (2022).

43. Wei, J. *et al.* The Hippo signaling effector WWTR1 is a metastatic biomarker of gastric cardia adenocarcinoma. *Cancer Cell International* **19**, 1–10 (2019).

44. Rao, S. *et al.* Inactivation of ribosomal protein L22 promotes transformation by induction of the stemness factor, Lin28B. *Blood, The Journal of the American Society of Hematology* **120**, 3764–3773 (2012).

45. Xie, W. *et al.* CD2AP inhibits metastasis in gastric cancer by promoting cellular adhesion and cytoskeleton assembly. *Molecular Carcinogenesis* **59**, 339–352 (2020).

46. Zhang, L. *et al.* CD2AP promotes the progression of glioblastoma multiforme via TRIM5-mediated NF-kB signaling. *Cell Death & Disease* **15**, 722 (2024).

47. Tsofack, S. P. *et al.* Low expression of the X-linked ribosomal protein S4 in human serous epithelial ovarian cancer is associated with a poor prognosis. *BMC Cancer* **13**, 1–12 (2013).

48. Paquet, É. R. *et al.* Low level of the X-linked ribosomal protein S4 in human urothelial carcinomas is associated with a poor prognosis. *Biomarkers in Medicine* **9**, 187–197 (2015).

49. Gonda, Y. *et al.* Expression profiles of Insulin-like growth factor binding protein-like 1 in the developing mouse forebrain. *Gene Expression Patterns* **7**, 431–440 (2007).

50. Butti, E. *et al.* Neural precursor cells tune striatal connectivity through the release of IGFBPL1. *Nature Communications* **13**, 7579 (2022).

51. Bronstein, J. M., Tiwari-Woodruff, S., Buznikov, A. G. & Stevens, D. B. Involvement of OSP/Claudin 11 in oligodendrocyte membrane interactions: Role in biology and disease. *Journal of Neuroscience Research* **59**, 706–711 (2000).

52. Quinn, J. P., Kandigian, S. E., Trombetta, B. A., Arnold, S. E. & Carlyle, B. C. VGF as a biomarker and therapeutic target in neurodegenerative and psychiatric diseases. *Brain Communications* **3**, fcab261 (2021).

53. Theunissen, F. *et al.* Intronic NEFH variant is associated with reduced risk for sporadic ALS and later age of disease onset. *Scientific Reports* **12**, 14739 (2022).

54. Zhang, K. *et al.* Imbalance of excitatory/inhibitory neuron differentiation in neurodevelopmental disorders with an NR2F1 point mutation. *Cell Reports* **31** (2020).

55. Tocco, C., Bertacchi, M. & Studer, M. Structural and functional aspects of the neurodevelopmental gene NR2F1: From animal models to human pathology. *Frontiers in Molecular Neuroscience* **14**, 767965 (2021).

56. Maheras, K. J. *et al.* Absence of Claudin 11 in CNS myelin perturbs behavior and neurotransmitter levels in mice. *Scientific Reports* **8**, 3798 (2018).

57. Türei, D. *et al.* Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology* **17**, e9923 (2021).

58. Vanhoutte, D. *et al.* Thrombospondin expression in myofibers stabilizes muscle membranes. *Elife* **5**, e17589 (2016).

59. Henderson, M. X. *et al.* Spread of α-synuclein pathology through the brain connectome is modulated by selective vulnerability and predicted by network analysis. *Nature neuroscience* **22**, 1248–1257 (2019).

60. Henderson, M. X., Trojanowski, J. Q. & Lee, V. M.-Y. α-Synuclein pathology in Parkinson's disease and related α-synucleinopathies. *Neuroscience letters* **709**, 134316 (2019).

61. Fan, Y. & Xiao, S. Progression rate associated peripheral blood biomarkers of Parkinson's disease. *Journal of Molecular Neuroscience* **65**, 312–318 (2018).

62. Wu, H. *et al.* A single-cell atlas of the substantia nigra reveals therapeutic effects of icaritin in a rat model of Parkinson's disease. *Antioxidants* **13**, 1183 (2024).

63. Wang, Y., Liu, H., Zhang, B.-S., Soares, J. C. & Zhang, X. Y. Low Bdnf is associated with cognitive impairments in patients with Parkinson's disease. *Parkinsonism & related disorders* **29**, 66–71 (2016).

64. Palla, G. *et al.* Squidpy: A scalable framework for spatial omics analysis. *Nature Methods* **19**, 171–178 (2022).

65. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* **42**, 293–304 (2024).

66. Brbić, M. *et al.* Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods* **19**, 1411–1418 (2022).

67. Clifton, K. *et al.* STalign: Alignment of spatial transcriptomics data using diffeomorphic metric mapping. *Nature Communications* **14**, 8123 (2023).

68. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053–1058 (2018).

69. Schott, M. *et al.* Open-ST: High-resolution spatial transcriptomics in 3D. *Cell* **187**, 3953–3972 (2024).

70. Segol, N. & Lipman, Y. On universal equivariant set networks. In *International Conference on Learning Representations* (2020).

71. Torres-Mendez, L. A., Ruiz-Suarez, J. C., Sucar, L. E. & Gómez, G. Translation, rotation, and scale-invariant object recognition. *IEEE Transactions on Systems, Man, and Cybernetics* **30**, 125–130 (2000).

72. Romero, D., Bekkers, E., Tomczak, J. & Hoogendoorn, M. Attentive group equivariant convolutional networks. In *International Conference on Machine Learning* (2020).

73. Wu, Y., Löwe, H., Carricato, M. & Li, Z. Inversion symmetry of the Euclidean group: theory and application to robot kinematics. *IEEE Transactions on Robotics* **32**, 312–326 (2016).

74. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations* (2021).

75. Vignac, C., Osman, N., Toni, L. & Frossard, P. Midi: Mixed graph and 3D denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 560–576 (Springer, 2023).

76. Chan, S. H. Tutorial on diffusion models for imaging and vision. *arXiv preprint arXiv:2403.18103* (2024).

77. Xu, M. *et al.* GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations* (2022).

78. Hoogeboom, E., Satorras, V. G., Vignac, C. & Welling, M. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning* (2022).

79. Chiaruttini, N. *et al.* ABBA+ BraiAn, an integrated suite for whole-brain mapping, reveals brain-wide differences in immediate-early genes induction upon learning. *Cell Reports* **44** (2025).

80. Muhlich, J. L. *et al.* Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* **38**, 4613–4621 (2022).

81. Patterson, H. Multi-modal whole slide image registration (wsireg) (2022).

82. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging* **29**, 196–205 (2010).

83. Shamonin, D. *et al.* Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Frontiers in Neuroinformatics* **7** (2014).

84. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).