UNIVERSITY OF CALIFORNIA,
IRVINE

Modeling and Deep Learning of Cellular Transcriptome and Epigenetic Regulations

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational, and Systems Biology

by

Honglei Ren

Dissertation Committee:
Professor Qing Nie, Co-Chair
Associate Professor Elizabeth L. Read, Co-Chair
Associate Professor Timothy L. Downing

2023

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First, I would like to thank my advisor Dr. Qing Nie, who always encourage me, give me the freedom and support to explore my interested directions, and teach me skills for being a independent researcher. I would also like to thank my advisor Dr. Elizabeth Read, who guided me to do research step by step, always be patient to answer all my questions, spending time to discuss the new results and ideas. I would also thank to my great collaborator and defense committee member Dr. Timothy Downing, for taking the time to evaluate this dissertation and providing valuable feedback.

I would like to thank the wonderful collaborators I have had a chance to work with, these include Dr. Timothy Downing, Dr. Marcelo Wood, Dr. Julien Morival, Dr. Zixuan Cang, Dr. Benjamin Walker, Dr. Cameron Gallivan, Carlene Chinn, Dr. Wei Zhao. Their contributions were of great importance in assembling the research presented in this dissertation.

I would like to thank the all of the current and previous Nie lab members (Benjamin Walker, Zixuan Cang, Axel Almet, Matt Karikomi, Peijie Zhou, Yutong Sha, Yuchi Qiu, Lihua Zhang, Suoqin Jin, Shuxiong Wang, Yangyang Wang) and Read Lab members (Cameron Gallivan, Rob Taylor, Kojo Bonsu, Junhao Gu) for their help and conversations.

I would also like to thank the funding sources that have supported me during my time at the University of California, Irvine. These include the National Science Foundation and the Simons Foundation. As well as on-campus centers that provided the facilities and resources to perform the research presented, including the NSF-Simons Center for Multiscale Cell Fate Research, the Center for Complex Biological Systems. In addition, the work in this dissertation was funded by National Science Foundation grant DMS176372 and DMS1715455, the National Institutes of Health grants U01AR073159, R01DE030565, and P30AR075047, and a Simons Foundation Grant (594598 QN).

I would like to thank The Royal Society, Springer Nature for permission to reprint portions.

Finally, I would like to thank to my family, who give me their unconditional trust and love to my PhD career and decision. Also, thanks to my friends who encourage and inspire me during my PhD, you provided a lot of support and relief during my most stressful times.

VITA

Honglei Ren

EDUCATION

Doctor of Mathematical, Computational and System Biology **2023**
University of California, Irvine *Irvine, CA*

Master of Software Engineering **2018**
BeiHang University *Beijing, China*

Bachelor of Software Engineering **2015**
BeiHang University *Beijing, China*

RESEARCH EXPERIENCE

Graduate Student Researcher **2018–2023**
Read Lab & Nie Lab at University of California, Irvine *Irvine, California*

Research on DNA methylation, Spatial Transcriptomics

REFEREED JOURNAL PUBLICATIONS

1. **Ren, H.**, Walker, B.L., Cang, Z. et al. *Identifying multicellular spatiotemporal organization of cells with SpaceFlow.* Nat Commun 13, 4076 (2022).

2. **Ren H.**, Taylor, R.B., Downing, T.L. and Read, E.L., 2022. *Locally correlated kinetics of post-replication DNA methylation reveals processivity and region specificity in DNA methylation maintenance.* J. R. Soc. Interface, 19(195), p.20220415.

SOFTWARE

SpaceFlow github.com/hongleir/SpaceFlow
A Deep Learning method for identifying spatiotemporal patterns and spatial domains from Spatial Transcriptomic (ST) data

DNAmCorr github.com/Read-Lab-UCI/DNA-methylation-kinetics-correlation
Code for the paper of Locally-correlated kinetics of post-replication DNA methylation reveals processivity and region-specificity in DNA methylation maintenance

ABSTRACT OF THE DISSERTATION

Modeling and Deep Learning of Cellular Transcriptome and Epigenetic Regulations

by

Honglei Ren

Doctor of Philosophy in Mathematical, Computational, and Systems Biology

University of California, Irvine, 2023

Professor Qing Nie, Co-Chair; Associate Professor Elizabeth L. Read, Co-Chair

Cellular process is meticulously regulated by means of transcription and additional epigenetic mechanisms. This regulation is further complicated by the communication between cells, which coordinate gene expression across multiple cells and tissues. Epigenetic modifications such as DNA methylation, histone modification can induce heritable changes in gene expression without alterations to the DNA. Dysregulation of epigenome or transcriptome often leads to various diseases, such as cancer, developmental disorders, and neurodegenerative diseases. Recent advances in single-cell RNA sequencing (scRNA-seq) and spatially resolved sequencing has provided us unprecedented insights into cellular heterogeneity, developmental trajectories, spatial organization, and cell-cell communications. However, how the cells are spatially and temporally regulated through transcriptome, epigenome, and intercellular communication, is still not clearly understood. In this dissertation, we studied this issue from three distinct perspectives. First, we investigated the maintenance of DNA methylation throughout the cell cycle. Specifically, by analyzing experimental data, we found the post-replication methylation maintenance rates are correlated between nearby CpGs in a region-specific manner. Through stochastic modeling, we derived evidences for genome-wide methyltransferase processivity in cells, and developed an approach to infer lengthscales of linear diffusion of DNA-binding proteins using the rate correlation. In the second project, we developed a deep learning method to identify the spatiotemporal organization of cells.

We further proposed the pseudo-Spatiotemporal Map (pSM) as a spatial counterpart of the pseudotime in scRNA-seq. We validated the accuracy of the pSM using public datasets and demonstrated its usefulness in revealing developmental sequences and providing insights into cancer progression. In the last project, we developed a method for summarizing the major cell-cell communication (CCC) patterns in spatial transcriptome data, which can greatly reduce the analysis burden for researchers who have to analyze thousands of spatial CCC patterns from ligand-receptors. This method also offers new biological insights into CCC through the interpretation for these patterns. For example, we provided the pattern associated ligand-receptors (LRs) which helps to explore potential interacting ligand-receptor network in a tissue specific manner. Overall, our studies provide an approach to reveal the DNA methylation dynamics in post-replication; a method to unravel the dynamics of spatiotemporal organization of cells; and a tool to understand how cells communicate to perform specific functions in tissue.

Chapter 1

Introduction

Cellular processes are regulated by a complex interplay between the transcriptome and epigenome. The transcriptome is the complete set of expressed genes in a cell or tissue, and it reflects the underlying genetic program that is executed by the cell. The epigenome encompasses all the chemical modifications to DNA and histones that control gene expression. This regulation complexity is further complicated by the communication between cells, which coordinate gene expression across multiple cells and tissues.

The transcriptome reflects the active genes in a cell or organism and their level of expression. RNA molecules can be transcribed from both protein-coding genes and non-coding regions of the genome, with the latter often playing regulatory roles in gene expression. RNA molecules can be further modified and processed to produce various types of functional RNA molecules, including messenger RNA, transfer RNA, and ribosomal RNA.

The epigenome includes modifications such as DNA methylation, histone modifications, and non-coding RNA molecules that can influence gene expression by affecting the accessibility of the DNA to the transcriptional machinery. These modifications can be dynamic and respond to environmental stimuli, leading to changes in gene expression that can drive developmental

processes or contribute to disease states.

Cell-cell communication is a fundamental process that allows cells to coordinate their activities and interactions with one another. This process plays a crucial role in various biological processes, including development, tissue homeostasis, and disease progression [1]. The mechanism of cell-cell communication involves the transmission of signaling molecules, such as as ligands, from one cell to another, which bind to specific receptors on the target cell and initiate a cascade of intracellular events leading to a cellular response. This ability enables cells to adapt to their environment, respond to stimuli, and perform specialized functions.

Spatial organization and temporal dynamics of cells play crucial roles in various biological processes such as development and cancer [2]. During development, cells undergo complex spatial and temporal changes, including proliferation, migration, differentiation, to form functional tissues and organs. Similarly, the proliferation, invasion, and metastasis of cancer also involves complicated spatial and temporal dynamics in tumor cells. These processes require precise coordination of cell behaviors in both space and time. Understanding these processes requires detailed knowledge of the molecular mechanisms that regulate the spatial and temporal organization of cells.

Recent advances in technology, such as single-cell RNA sequencing and spatially resolved transcriptomics, have enabled the analysis of gene expression patterns at the single-cell level and in their spatial context. This allows for the identification of cellular states and interactions that contribute to spatial organization and temporal dynamics [3]. Additionally, imaging techniques, such as live-cell imaging, provide insights into the dynamic behavior of cells in real-time.

However, how the cells are spatially and temporally regulated through transcriptome, epigenome, and intercellular communication, is still not clearly understood. In this dissertation, we studied this issue from three distinct perspectives. Specifically, we investigate the mechanism

of DNA methylation maintenance in post-replication through bioinformatic analysis and stochastic modeling; Additionally, we develop approaches to study the spatiotemporal dynamics of cells during development and disease progression using spatial transcriptome data. Lastly, we build tools to identify major cell-cell communication patterns in spatial.

Chapter 2

Locally correlated kinetics of post-replication DNA methylation reveals processivity and region specificity in DNA methylation maintenance

2.1 Introduction

DNA methylation is an important epigenetic modification that plays a critical role in development, aging, and cancer, and it is well conserved among most plants, animals and fungi [4, 5]. In mammals, DNA methylation occurs predominantly in the cytosine-phosphate-guanine (CpG) dinucleotide context. Across most of the mammalian genome, CpGs occur with low frequency, except for regions called CpG islands (CGIs), which are often associated

with promoters [6]. Methylated promoters are associated with transcriptional repression, pointing to a role for DNA methylation as a stable and heritable chromatin mark to program alternative gene expression states [7, 8].

The inheritance and maintenance of methylation patterns across cell cycles is important in development and throughout organismal lifespan. Methylation patterns encode information related to gene expression [9, 10], differentiation [11], and genomic imprinting [12]. Failure in maintenance and transmission of such patterns can lead to aberrant gene expression, and diseases including cancer [13], developmental abnormalities, and even death [14]. The classical model of DNA methylation maintenance introduced the idea that the symmetrical nature of the CpG dinucleotide provides a biomolecular structure whereby DNA methylation could be inherited across a single CpG site by the activity of a, then posited, "maintenance" methyltransferase enzyme [12, 15]. The mammalian DNA methyltransferase DNMT1 was subsequently found to serve as the primary maintenance enzyme [16]. However, the classical model has been refined in a number of ways based on updated understanding of the biochemical properties and genomic activity of methyltransferases (reviewed in [17]). For instance, a wealth of evidence suggests that the efficiency and specificity of methylation enzymes are not sufficient to support the observed high fidelity of maintenance, within the classical, independent-site model [18, 19, 20, 21, 22].

Interdependence, or coupling, of maintenance methylation activities imparted on CpGs located within close proximity can provide some reconciliation between the known biochemistry of methylation reactions and the observed stability of the genomic methylation landscape. Recent findings of preferential recruitment of DNMT1 [23], and faster maintenance methylation rates [24], at sites with more neighboring hemimethylated CpGs are clearly at odds with an independent-CpG-site model. CpG interdependence has been suggested to occur via various molecular mechanisms, including DNMT1 processivity [25, 26, 27] (in which an enzyme can methylate multiple neighboring CpGs on nascent DNA sequentially) and coop-

erative interactions, e.g., with UHRF1, which localizes, and in turn helps recruit DNMT1 molecules, to hemi-methylated CpGs [28, 29]. Mathematical modeling has also suggested the importance of CpG interdependence, also called collaboration, both in maintenance and *de novo* methylation, for long-term collective stability of methylated and unmethylated genomic regions [30, 31, 22, 32, 33, 34].

A quantitative and mechanistic understanding of CpG interdependence during maintenance methylation *in vivo* is lacking. The genomic lengthscales over which CpG coupling occurs are not well understood. It is not yet known to what extent processivity, versus other mechanisms of CpG interdependence, influences dynamics of maintenance methylation. Nor is it yet well-understood how local genomic context influences these mechanisms *in vivo*. In this study, we address these questions by elucidating CpG-coupled-dynamics in maintenance methylation by use of statistical inference, bioinformatics, and stochastic modeling. We leverage experiments that measured methylation status of nascent-strand CpGs across post-replication timescales, genome-wide [35]. From these data, we infer how the rates with which individual CpGs acquire methylation, post-replication, are correlated on nearby sites in different regional contexts. Using stochastic models and theory of proteins diffusing along DNA, we demonstrate that the rate correlation as a function of genomic distance provides a mechanistic fingerprint for post-replication enzymatic processes. Our method provides a novel way to infer lengthscales of linear diffusion of DNA-binding proteins, and it provides the first direct evidence for genome-wide methyltransferase processivity in cells. Comparing simulations to data allows us to extract quantitative insights from data, including the relative strengths of processive versus non-processive coupling mechanisms in different genomic regions, and the length of processive steps.

2.2 Results

2.2.1 Overview of Method

The methodology of this paper can be summarized as follows. We reanalyzed data from Whole Genome Bisulfite Sequencing (WGBS) [36] and Replication-associated Bisulfite sequencing (Repli-BS)[35] in human Embryonic Stem Cells (hESCs) using a combination of data-driven statistical inference and hypothesis-driven stochastic modeling. First, Maximum Likelihood Estimation was used to infer per-CpG post-replication remethylation rates from Repli-BS experiments, following our previously developed method [37]. We analyzed the correlation of these data-inferred rates on nearby CpGs in different genomic contexts, such as Enhancer, Promoter, etc., to study regional differences in maintenance kinetics. Next, we studied the association of the strength of nearby-CpG remethylation-rate correlations with other local genomic/epigenomic features. To aid interpretation of the experiment-derived correlation functions and their regional differences and associations, we developed region-specific stochastic models of post-replication DNA methylation maintenance kinetics. Using these models, we generated simulated bisulfite sequencing datasets under different mechanistic hypotheses, and we compared the resultant *in silico*-generated rate correlation functions to those of the experimental data, focusing both on qualitative and quantitative features, such as shape and lengthscale. Combining theory with stochastic simulation, we developed a method to separate processive and non-processive contributions to the experiment-derived rate correlations.

2.2.2 Post-replication remethylation rates are correlated on neighboring CpGs, and correlation varies with genomic context

In our previous work, we developed a statistical inference procedure to obtain single-CpG post-replication maintenance methylation rates (here denoted "remethylation rates") from Repli-BS data [37], which further supported the variability of remethylation kinetics across the genome.

In this paper, we focus on local correlation of methylation kinetics, obtained from Repli-BS data, and analyze the data-derived correlation using biophysical models of enzymatic methylation reactions. The data-inferred kinetic parameters quantify the rate of accumulation of methylation at each CpG site across hESCs in the measurement set over the experimental timecourse of 0-16 hours post-replication. We correlate single-CpG remethylation rate constants (denoted k_i, for the rate at the ith CpG, obtained by Maximum Likelihood Estimation, see Methods) on pairs of CpGs, as a function of the genomic distance between them.

DNA methylation levels (also hereon denoted methylation "states", or the fraction of cells exhibiting methylation at an individual CpG site) on neighboring CpGs are also correlated [38, 33]. This methylation *state correlation*, as a function of genomic distance in basepairs, reflects information such as the size of persistently methylated (or unmethylated) domains. As such, it reflects the generally static methylation landscape in a given cell type.

We observe both common and distinct features in the correlation functions, when comparing different genomic regional contexts for both data modalities (i.e., rate-correlations and state-correlations). Common to all regions, and to the genomic average with no region-filtering, remethylation rate-correlation decreases rapidly with distance until approximately 100 bp (red curves in Figure 2.1a). After 100 bp, a slower decay is evident in all regions. However, this steep decrease in correlation and rapid change in decay is not observed in correlation

of methylation state (navy curves), which generally shows a slower (persisting to >1 kb), smooth decay, albeit with wide variation in decay-lengths between regions. The existence of such a pronounced difference between methylation state and rate correlation, persistent in all genomic regions, suggests that the methylation rate correlations are not determined by the methylation landscape itself, and raises the question of what dynamic processes determine rate correlations.

Some features of the correlation functions vary between the genomic regions studied. We quantify the average magnitude of short-range correlation by the mean correlation to 1 kbp. These magnitudes are highly variable across genomic contexts and across modalities. Genome-wide, the magnitude of rate-correlation is low (0.06) compared to that of state-correlation (0.82). Across regions, CGIs (CpG Islands) show the highest magnitude of state-correlation (0.85) while transcription factor binding regions (TFBR) shows the highest magnitude of rate-correlation (0.15). By contrast, the lowest average magnitude of state-correlation and of rate-correlation are both found within SINEs (Short Interspersed Nuclear Elements) regions with 0.18 and 0.035 correlation, respectively. We hereafter refer to these differences in remethylation rate- and state-correlations according to genomic context as "genomic region specificity".

In general, the results for methylation state indicates that methylation on CpGs within 1kb are highly correlated genome-wide. Exceptions are within SINEs and LINEs (Long Interspersed Nuclear Elements), where state-correlation has significantly decayed by 1 kbp.

Although the methylation rate correlations show an apparently uniform sub-100-bp decay across regions, differences are visible in their longer-lengthscale (>100 bp) decay profiles (Figure 2.1b). In particular, CGIs and TFBRs (Transcription Factor Binding Regions) show the strongest long-distance correlation, persisting near or above 0.1 past 1 kbp. We further separated the analyzed CpGs from CGIs and TFBRs into "within" and "across" region pairs (2.1c). "Within" correlation is computed for CpGs that are within one contiguous

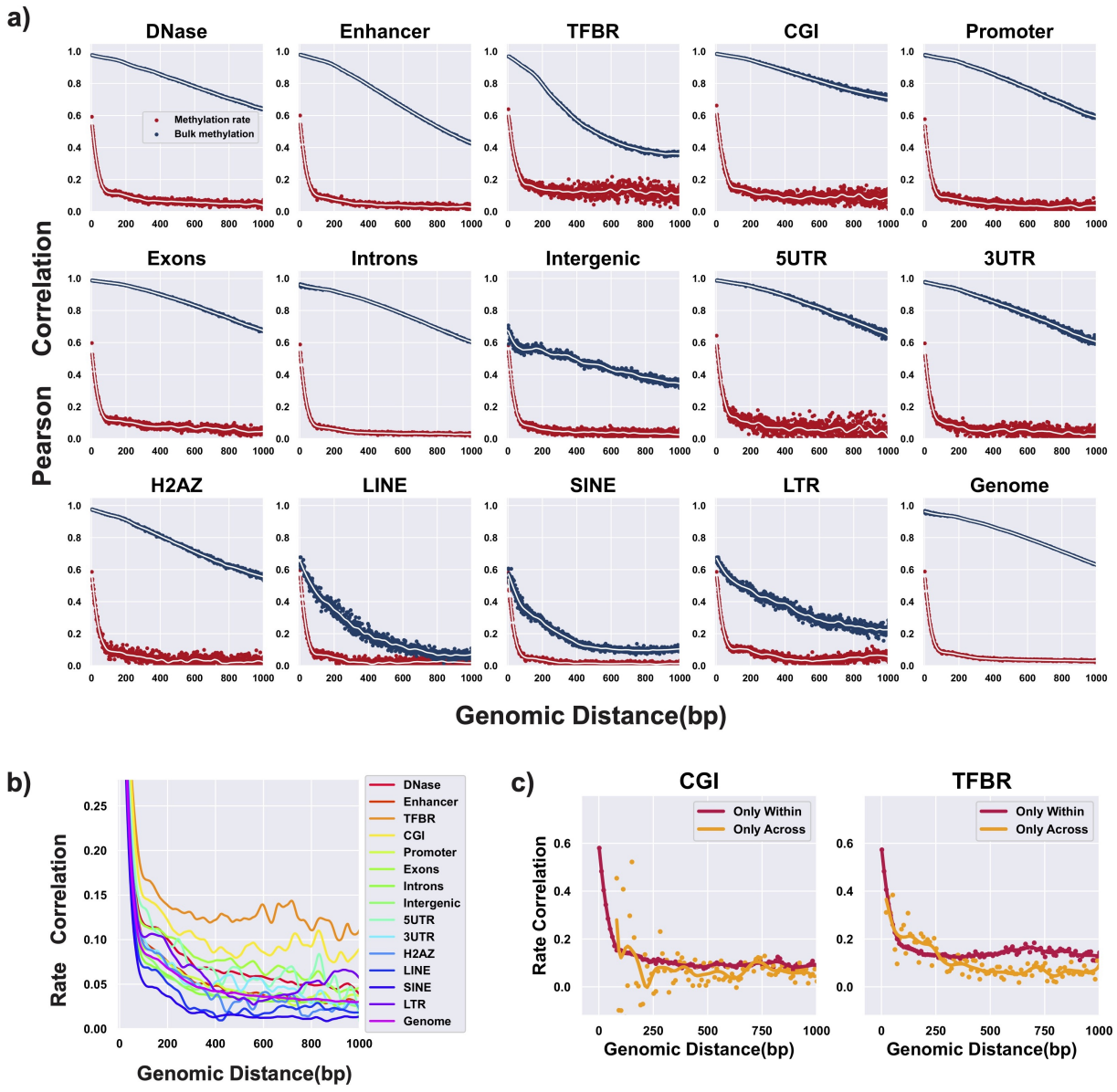

Figure 2.1: **Methylation states and post-replication remethylation rates are correlated on neighboring CpGs, and correlation varies with genomic regional context.** **(a)** Correlation of remethylation rates (red, y-axis) and correlation of percent methylation (i.e., bulk methylation state, blue, y-axis) of nearby CpGs at given genomic distances (x-axis), separated in panels by genomic regional context. Each scatter point at distance d represents the correlation coefficient of rate-pairs $\{k_X, k_Y\}$ derived from all CpG pairs in the region with intervening distance d. Dots: raw rate correlation in base pair resolution, lines: smoothed correlation curve by LOESS(LOcally Estimated Scatterplot Smoothing,[39]) with 100 bp span. **(b)** Correlation of remethylation rates with curves from different genomic regional contexts overlapped and y-axis zoomed. Only the LOESS smoothed curves are shown for clarity. **(c)** Rate correlation functions computed retaining only pairs of sites identified either as within the same localized region (red) or not within the same localized region (orange).

region (as defined by the filtering protocol), whereas "across" correlation only retains pairs of sites that are assigned to non-contiguous regions. The resultant rate correlation functions indicate generally stronger contribution of within- versus across-region correlation (note that the number of datapoints for "across" correlations is lower for short distances, since nearby CpGs are more likely to be in contiguous regions).

In all, these results indicate different magnitudes and decay-lengths of inter-CpG coupling in maintenance methylation. In particular, they point to more pronounced CpG collectivity of maintenance kinetics in contiguous CGIs and TFBRs. That is, neighboring CpGs in CGIs and TFBRs tend to have more correlated remethylation rates and thus more similar methylation across post-replication time, as compared to other regions. However, all regions showed correlated rates to some degree.

2.2.3 Regional correlation of DNA methylation maintenance kinetics is increased with CpG density and chromatin accessibility, and decreased with higher bulk methylation levels

To investigate the factors associated with the observed region specificity, the association between remethylation rate correlation and other local genomic characteristics were examined. For each individual CpG in the dataset, a measure of the local chromatin accessibility is collected from DNaseI hypersensitivity data in ENCODE/OpenChrom (see Materials and Methods). The local CpG density surrounding a given site is calculated by the number of neighboring CpGs within a 500bp window. "Bulk" methylation refers to measurement of DNA methylation in human embryonic stem cells(HUES64) measured by Whole Genome Bisulfite Sequencing(WGBS). We plotted the magnitude of the remethylation rate correlation (0-1000 bp) for each region versus average chromatin accessibility (e.g., DNAse hypersensitivity enrichment) (Figure 2.2a), local CpG density (Figure 2.2b), and WGBS ("bulk")

methylation percentage (Figure 2.2c). We calculated the Pearson correlation coefficients (and associated p-values) between the mean rate correlations and three factors using the 'stats.pearsonr' in the Scipy package [40]. Linear correlation was found between the magnitude of remethylation rate correlation and these three factors. We found the rate correlation is positively correlated with DNase level and CpG density (albeit weakly), while it is inversely correlated with bulk methylation level. These results suggest that the genomic regional differences observed in rate correlation in Figure 2.1 may be driven globally by variation in chromatin accessibility, CpG density, and background methylation landscape.

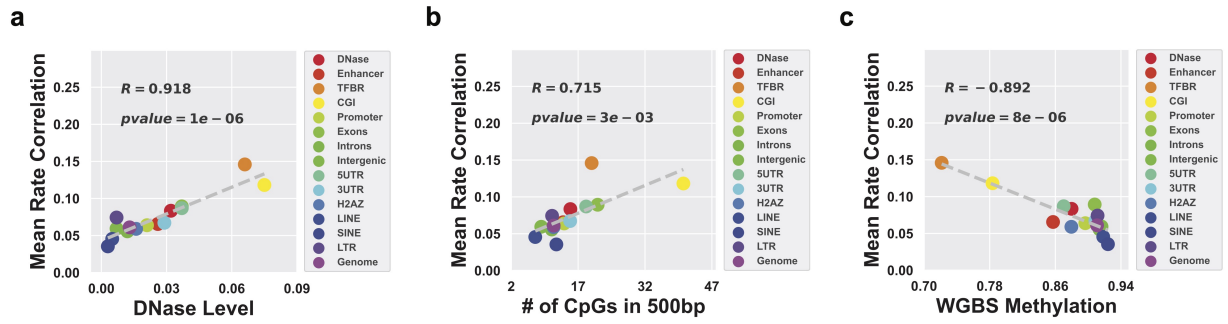

Figure 2.2: **DNA methylation maintenance rates show higher local correlation in genomic regions with higher chromatin accessibility, CpG density and lower bulk methylation levels.** Mean remethylation rate correlation by region, plotted versus other quantified, localized genomic measurements from independent measurements: **(a)** Magnitude of mean remethylation rate correlation (equal to the average over rate correlation for all inter-CpG distance < 1 kbp (i.e., integers from 2- 1000) in the given region) versus mean regional chromatin accessibility, as quantified by DNase level; **(b)** Versus mean local CpG density (mean number of neighboring CpGs within a 500bp window); **(c)** Versus mean WGBS ("bulk") methylation level. Note that the datasets are site-matched, so the analysis is restricted to sites that tend to have intermediate to high methylation, since these are the sites for which remethylation rates are available.

2.2.4 Local methylation correlation varies across post-replication time and shows persistent region specificity

In order to investigate the origins and regional differences of methylation correlation, we focus on three representative genomic regions characterized by high (CGI), medium (Enhancer) and low (SINE) methylation state correlation. In figure 2.3, we plot three types of correlation functions for these three regions: in addition to bulk WGBS correlation and rate correlation, we also plot correlation of methylation state in "nascent" DNA, which contains a subset of measurements from the full temporal Repli-BS dataset, corresponding to the 0-hour timepoint of the original pulse-chase experiment[35]. Thus, "nascent" here refers to methylation readout ≤ 1 hour post-replication.

Note that the red rate and dark blue state correlations are derived from the same data as in Figure 2.1, but in contrast to Figure 2.1, the curves in each panel of Figure 2.3 are site matched. In practice, remethylation rates are available for only a subset of CpGs, as compared to WGBS measurements. This is because the rate constant is undefined where no methylation is measured, and is often unidentifiable when methylation is very low. Thus, there is significant overlap between the sites for which rates are not available and sites for which WGBS percentages are 0 or near 0. We reasoned that some of the methylation state correlation in Figure 2.1 could arise from the bimodal nature of methylated and unmethylated regions. For a more direct comparison to rate correlation on a site-specific basis, we thus filtered to a common set of sites, which in practice retains mostly sites with intermediate to high methylation in WGBS. After filtering to these common sites, we indeed observe some decrease in state correlation (dark blue curves in Figure 2.1a versus 2.3), but rate correlation remained persistently lower than state correlation in all regions.

We observe that methylation state correlation in nascent DNA is generally lower than that in bulk DNA, suggesting that state correlation increases over post-replication time. (Bulk

WGBS measurements reflect temporal variability from cells in various stages of the cell cycle and from differences in replication-timing across the genome (meaning it largely contains matured DNA strands), whereas the nascent data in principle captures reads within one hour post-replication [35]). In all studied regions, the nascent methylation correlation is intermediate between rate and bulk state correlation (with the exception of very short distances in SINEs).

The trend in genomic region specificity of methylation correlation is persistent across the three types of correlation functions. This in turn suggests that the region specificity is persistent across post-replication time. For example, CGIs consistently show the highest correlation compared to the other regions in rate, bulk methylation, and nascent methylation. Conversely, SINE consistently shows the lowest correlation.

These results suggest that the processes that govern coupling (or interdependence) of methylation among neighboring CpGs differ depending on the genomic regional context, and that these processes are region-specific already at early post-replication timepoints. Additionally, the distinct shapes and magnitudes of the CpG-neighbor correlations across time suggest that different processes control CpG-neighbor-interactions at early versus late post-replication times.

We hypothesize that the three correlation modalities can be interpreted as follows: rate correlation reflects the dynamic mechanisms of maintenance methylation, thus shedding light on early post-replication-time processes. In contrast, bulk methylation state correlation largely reflects the steady-state methylation landscape, i.e. reflecting the balance among methyl-reading/writing/erasing processes operating across post-replication time to regulate the methylation landscape, but largely reflecting the stable methylation landscape of a given cell type. Nascent methylation state reflects a mixture of the two, as the experiment "captures" CpGs in transit between their state immediately (up to one hour) post-replication and steady state. In the following sections, we test this hypothesis by use of computer

simulations and model-guided data analysis.

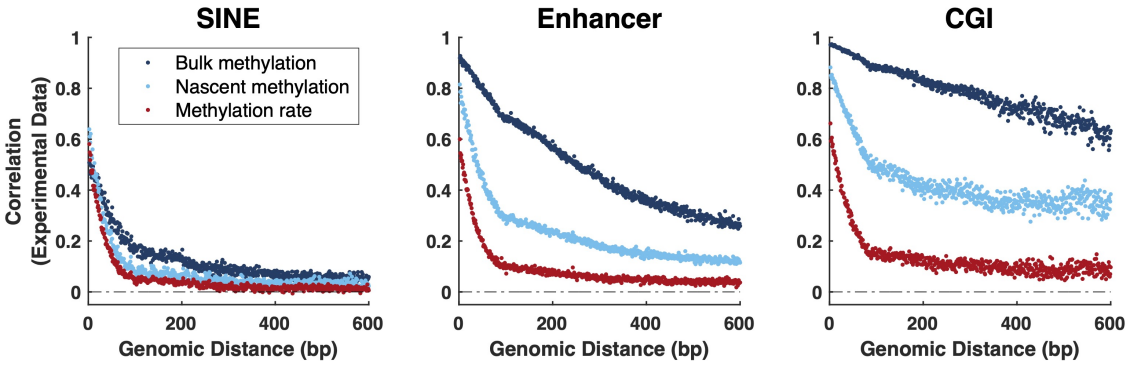

Figure 2.3: **DNA methylation correlation increases over post-replication time and shows persistent region specificity across different data modalities.** Data from three representative regions and three methods of extracting methylation correlation; all three curves in each panel are derived from the same set of CpGs. Dark blue: correlation of methylation state of neighboring CpGs from Whole Genome Bisulfite Sequencing Experiments (WGBS, "Bulk"). Light blue: Correlation of methylation state of neighboring CpGs from WGBS on nascent DNA from Repli-BS-seq experiments (i.e., $<$ 1 hour post-replication) ("Nascent"). Red: Correlation of remethylation rates inferred from Repli-BS-seq experiments. Data for rate-correlation (red) is identical to that of Figure 2.1a. Data for bulk methylation (WGBS) is also derived from the same dataset as Figure 2.1a, but filtered to retain only those CpG sites for which a rate was available (determined by per-site read-depths, see Methods).

2.2.5 Rate correlation provides a mechanistic fingerprint for enzyme kinetics

We use simulations to generate synthetic data mimicking the various bisulfite sequencing datatypes (rates from Repli-BS, nascent methylation from Repli-BS, and bulk methylation state from WGBS). We then compute the correlation functions for the synthetic data. Figure 2.5 shows simulation-derived correlation functions for three representative genomic regions from Chromosome 1 for two mechanistic models termed Distributive and Processive. Briefly, DNMT1 binds to nascent CpGs and catalyzes the addition of a methyl group. In the Distributive model (Figure 2.4b), the enzyme unbinds after the catalytic step, and must

independently rebind to an available hemi-methylated CpG in order to catalyze a subsequent methyl addition. In the Processive model (Figure 2.4c), the enzyme can remain bound to DNA after methylating a CpG, and can travel along DNA (via a one dimensional diffusive random walk) to reach neighboring hemi-methylated CpGs and again catalyze methylation. The random walk occurs with with 1D diffusion coefficient, D, and the enzyme potentially unbinds before reaching its target with rate k_{off}. The kinetic parameters for both models are given in the Supplement.

In both Distributive and Processive mechanisms, the simulated bulk methylation correlation approaches the input methylation landscape in the three regions. This is expected, because the input landscape dictates parental strand methylation in the model, and the model assumes that as time progresses, the nascent strand methylation will ultimately match the parental strand input (i.e., it assumes perfect maintenance in the long-time limit). In contrast, the shapes of remethylation rate correlation are distinct from those of the input methylation landscape, and are dependent on the model mechanism. In all simulated regions, nascent correlation is intermediate between bulk state and rate, in agreement with experiments. The Distributive model produces no correlation in remethylation rate, in agreement with our previous results [37]. In contrast, the Processive model produces a nonzero rate correlation that appears to qualitatively reproduce the experimentally observed rapid decay of correlation with genomic distance up to approximately 100 bp. Neither model reproduces the low but persistent correlation visible at distances greater than 100 bp in CGI and Enhancer in Figure 2.3.

The simulated correlation functions provide a possible explanation for the discrepancy between methylation state and rate correlations observed from experiments (Figures 2.1 and 2.3). Namely, they suggest that the methylation rate correlation shape is dictated by enzymatic maintenance mechanisms, independent of the background methylation landscape. That is, simulations support that mechanistic insights on maintenance methylation can be

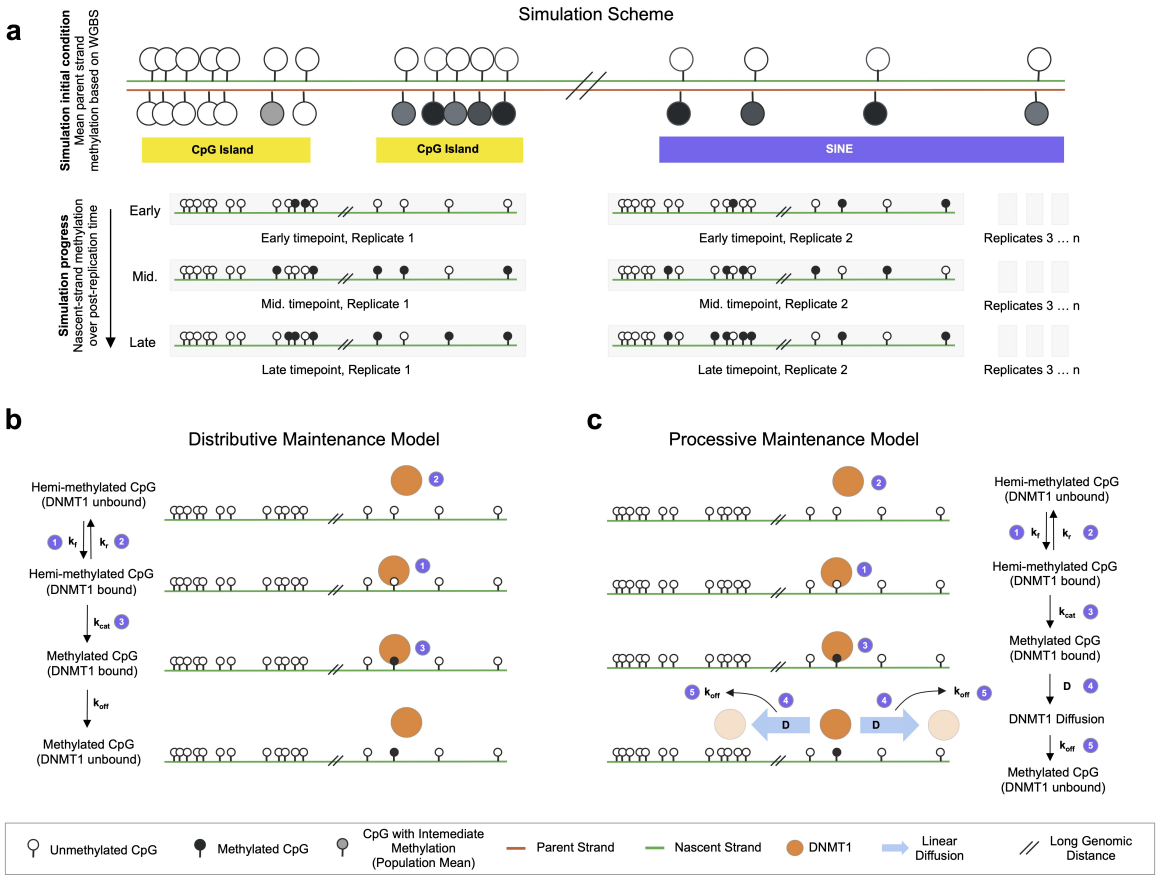

Figure 2.4: **Schematic of stochastic simulation of DNA methylation maintenance in different genomic regions, according to either Distributive or Processive mechanisms. (a)** Each simulation models a strand of N CpGs ($N = 25000$ to 75000) where the CpG positions and the parent-strand methylation at the initial condition are taken from WGBS measurements in a given genomic region (e.g., CGI or SINE). Multiple strand replicates are simulated over time. Immediate-post-replication DNA is assumed to be unmethylated on the nascent strand at all sites, and the binary methylation status of the parent strand sites are sampled probabilistically from the input methylation landscape (mean methylation level in WGBS). During the simulations, DNMT1 targets hemimethylated CpGs for methylation, according to either a Distributive **(b)** or Processive **(c)** mechanism. The Distributive mechanism assumes that the enzyme binds to each hemi-methylated CpG independently. After catalyzing methylation, the enzyme immediately unbinds from DNA (we assume $k_{\text{off}} \gg k_{\text{cat}}$, such that methylation and unbinding are treated as a single reaction). To reach a subsequent CpG, the enzyme must independently rebind with rate k_f. The Processive mechanism assumes that, after catalyzing methylation at a CpG, DNMT1 can remain bound to DNA and reach nearby hemimethylated CpGs by linear diffusion along DNA.

derived using the rate correlation, since it effectively separates correlations introduced by early post-replication-time processes from those operating at longer timescales (and which

dominate the bulk WGBS correlation). Thus, the rate correlation is a useful new quantity, which can be used to distinguish between hypothesized mechanisms. In the processive model, the rapid dropoff of rate correlation is due to the enzyme's intrinsically limited capability to process along DNA over long lengthscales. As such, a pair of CpGs can be strongly correlated in their methylation state (as dictated by parent-strand methylation, and thus reflective of the correlated methylation landscape), while showing low correlation of their kinetic rates, if their distance is outside the enzyme's processivity range. However, the simulations also demonstrate that, while the Processive mechanism partially recapitulates experiment-derived rate correlations, it cannot explain the longer-distance, slow rate correlation decay in CGI.

2.2.6 Rate correlation in Processive model decays exponentially, dependent on 1D diffusion and unbinding rates

We used simulations and theoretical modeling to determine whether quantitative insights could be derived from the experimental rate correlations. We first investigated how rate correlations arising from the Processive mechanism depend on model parameters. In an idealized theoretical model, we find that the rate correlations are related to the diffusion constant D and the unbinding rate, k_{off}, as:

$$Corr(X, Y|d) \propto \mathrm{e}^{-d\sqrt{k_{off}/D}} \tag{2.1}$$

where X, Y are pairs of remethylation rates on sites that are distance d apart (along the DNA strand). See the Supplemental Information for justification of Equation 2.1.

We find a good match between the above analytical theory and the stochastic simulations (Figure S2). We performed regional simulations for varying values of D/k_{off}, and then processed the simulated data through the MLE inference pipeline and computed the correlation

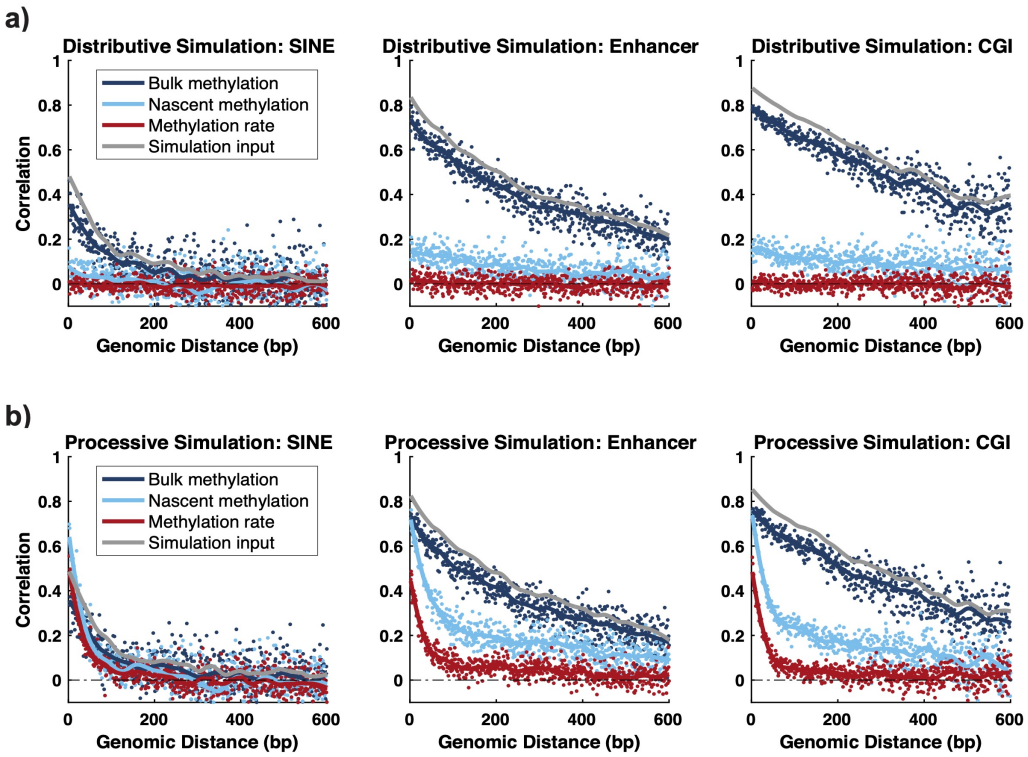

Figure 2.5: **In simulations, rate correlation but not state correlation depends on enzymatic mechanism; only Processive model displays nonzero rate correlation.** Simulated correlation functions for three representative regions (SINE, Ehnhancer, CGI) and using the Distributive (top) or Processive (bottom) mechanism. Simulations are performed using experimental regional methylation landscapes for SINE, Enhancer, and CGI as initial condition inputs (grey curves). The simulations provide synthetic data corresponding to each of the three experimental data modalities, as in Figure 2.3. Synthetic data are processed in the same way as the experimental data to compute the inferred remethylation rates and Pearson correlation functions. Both models recapitulate temporal trends in rate correlation seen in experiments, Figure 2.3 (correlation in rate $<$ nascent state $<$ bulk state). Only the Processive model captures nonzero rate correlation sub-100-bp; neither model captures low, persistent correlation $>$ 100 bp observed in Figure 2.3.

function. We then fit the correlation functions to a single exponential decay, and observe a good match between the fitted decay constant and the theoretically predicted value of $\sqrt{k_{off}/D}$. These results demonstrate that the predicted exponential correlation holds, despite potentially complicating factors in the stochastic model (e.g., the finite time required for diffusion, many CpG sites and multiple enzymes acting simultaneously on one strand, etc.), and uncertainty introduced by the MLE-fitting pipeline. The simulations also predicted that the exponential decay length was not strongly affected by the length of measured reads in the Repli-BS experiments (Figure S3), though some spurious correlation arising from read-length was observed *in silico*. All in all, these results support the finding that processivity, in a linear diffusion model, is consistent with exponential decay of rate correlations obtained from Repli-BS, and that the decay constant can be interpreted as $\sqrt{k_{off}/D}$.

2.2.7 Mathematical model for diffusion-based and region-based components of methylation rate correlation

The Processive model can explain exponential decay of rate correlation, but cannot by itself explain the significant correlation observed past 100 bp in strongly correlated regions such as CGI (Figure **??**). Nor can it explain the association of this correlation to other regional genomic characteristics (bulk methylation levels, CpG density, chromatin accessibility). Based on the observations in Fig **??**, we reasoned that this additional rate correlation can be understood to result from correlated methylation times on neighboring sites due to a variety of regional features, that we describe as

$$Corr_{\text{region}}(X, Y|d) = \phi(\text{features}(d)) \tag{2.2}$$

and refer to as "region-based correlation". That is, this component of the correlation function can be expressed as a function ϕ, which depends somehow on local features, e.g., related to the

chromatin landscape. These features themselves held some distance-dependent correlation. Consideration of additional correlating factors (beyond diffusion) equates to a revision of the idealized theoretical model above. Now, in event 1, if the neighboring site is not reached by diffusion, but rather the two sites are methylated in separate events, their times τ_i and τ_j are nevertheless correlated in a way that depends not only on their distance apart, but also on various other features of their location within the genome. We do not propose a mechanistic model for this additional correlation, but label it as ϕ, and estimate it for each region, based on the data. Given these two contributions, which we term diffusion-dependent (denoted θ) and region-dependent (ϕ), the total correlation from both contributions is given by:

$$TotalCorr(X, Y|d) = \theta + (1 - \theta)\phi. \tag{2.3}$$

where the above equation can be understood by probabilistic arguments, i.e., the probability that a neighbor is reached by diffusion is θ (given by Eq.2.1); additional correlation ϕ is only present when the neighbor is *not* reached by diffusion, with probability $1 - \theta$. This mathematical decomposition of the rate correlation is shown schematically in Figure 2.6a. The model of equations 2.1, 2.2, and 2.3 makes a prediction: that if ϕ (the component of correlation due to genomic regional features) can be estimated from the data, then the remaining correlation (θ) should decay exponentially with distance, and should not depend on local genomic context.

2.2.8 Separation of experimental remethylation rate correlation functions into diffusion-dependent and region-dependent components quantifies genome-wide processivity of DNMT1 *in vivo*

We developed an approach to estimate ϕ as follows. The total experiment-derived correlation functions are computed from a list of cytosine positions and their inferred remethylation rates. Denote x_n^m as the position of site n, identified as belonging to region m (where m is one of the fourteen regions of Figure 2.1, and $n \in [1, N]$ given that data is available for N sites in a region of interest). Denote the remethylation rate as k_n^m, giving a list of pairs $\{x_n, k_n\}_m$ for each region, from which the correlation is computed. We also have additional genomic feature data: let q_n^m, r_n^m, s_n^m denote the measured bulk methylation level, the local CpG density, and the chromatin accessibility acquired from independent datasets (see Methods). We then perform unsupervised k-means clustering on the features $\{q, r, s\}_m$ to obtain clusters of sites that are (i) previously assigned as belonging to the same type of genomic region m, and (ii) share more fine-grained similarity in terms of their bulk methylation level, CpG density, and chromatin accessibility. We then randomly shuffle the nucleotide positions within each subcluster, and denote these shuffled positions as $\tilde{x}_n$. We now have a new list, $\{\tilde{x}_n, k_n\}_m$, where the true nucleotide positions have been randomized, but their reassigned position is still similar to the true position in terms of the features $\{q, r, s\}_m$.

We recalculate the correlation functions for each of the regions with the new lists $\{\tilde{x}_n, k_n\}_m$. We label this new correlation function as the region-dependent component ϕ, reasoning that it captures rate correlation that can be attributed to the regional features, according to the model of Equation 2.2. We then extract θ from the total correlation using equation 2.3. The results of this decomposition are plotted in Figure 2.6b (a more detailed view in regions CGI, Enhancer, and SINE is shown in Supplemental Figure S4). We find that ϕ is more variable

between different genomic regions, as compared to the component θ, which appears nearly uniform across genomic regions and in all regions shows rapid decay within $\approx$ 100 bp. We find that θ is generally well fit by a single-exponential decay and the fitted decay constants are similar across genomic regions, with a mean decay constant of 0.028 (ranging from .023 to .032, in TFBR and SINE, respectively), corresponding to a mean decay length of 36 basepairs (31 to 43 bp, respectively), Figure 2.6c. Importantly, this decomposition approach does not impose any *a priori* assumptions on the functional form of θ. The results confirm our hypothesis, that after removal of the correlation component dependent on regional features of the chromatin landscape, the remaining correlation decays exponentially with distance.

Our model predicts that the fitted exponential decay constant should be reflective of parameters of enzyme diffusion, namely, equal to $\sqrt{k_{off}/D}$. Thus, the model predicts that the decay constant is insensitive to other factors, such as variability of inter-CpG spacing. We confirm this in the experiment-derived correlation functions, by comparing the fitted decay lengths in each region to the inter-CpG distance distributions in each region (Figure 2.6d). Despite the significant differences in CpG density in the different regions, the decay lengths are generally constant, i.e., the median inter-CpG distance is 10 and 161 bp in CGI versus Intergenic regions, with corresponding fitted decay lengths of 39.6 and 35.6 bp, respectively.

Note that the exponential fit is not perfect, as evident by slight discrepancies in fit constants obtained for different window sizes. This discrepancy is due to low, persistent, nonzero longer-range correlation visible in θ in some regions (Figure S4), which we attribute to our model of ϕ, which was based only on three genomic features, and thus likely did not fully account for all region-based correlation. We therefore report fit constants from the 100bp window, to estimate the short-range decay while minimizing contamination from long-range residual correlation. However, the quantitative impact of different fit window sizes on estimated constants is relatively minor, as seen in Figure 2.6c.

All in all, these findings support that θ reflects the diffusion-based contribution to the cor-

relation function, because (i) it is the explicitly distance-dependent part that remains after removing correlation attributed to the other features $\{q, r, s\}_m$, (ii) it decays exponentially, (iii) it is uniform across the genome, consistent with processivity being an inherent property of DNMT1's mode of action, and thus uniform across genomic regions. If θ is interpreted as reflecting processivity of the enzyme according to a 1D diffusion model, then diffusion parameters can be obtained from the experiment-derived correlation function. We thus estimate D/k_{off} to be 1300 bp^2 and the length of a processive step of DNMT1 to be 36 basepairs, on average, across the genome.

2.3 Discussion

2.3.1 Summary of key results and methodological contribution

We have analyzed correlations among CpG sites in the genome obtained from estimated kinetics of post-replication DNA methylation and from WGBS in hESCs. We find that post-replication remethylation rates on nearby CpGs are correlated, and the nature of this correlation can shed light on molecular mechanisms of maintenance methylation, when analyzed in conjunction with stochastic simulation and mathematical models. We summarize our key results as follows: (i) Methylation rate correlation is a new genomic quantity, which contains information distinct from that contained in methylation state correlation. In particular, the rate correlation reveals mechanistic information on enzymatic processes. (ii) Some, but not all, of the rate correlation observed on nearby CpGs decays exponentially with genomic distance, and is consistent with processive activity of DNMT1 according to a linear diffusion model. Our analysis indicates an average distance of 36 bp between nearby CpGs that are methylated processively after DNA replication. (iii) In addition to evidence of processivity, we also discovered additional correlation, not consistent with a processive

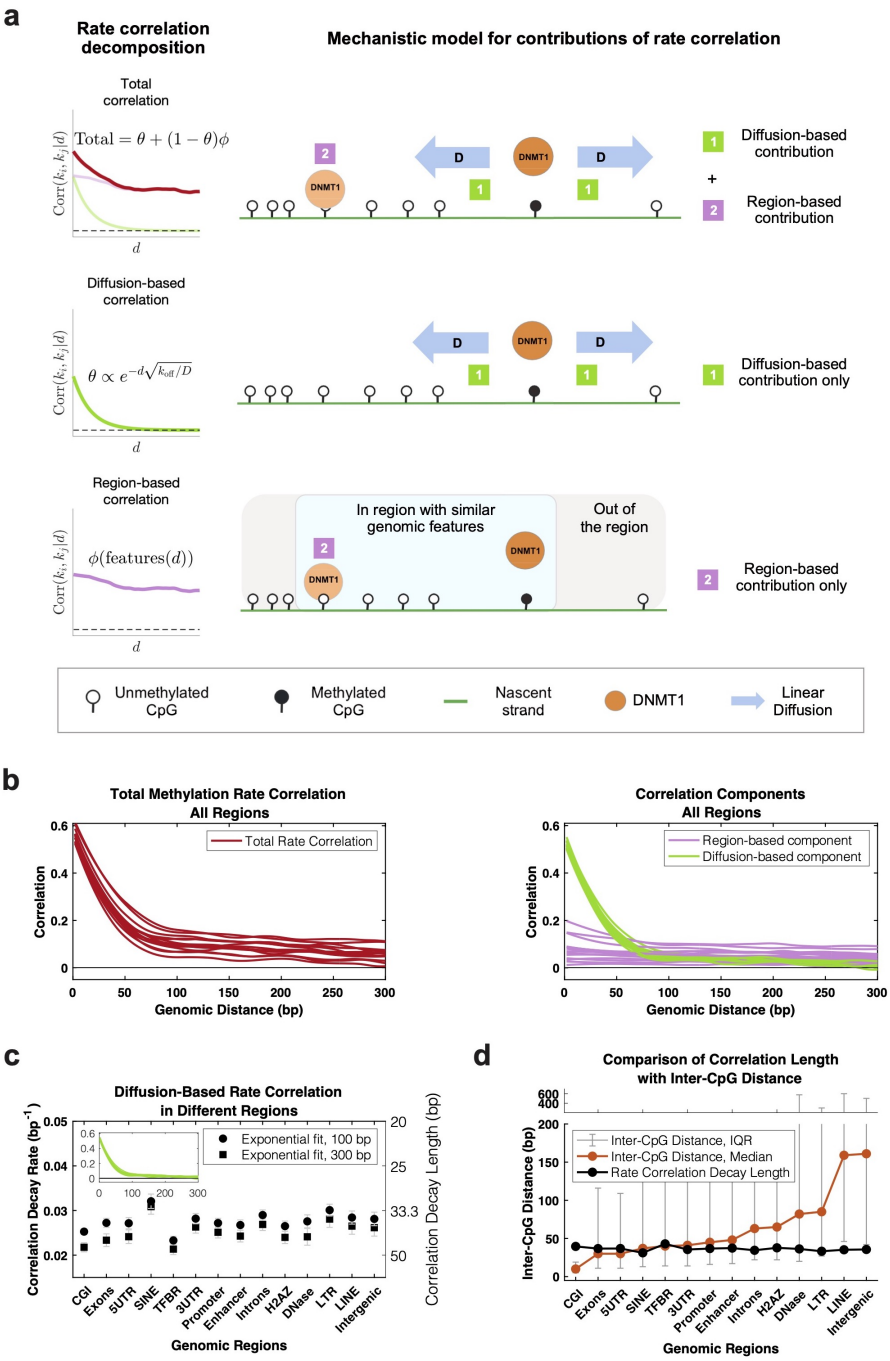

Figure 2.6: **Decomposition of experimental rate correlation into processive and nonprocessive components allows estimation of DNMT1 diffusion parameters from data.** **(a)** Mathematical model of mechanistic contributions to rate correlation arises from diffusion-based (i.e., processive) and region-based (i.e., nonprocessive) processes. **(b)** Rate correlation from experimental data are plotted in red, and the corresponding decomposed components, diffusion-based components in green and region-based components in pink. **(c)** Single-exponential fit decay constants for the green curves (θ). **(d)** Comparison of the inter-CpG distances and extracted rate correlation decay length (from θ) of each genomic region

model, that is dependent on the genomic regional context. Among studied regions, CGIs and TFBRs showed the most significant contribution of this non-processive (termed "region-based") correlation, which persisted past 1 kbp. In contrast, SINE and LINE showed the least contribution, suggesting that CpG interdependence in these regions results nearly entirely from linear diffusion (processivity) of the enzyme. (iv) Further analysis of the region-based rate correlation indicates that much of the regional variation can be attributed to variation in three genomic/epigenomic features: chromatin accessibility, CpG density, and bulk methylation.

More generally, we show how combining "top-down" mathematical modeling (i.e., data-driven, using statistical inference) with a "bottom-up" approach (i.e., using hypothesis-driven, or mechanistic models) can be used to glean kinetic insights from a measurement technique that affords genome-wide readout of the post-replication methylome over time [35]. A major new contribution of our paper is the development of a method for using data-derived rate correlations on genomic sites as a quantitative fingerprint of diffusive and nondiffusive enzyme kinetics. This method could be applied to other datasets in the future.

A number of experimental techniques have recently been developed, employing nucleoside-analog labeling of replicating DNA, followed by isolation/immunoprecipitation and sequencing [24, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50]. Our study shows that statistical correlations from such measurements have potential to yield quantitative insights into molecular mechanisms governing epigenetic inheritance, even when time resolution is coarse.

Diffusion of proteins along DNA has been an area of intense study, notably in the context of how transcription factors locate target sites (reviewed in [51, 52]). Direct measurement of linear diffusion of proteins along DNA *in vivo* has been achieved by single-molecule tracking [53, 54]. Our study shows that, for enzymes with ubiquitous target sites on DNA, correlations from temporal sequencing can also yield quantitative details on protein diffusion along DNA within cells (albeit indirectly), without recourse to labeling and microscopy.

2.3.2 Processivity of DNMT1

While a number of studies have discovered evidence of DNMT1 processivity previously in *in vitro* systems [55, 26, 27, 25, 31, 56], our study is the first to our knowledge to uncover a quantifiable signature of methyltransferase processivity in genome-wide, mammalian cell-based measurements. Thus our findings shed new light on enzymatic processivity within the context of maintenance methylation *in vivo*. Our findings are consistent with a picture wherein DNMT1 performs processive catalysis regardless of genomic context and in a quantitatively consistent manner (i.e., with relatively uniform lengths of processive steps between CpG targets, genome-wide). Our analysis does not afford direct estimation of linear coefficient D, but rather the ratio D/k_{off}. We are not aware of any existing quantitative estimates or measurements of DNMT1 linear diffusion coefficient *in vivo*. If we assume that DNMT1 has D within similar range to other measured DNA-binding proteins (of order 10^5 to 10^7 bp ^{2}/s [57, 52, 58]), our results would indicate average residence times of roughly 10^{-4} to 10^{-1} seconds for the enzyme when it is nonspecifically bound to non-CpG sites, en route to catalytic sites.

Previous *in vitro* estimates of the length of processive runs of DNMT1 varied widely. For example, Vilkaitis, et al reported processive runs as long as 520 bp [26] while Goyal et al. reported processive runs of over 6000 bp [31]. Our estimates based on the Repli-BS dataset are much shorter than these, with the average length of a processive run being about 36 basepairs for nearby CpGs. One possible explanation for the discrepancy is that our estimate is from experiments in cells, where the *in vivo* chromatin environment, replication machinery, and full complement of DNA-binding proteins are present. These could limit free diffusion of DNMT1 along DNA. Transcription factors are also thought to search for targets partly by 1D diffusion along DNA, and the effect of crowding has been considered [52]. While *in vitro* estimates for 1D sliding lengths of various DNA binding proteins are as high as 20 kb [57], *in vivo* sliding length of the lac repressor was found to be short, at

45 bp [53]. How the chromatin environment affects the motion of eukaryotic DNA binding proteins is still poorly understood [59]. In particular, the question of how methyltransferase processivity is affected by the crowded environment of the cell warrants further study.

Our analysis of the exponential contribution to rate correlation can not be directly or certainly attributed to the action of DNMT1 alone. For example, PCNA plays a role in recruitment of DNMT1 to replication foci[60], and this protein also can diffuse linearly along DNA [58] (although DNMT1 processivity does not rely on PCNA [26]). We cannot exclude the possibility that some mechanism other than intrinsic DNMT1 diffusion along DNA gives rise to the observed rate correlations, though this interpretation is consistent with the uniform lengthscale observed across different chromatin environments.

2.3.3 Nonprocessive CpG coupling

We find that additional mechanisms affect post-replication remethylation rate correlation where the local genomic/chromatin landscape allows it. We find this non-processive (i.e. non-exponential) source of rate correlation to be most prevalent where chromatin is locally open, CpGs are dense, and the background methylation is relatively low. Although our study does not attempt to define the mechanistic basis of the non-processive component of the rate correlation, various mechanisms can be proposed based on our observations and on previous literature. For example, any mechanism whereby DNMT1 reaches its CpG target through cooperative interactions with other molecules could be speculated to have kinetics dependent on the local genomic and epigenomic context. Neighbor rate correlations could then be sensitive to local context and have lengthscales determined by the the cooperative molecular interactions, rather than being solely dependent on linear inter-CpG genomic distances, in contrast with the processive mechanism. For example, recruitment of DNMT1 to replicating DNA by UHRF1 [28, 29] likely results in context-dependent kinetics, since

UHRF1 targeting is dependent on both histone state and the presence of hemi-methylated CpGs [61]. A recent finding of monoubiquinated histone H3 helping recruit DNMT1 to DNA stretches with multiple, but not one, hemimethylated CpGs [23] supports the idea that UHRF1 helps direct DNMT1 to CpG-dense regions, and is consistent with our observation of higher rate correlation in CpG-dense regions. Our findings may also be consistent with a nucleation model, in which the initial binding of DNMT1 to replicating DNA occurs on nucleosomes, directed by UHRF1, after which DNMT1 reaches nearby CpGs processively [62]. If the initial binding events of separate DNMT1s on nearby nucleosomes are correlated, such correlation would contribute on a lengthscale on the order of hundreds of basepairs, while shorter-lengthscale correlation would be introduced through processivity.

In addition to UHRF1-mediated mechanisms, additional factors are likely to play a role in the non-exponential correlation we observe. First, DNA is not one-dimensional; DNMT1 could reach nearby CpGs by facilitated diffusion (combining 1D diffusion along DNA with 3D diffusion in the nucleosol to nearby sites)[52], or by intersegmental transfer, similar to transcription factors [51]. Our processive model assumes a 1D substrate, but our results hint at sensitivity of maintenance kinetics to 3D DNA structure in the weak appearance of peaks consistent with nucleosome spacing (Figure 2.1). Finally, transient binding of post-replication DNA by transcription factors could introduce correlation into maintenance methylation kinetics, as transcription factors could transiently block access to CpGs by DNMT1 and thus delay remethylation. Such a mechanism could explain why we observe the most pronounced region-based correlation in TFBRs.

2.3.4 Relationship between WGBS and rate correlation

Our study reveals a significant difference between the local correlation of methylation state versus rate. We interpret this result as being due to the different information content of the

two: WGBS experiments capture reads that largely reflect the stable methylation landscape (though still partially influenced by replication-associated temporal variability [35]), while the Repli-BS-derived methylation rates reflect the transient dynamic processes occurring post-replication. Of note, there appears to be some relationship between the two, as we observed similar trends in region specificity in methylation state and rate (specifically in the region-specific, or non-processive, contribution to correlation). We furthermore note that, when we restrict the analysis to common sites (thus retaining CpGs with intermediate to high methylation levels, as rates are only inferrable for these sites), the similarity between rate- and state-correlation is increased (i.e., as in Figure **??** versus Figure 2.1, where, for example, the decay of bulk state-correlation in SINE is nearly as rapid as the decay of rate-correlation). This suggests partial cross-talk between the transient post-replication methylation events (including processivity) and the stable methylation landscape. Indeed, it was previously reported that CpG co-methylation decays within tens to hundreds of bp, with enzyme processivity proposed as its mechanistic origin [38].

However, our results also show that, in some regions (such as CGI), methylation state correlation is significantly longer-lived than the processive lengthscale and significantly higher than the rate correlation in total, and so the relationship between state- and rate-correlation is not directly clear. As our analysis is based only on replication-associated methylation reactions, it seems likely that our rate correlations are largely indicative of DNMT1-mediated maintence methylation, which cannot by itself shape the methylation landscape. The stochastic models of this paper focus only on sub-cell-cycle timescales and methylation maintenance, and thus assume that the methylation landscape (i.e., state correlation) is a static property dictated by the parent-strand landscape at the time of DNA replication. However methylation models of multi-cycle dynamics suggest that differences in kinetic parameters, including maintenance kinetics as well as demethylating and *de novo* methylation, shape the global methylation landscape across mitotic cycles [22, 63]. Our method for obtaining data-driven kinetic correlations could therefore be useful in the future to further improve these types of

multi-cycle mathematical models.

2.3.5 Implications for stability of the methylation landscape

A number of mathematical modeling studies have provided support for the presence of interactions (also called "collaboration") among CpGs in dynamic methylation processes, including in maintenance, de novo methylation, and demethylation reactions [30, 22, 33, 34]. These models, in which a CpG is in some way affected by the state of nearby CpGs, built upon the earlier, so-called "standard" model, wherein each CpG was considered to be independently targeted by methyltransferases [64]. Crucially, as the cited studies showed, these interactions provide the necessary nonlinearity to enable bistability in the dynamic system. That is, they enable the same family of "reader" and "writer" enzymes to simultaneously maintain distinct states of hyper- and hypo-methylation on groups of CpGs in different parts of the genome, thus mimicking observed methylation patterns. While these models tend to be phenomenological in nature (i.e., capturing dynamic phenomena without necessarily encoding detailed molecular mechanisms), processivity can be considered to be one type of molecular mechanism that contributes to inter-CpG interactions. Indeed, mathematical modeling also supports the idea that diffusive processivity enhances multi-generational stability of methylation patterns [31], just as does other mechanisms of CpG interaction [22]. It follows that interdependence of CpG methylation kinetics, as quantified by correlation in this paper, has relevance to human ageing and disease, since instability of the methylation landscape has been linked to both[65, 66].

The mathematical model of Haerter et al. predicted that local (nearest-neighbor) interactions of methylating reactions was sufficient to achieve stable propagation of methylation states over multiple generations, though longer-range interactions were required for demethylating reactions [22]. In the present study, we find that maintenance methylation occurs largely

independently in regions that are CpG sparse and show low region-dependent correlation, such as LINE. A nearest-neighbor-only model is consistent with our findings in regions such as 5UTR and SINE, where the processive lengthscale is on the same order as the typical inter-CpG distance. In CGIs, the processive lengthscale is longer than the inter-CpG distance (median 10 bp in the analyzed data), suggesting that CpGs in CGIs effectively interact beyond nearest neighbors. Strong coupling of methylation in CGIs is consistent with faster maintenance kinetics in CpG-dense regions, as has recently been reported [24]. TFBRs showed inter-CpG distances similar to the processive lengthscale, however here processivity is compensated by weak but longer range (region-based) coupling (also evident in CGIs). These findings may predict enhanced stability of methylation in TFBRs and CGIs across mitotic cycles, although better understanding of the interplay of correlation lengthscales for *de novo* and demethylating reactions is needed.

2.3.6 Limitations of our study

Our study shows that novel experimental techniques that probe replication-associated dynamics genome-wide can yield surprisingly detailed dynamic insights, despite the limited time resolution. However, the statistical inference approach is nevertheless limited. First, each site-specific inferred rate constant is the result of a fit of the data to a stochastic Poisson process; undoubtedly this is a simplistic model for dynamics that could potentially be temporally complex in reality, e.g., non-exponential or even non-monotonic. Thus, individual estimates can be error-prone for a number of reasons, from the inability of the simplistic model to capture complex dynamics, to the limited time-resolution or sampling depth for a given site, or all of the above. Because of the inherent difficulty in characterizing dynamics in detail at any individual CpG genome-wide, we focused here on general features of the rate correlation functions that are robust across a given genomic region. In this way, the whole-genome nature of the data partially compensates for the sparse temporal resolution. We

also ensured that our inference and analysis pipeline performed well on synthetic datasets, generated from simulations. Going forward, it may be possible to yield more detailed dynamic insights with deeper sampling and more fine-grained time resolution (as achieved in recent experiments [24]), which could enable further investigation into detailed, local kinetic correlation.

Our study was performed on data from a particular cell line (HUES 64). Cell-type-specific differences in human methylomes have been reported (e.g., [67]). The hESCs analyzed here were found to have bimodal CpG landscapes similar to those of somatic cell-types [68], though significant intermediate methylation is also present and has been attributed to proliferation[35]. Although our mathematical modeling suggests that the dominant short-range correlation lengthscale we observed relates to enzymatic properties, and is thus likely not sensitive to cell-type-specific differences in methylation landscapes, further studies on other cell types will be needed to determine the generalizability of our findings.

While we focus our models around the enzyme kinetics of DNMT1 (for simplicity and because it is the dominant methyltransferase responsible for carrying out maintenance activity), we acknowledge that our rate correlations are likely impacted by the presence of other (*de novo*) methyltransferases, which have been known to associate with highly methylated CpG-dense location (i.e., CGIs) [69]. Although our mechanistic models rigorously incorporate 1D diffusion, they lack the dynamic interplay of *de novo*, maintenance, and demethylation reactions that has been studied in mathematical models previously [22, 17]. Our approach could be applied to different cell lines in the future, e.g., with specific methyltransferases disrupted, to further disentangle the molecular basis of kinetic correlation.

Another limitation of the model is that it does not directly include the dynamics of DNA replication, nor account for the presence of multiple distinct origins of DNA replication in the genome. Replication timing of distinct loci could introduce kinetic correlations in DNA methylation that were not directly analyzed in our study. However, it is possible that rate

correlation reported here could be related to replication origins, as CGIs have been partly linked to replication origin activity [70] and show more pronounced region-specific correlation in our study.

2.4 Methods

2.4.1 Site-specific post-replication methylation kinetics inference

The post-replication methylation data (Repli-BS data) of Human Embryonic Stem Cells (HUES 64) was downloaded from GSE82045. In the Repli-BS experiments [35], cells were pulsed for one hour with bromodeoxyuridine (BrdU). Then, bisulfite sequencing of BrdU-labeled DNA captured CpG methylation reads from DNA that was replicated during the pulse interval. The pulse-chase experiment captured methylation level of CpGs at timepoints 0, 1, 4, and 16 hours post-pulse, thereby giving a genome-wide temporal readout of CpG-methylation over sixteen hours post-replication.

The Maximum Likelihood Estimation (MLE) procedure for inferring per-CpG post-replication methylation rates from Repli-BS data is described in detail in [37]. Briefly, the temporally distributed binary read-data (methylated-1 or unmethylated-0) at each CpG site is fitted by a Poisson process, with each site i characterized by two inferred constants, k_i and f_i, which represent the rate at which methylation accumulates at the site over the course of the experiment, and the steady-state (or long-time) fraction of cells in the measurement set that exhibit methylation at site i, respectively. We hereon refer to the inferred parameters k_i as the "post-replication remethylation rates" or simply the "remethylation rates". Details of the inference approach can be found in the Supplement (Extended Methods). Note that, while the per-site inferences are obtained based on an analytical, independent Poisson process model, the inferred rate parameters can nevertheless be used to investigate more complex

types of dynamics and inter-site dependencies through correlations that are observed among inferred parameters on nearby CpGs.

The ability to infer a remethylation rate for a given CpG site, and the uncertainty associated with that inferred rate, depends on the read-depth of the experimental data, which varies across sites and across timepoints. Details of uncertainty quantification can be found in the Supplement and in our previous study [37]. We estimated on average 30% error in any given estimate of rates k_i. We validated our method by ensuring that ground-truth rate correlations (obtained from simulated data), could be accurately recovered by the MLE inference pipeline.

2.4.2 Annotations of genomic regions

The GRCh37/hg19 genome was used as the reference genome in this paper. The region annotations for genes, promoters, exons, introns, 3'UTRs and 5'UTRs are downloaded from the UCSC Genes track in UCSC Table Browser, whereas the LINEs (Long Interspersed Nuclear Elements), SINEs (Short Interspersed Nuclear Elements), LTRs (Long Terminal Repeats) were extracted from RepeatMasker track. CGIs (CpG Islands), Enhancers were downloaded from CpG Island track and GeneHancer track, respectively. The promoters in this paper were defined as regions 2000bp upstream and 200bp downstream of transcription start sites (TSS). Local CpG density for a site was defined as the number of neighboring CpGs in a 500bp window centered at that site.

The chromatin accessibility data was retrieved from ENCODE/OpenChrom(Duke University) H1 cell line. The regions of transcription factor (TF) peaks or TFBR denoted in this paper were acquired from ENCODE ChIP-seq clusters for 161 TFs in H1 cells. The Whole Genome Bisulfite Sequencing (WGBS) dataset used in this study was retrieved from GSM1112841.

2.4.3 Region-specific stochastic simulations of post-replication maintenance methylation

In order to gain further biological insight from the experiment-derived methylation correlation functions, we perform region-specific stochastic simulations of maintenance methylation (Figure 2.4), and use these simulations to generate synthetic data analogous to the various experimental bisulfite sequencing data modalities. From these synthetic data, we compute regional correlation functions and compare to those derived from experiments. Briefly (see Methods), the simulations track nascent-strand methylation status of stretches of sequentially positioned CpGs, numbering on the order of tens of thousands. In contrast to mathematical models that treat the interplay of *de novo*, maintenance, and demethylating reactions, e.g. [22], we apply minimal models of single-site-resolution DNMT1-mediated methylation on post-replication timescales. Each stochastic simulation tracks the binary (methylated or not methylated) status of the "nascent-strand" CpGs. At the start of the simulation (representing exactly time 0 with respect to DNA replication at that site-i.e., the time of nucleotide addition), all nascent CpGs are assumed to be unmethylated. The presence or absence of methylation on cytosine bases on the opposing parental-strand at time 0 is determined probabilistically from a data-derived regional methylation landscape that acts as the simulation input. If the parental cytosine is methylated at time 0, then the CpG is considered hemimethylated and the nascent cytosine is assumed to be a target for DNMT1-catalyzed methylation, and it will acquire methylation stochastically at some post-replication timepoint, according to the chemical reaction kinetics encoded in the model. If the parental cytosine is unmethylated at time 0, then DNMT1 does not target the nascent cytosine for methylation, and the site will remain fully unmethylated. In this way, the model tracks only unidirectional maintenance methylation, and does not include active demethylation reactions. It also does not account for any de novo methylation activity. The simulation tracks post-replication timescales (following experiments, to approximately 16 hours), up to

but not including subsequent replication events.

Region-specificity is encoded at the start of the simulation in two ways: (1) the CpG positions and (2) the local methylation landscape, meaning: the probability of the nascent CpG to be a target for DNTM1-catalyzed methyl-addition due to the presence of methylation on the parental strand CpG dinucleotide. Both of these quantities are derived from experimental WGBS data with a regional filter to retain only CpGs in the desired region. Thus, in simulating a CGI region, the ith simulated CpG ($i \in [1..N]$) has a genomic position x_i and a probability f_i to be targeted for methylation. We obtain both x_i and f_i from WGBS data from hESCs, where x_i is the integer site-ID for the cytosine (which is identified as being located within a CGI) and $f_i \in [0, 1]$ is taken to be equal to the measured methylation fraction at that site. For example, if a given CpG in the dataset has a WGBS-measured methylation fraction of 0.8, then the model assigns the parental cytosine to be methylated at the start of the simulation with a probability equal to 0.8. Strands are simulated in replicate. With sufficient replicates, the simulation eventually recapitulates the experiment-derived input methylation landscape, if it is run for a period of time that is sufficiently long. That is, the simulation assumes perfect recapitulation of parent-strand methylation by DNMT1 eventually, though the time at which each hemi-methylated CpG gains nascent strand methylation is stochastic. In practice, the number of replicates and sampled time-points are chosen to match those of the experimental data (see Methods). Note that the WGBS-data-derived landscape likely reflects some degree of replication-associated temporal variability [35], rather than a true steady-state. Nevertheless, use of the WGBS-background-methylation landscape as the simulation input allows us to encode realistic region-specific differences in CpG densities and qualitative differences in bulk methylation levels.

Region-specific simulations of single-CpG stochastic enzyme-kinetic models were carried out using two candidate mechanisms, the Distributive model and the Processive model (Figure 4). The model reactions and associated rate parameters are graphically depicted in Figures

4b and 4c. The Distributive model was simulated using the Gillespie Stochastic Simulation Algorithm [71]. To incorporate 1D diffusion into the Processive model, we used a First Passage Time Kinetic Monte Carlo algorithm inspired by [72]. The methylation maintenance model and simulation method is based in part on our previous simulation studies [37]. In the present paper, we refined our Processive model and simulation algorithm to rigorously incorporate physics of 1D diffusion (also called "sliding") of proteins along DNA, with unbinding [52], while enabling simulation of large numbers of CpGs. Our analytical results on enzyme diffusion and simulation algorithm are presented in the Supplement (Extended Methods), along with further details of the models. Parameter values are chosen to be in line with experimentally measured values for DNMT1 where possible [73], and also to match features of the Repli-BS data (see Supplemental Table S1 for details).

Simulations are performed for stretches of N CpGs, ($N = 25000 - 75000$). Simulations mimic two types of experimental data modalities: WGBS and Repli-BS. To simulate WGBS experiments, for a region of CpGs, the simulation is initialized at post-replicaton time=0, and then read out at randomly sampled timepoints between zero and 24 hours later, to reflect the variable post-replication timings of bulk cells in WGBS experiments. Ten simulation replicates are combined to generate estimates of average per-site methylation levels. To mimic the Repli-BS experiment, the methylation status of CpGs in the simulation were read out at timepoints sampled from intervals matching pulse-chase experiments[35], including uncertainty with respect to true post-replication timing. That is, given the finite BrdU pulse-length, the 0-hour experimental timepoint is assumed to correspond to $t \in [0-1]$ hours post-replication, the 1-hour experimental timepoint corresponds to $t \in [1-2]$ hours post-replication, etc.. Therefore, timepoints of simulation readout were sampled from $t_{chase} + r \in [0, 1]$ hour, where experimental timepoints t_{chase} were $0, 1, 4, 16$ hours and r is a uniformly distributed random number. The number of simulations and read-outs at each timepoint were chosen to mimic the distribution of experimental read-depths. The synthetic Repli-BS data were then processed with the same MLE procedure as the experimental data, to

infer per-CpG methylation rates. Simulations were used to validate our statistical inference procedure. We tested that ground-truth correlation functions produced by the two models could be recapitulated by the inference procedure.

Chapter 3

Identifying Multicellular Spatiotemporal Organization of Cells with SpaceFlow

3.1 Introduction

The spatiotemporal pattern of gene expression is critical to unraveling key biological mechanisms from embryonic development to disease. Recent advances in spatially resolved transcriptomics (ST) technologies provide new ways to characterize the gene expression with spatial information that the popular nonspatial single-cell RNA-sequencing (scRNA-seq) method is unable to capture. The majority of current ST technologies may be categorized into *in situ* hybridization (ISH)-based and spatial barcoding-based, varying in gene throughput and resolution [74, 75, 76]. ISH-based methods can detect target transcripts at the sub-cellular resolution, such as Multiplexed Error-Robust Fluorescence ISH (MERFISH) and sequential fluorescence ISH (seqFISH), for about 100-1000 and 10,000 genes respectively

[77, 78]. Spatial barcoding-based methods can capture the whole transcriptome with varying spatial spot resolutions, such as Visium in 55 μm, the Slide-seq in 10 μm [79], and the spatiotemporal enhanced resolution omics' sequencing (Stereo-seq) in nanometer (subcellular) resolution [80].

Many methods developed for non-spatial transcriptomic data such as scRNA-seq or bulk spatial transcriptomics data [81, 82] may provide insights in designing approaches for ST data at single-cell resolution through recasting the relevant tasks in a spatial manner. For example, the identification of spatially variable genes in ST data [83, 84] can be viewed as the spatial extension of the highly variable genes in scRNA-seq data. Similarly, methods have been developed to identify spatial domains in ST data [85], the analog of cell clustering in scRNA-seq data analysis, but using spatial information to produce spatially coherent regions. Giotto [86], BayesSpace [87], and SC-MEB [88] use Markov random fields to model the related gene expression in neighboring cells. stLearn utilizes morphological information to perform spatial smoothing before clustering [89]. MULTILAYER uses graph partitioning to segment tissue domains [90]. MERINGUE performs graph-based clustering using a weighted graph that combines spatial and transcriptional similarity [91]. SpaGCN [92], SEDR [93], SCAN-IT [94], stMVC [95], and STAGATE [96] build deep auto-encoder networks to learn low-dimensional embeddings of both gene expression and spatial information, and segment domains through embedding clustering. RESEPT learns a three-dimensional embedding from ST data by a spatial retained graph autoencoder and treats the embedding as a 3D image, identifying domains through image segmentation using a convolutional neural network [2].

The domain segmentation methods reviewed above are the ST counterpart of cell clustering in scRNA-seq data analysis. Contrary to discrete clustering, another powerful analysis in scRNA-seq is the concept of continuous pseudotime which can represent developmental trajectories. The dynamics of many developing systems such as regeneration and cancer

progression are often spatially organized [2, 97]. The ST data thus provides an opportunity to simultaneously reveal both spatial and temporal structures of development. While pseudotime methods for scRNA-seq can be directly applied to ST data, the resulting trajectory may be discontinuous in space. stLearn combines non-spatial pseudotime with spatial distance by simple average, as well as filters connections between clusters inferred by scRNA-seq trajectory inference methods using a spatial distance cutoff, but the resulting connections are limited by the initial pseudotime trajectories inferred without using spatial information [89]. There is thus a demand for computational tools for integrative reconstruction of fine-resolution spatiotemporal trajectories from ST data which is continuous both in time and space.

As pseudotime trajectories are traditionally computed from a low-dimensional embedding of transcriptomic data [98], the computation of spatiotemporal trajectories can be viewed as a problem of constructing spatially-aware embeddings of ST data. Multiple strategies for computing spatially-aware embeddings may be used such as Hierarchical SNE [99], Hierarchical UMAP [100], dual embedding [101]. Additionally, deep graph neural network-based approaches, such as DeepWalk [102], Variational Graph Auto-Encoder (VGAE) [103], Graph2Gauss [104], and Deep Graph Infomax (DGI) [105], while computationally more expensive, have been utilized for ST data due to their flexibility to model and learn non-linear and complex salient spatial dependencies between genes and cells.

In this work, we develop a framework to reveal continuous temporal relationships with spatial context using ST data. By combining a DGI framework with spatial regularization designed to capture both local and global structural patterns, we extract a spatially-consistent low-dimensional embedding and construct a pseudo-Spatiotemporal Map (pSM), representing a spatially-coherent pseudotime ordering of cells that encodes biological relationships between cells, along with a region segmentation. We compare SpaceFlow with five existing methods on six ST datasets, demonstrating competitive performance on benchmarks, and use SpaceFlow

to reveal evolving cell lineage structures, spatiotemporal patterns, cell-cell communications, tumor-immune interfaces and spatial dynamics of cancer progression.

3.2 Results

3.2.1 Overview of method

SpaceFlow takes Spatial Transcriptomic (ST) data as input (Figure 3.1a) and outputs a spatially-consistent low-dimensional embedding, domain segmentation, and pseudo Spatiotemporal Map (pSM) of the tissue. The input ST data consists of an expression count matrix and spatial coordinates of cells or spots. The output embedding encodes the expression of ST data so that nearby embeddings in the latent space reflect not only the similarity in expression but also spatial proximity. The domain segmentation characterizes the spatial patterns of tissue without the need for histological or pathological knowledge. The pSM is a map that represents the pseudo-spatiotemporal relationship of cells in ST data.

Before applying the deep graph network, a Spatial Expression Graph (SEG) is constructed (Figure 3.1b) with nodes in the graph representing cells with expression profile attached, while edges model the spatial adjacency relationship of cells (Figure 3.1b). In addition, an Expression Permuted Graph (EPG) is constructed by randomly permuting the nodes in SEG and used as negative inputs for the network. To encode the SEG into low-dimensional embeddings, a graph convolutional encoder is built with Parametric ReLU (PReLU) as activation (Figure 3.1c). The graph convolutional encoder applies a weighted aggregation to the expression of a cell with its spatial neighborhood to capture local expression patterns into embeddings. We utilize a Deep Graph Infomax (DGI) framework to train the encoder [105], which optimizes a Discriminator Loss (Figure 3.1d bottom) to learn to distinguish the embeddings from SEG and EPG input. Compared to other GCN architectures, this allows

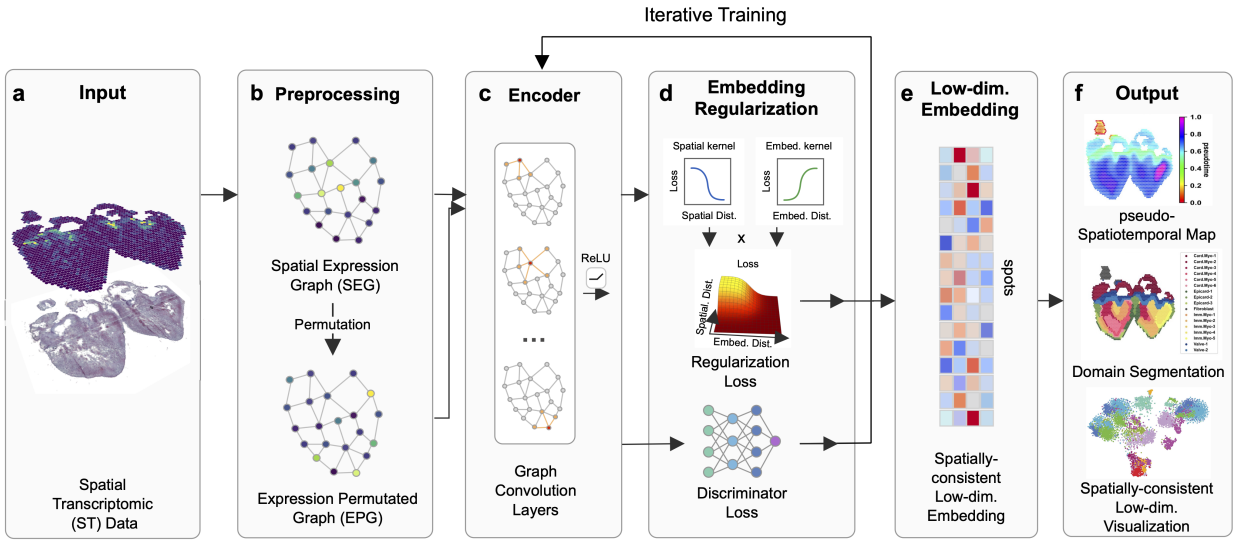

Figure 3.1: **Overview of SpaceFlow. (a)**. The input ST dataset consist of an expression count matrix and spatial coordinates of spots/cells. **(b)**. A spatial expression graph (SEG) is constructed as the network input, with edges characterizing the spatial neighborhood, and nodes representing cells/spots with expression profiles attached. By randomly permuting the nodes in SEG, Expression Permuted Graphs (EPG) are built as negative samples. **(c)**. A two-layer GCN encodes the SEG or EPG input into low-dimensional embeddings. **(d)**. The embeddings are regularized for spatial consistency. With the Spatial Regularization loss and the Discriminator loss, the encoder is iteratively trained until convergence. **(e)**. The low-dimensional embedding is obtained from the trained encoder. **(f)**. The output consists of the pseudo-Spatiotemporal Map (pSM), domain segmentation, and the visualization of low-dimensional embeddings.

the encoder to learn embeddings that emphasize specifically the spatial expression patterns that corresponding to meaningful structure as opposed to those due to non-spatial variation or noise.

Distant cells of the same cell type may exhibit a high degree of transcriptional similarity even when in very different parts of tissue. Consequently, in order to produce embeddings that most meaningfully represent spatial structure, one needs spatial consistency in embeddings, meaning that the latent space embeddings should be distant not only if their expression profile is distinct, but also when the expression is similar but their spatial locations are distant. We use embedding regularization to enforce this structure in the latent space (Figure 3.1d), which takes the spatial distance matrix and the embedding distance matrix of cells

or spots as input. These two matrices are then input into linear kernels (Spatial kernel and Embedding kernel) to calculate the loss of each cell pair based on the spatial or embedding distance. The spatial losses and embedding losses of cells from these kernels are then combined to produce the final regularization loss which is added to the discriminator loss used to train the encoder (Methods). The learned low-dimensional embeddings (Figure 3.1e) for the ST data are then used in downstream analysis, including the pseudo-Spatiotemporal Map (pSM), domain segmentation, and low-dimensional visualization (Figure 3.1f) to analyze spatiotemporal patterns of tissues.

3.2.2 Comparison of SpaceFlow with five existing methods for ST data at spot resolution

To evaluate the quality of the SpaceFlow embeddings, we compared it with five existing methods for unsupervised segmentation on ST data: one non-spatial method Seurat v4 [106], and four spatial methods Giotto [86], stLearn [89], MERINGUE [91], and BayesSpace [87] on a 10x Visium human Dorso-Lateral Pre-Frontal Cortex (DLPFC) dataset consisting of twelve samples [107]. Spots are annotated as one of six layers (layer 1 through layer 6) or white matter, and these annotations are used as the ground truth for benchmarking.

To compare the domain segmentation performance quantitatively, we used the adjusted Rand Index (ARI) to measure the similarity between the inferred domains and the expert annotations across all twelve sections (Figure 3.2a). SpaceFlow shows a 0.427 median ARI score, the second-highest across the six methods, slightly lower than the BayesSpace, which has 0.438 median ARI. MERINGUE shows the lowest median ARI score (0.232), followed by Seurat (0.300) and then Giotto (0.332), and stLearn (0.369). Interestingly, the DGI method without the spatial regularization used in SpaceFlow shows a significant decrease in ARI, with a 0.332 median score, indicating that spatial regularization does effectively improve the

domain segmentation of DGI.

Next, we performed a more detailed analysis on section 151671 (Figure 3.2b-f). We first computed the domain segmentation for each method and visualized the output compared to the expert annotation (Figure 3.2b). It is seen that all methods fail to capture the subtle structure of Layer 4 (L4), suggesting that this ST data does not have the necessary spatial resolution to capture the L4 structure. Both SpaceFlow and BayesSpace can capture all the remaining structures (L3, L5, L6, and WM) observed in the annotation. Moreover, BayesSpace identified the outer ring of the WM as an additional structure, whereas SpaceFlow found a different structure at the top right part of Layer 3 (labeled as domain 3 in orange). The structure found in SpaceFlow is consistent with the domains from Giotto, stLearn, and MERINGUE. stLearn also identified L5, L6, and WM that are consistent with the annotation but with noisy boundaries between domains. Giotto and MERINGUE identified the L6 and WM domain but are unable to identify the boundary of L5 along with the same noisy boundary issues. Seurat showed an overall disordered domain structure and can barely capture the white matter (WM) structures. The DGI method (Figure 3.2b), like SpaceFlow but without spatial regularization, showed a layered structure with inconsistent boundary shape and non-contiguous domains, reinforcing the importance of spatial regularization. Similar results can also be observed in samples 151507 and 151673 (Supplementary Fig. 1a, c).

We next compared the low-dimensional embeddings from Seurat, stLearn, DGI, and SpaceFlow (the segmentation methods Giotto, MERINGUE, and BayesSpace do not produce embeddings), applying UMAP to the embeddings to produce a two-dimensional visualization of all spots colored by the layer annotation. We observe that SpaceFlow embeddings produce embeddings that clearly separate the spots by layer when compared to stLearn, DGI, and Seurat (Figure 3.2c). As the separation between low-dimensional embeddings of the regions provides an upper limit on the ability of segmentation to separate the regions, this shows that the incorporation of spatial regularization produces more distinct embeddings between

different layers and thereby a greater ability to distinguish them in downstream analysis.

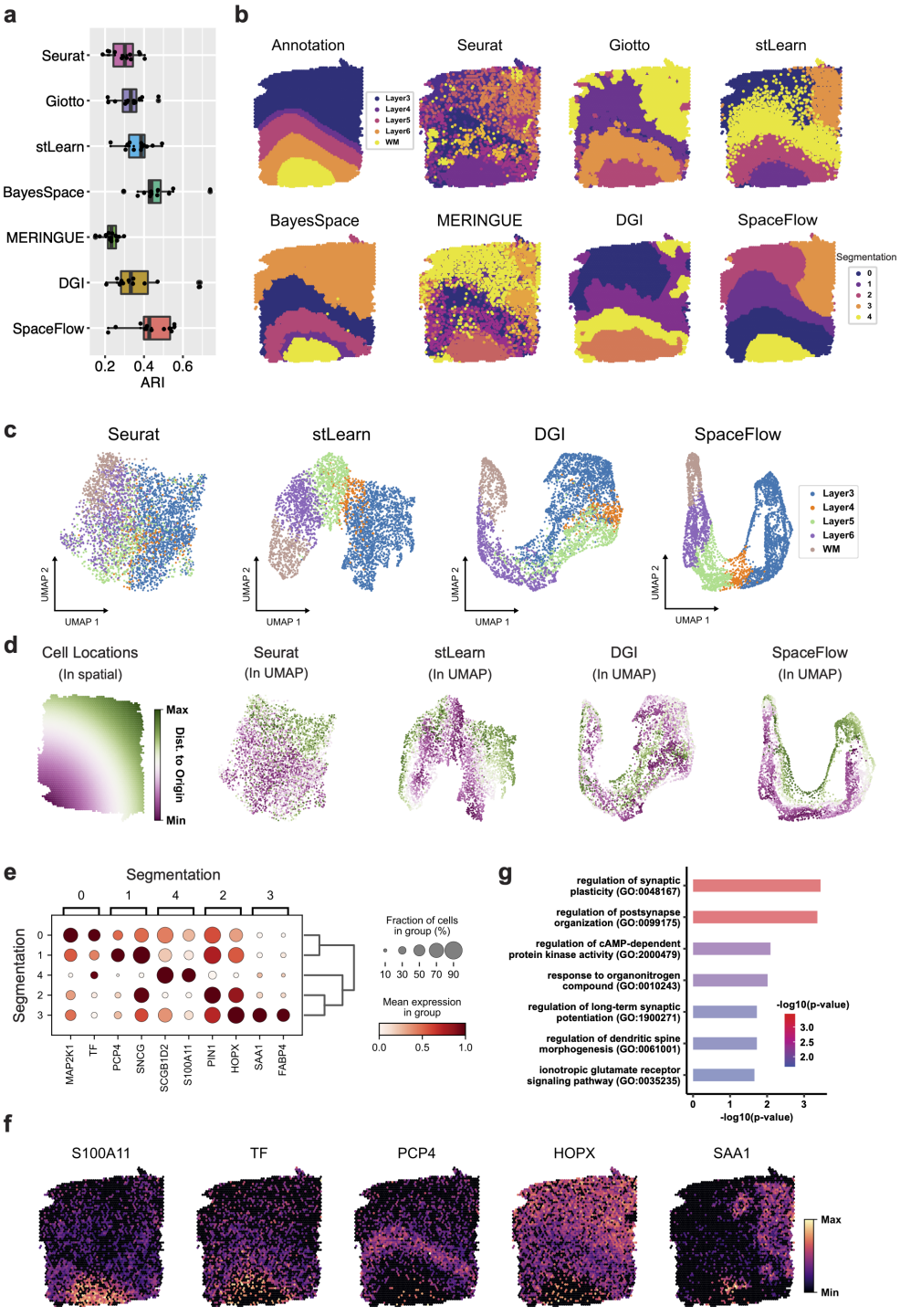

Figure 3.2: **Comparison with five unsupervised methods shows that SpaceFlow can identify biological meaningful spatial domains and generate spatially-consistent low-dimensional embeddings. (a)**. Boxplot of clustering accuracy in all sections of the LIBD human dorsolateral prefrontal cortex (DLPFC) ST dataset [107] (n=12 sections) in terms of adjusted rand index (ARI) scores for seven methods. In the boxplot, the center line, box limits and whiskers denote the median, upper and lower quartiles, and 1.5× interquartile range, respectively. **(b)**. Domain segmentations of cortical layers and white matter by annotation (top left panel) and by seven different methods (other panels) using section 151671 of DLPFC data. **(c)**. UMAP visualizations for DLPFC data section 151671, using low-dimensional embeddings from Seurat, stLearn, DGI, and SpaceFlow colored by the layer annotation of spots. **(d)**. Cell spatial locations (left panel) and UMAP visualizations (right four panels) colored by the Euclidean distance between spot/cell and origin (0,0), which is the left bottom corner of the first panel. **(e)**. Dot plot of the gene expression of domain-specific markers. The dot size represents the fraction of cells in a domain expressing the marker and the color intensity represents the average expression of the marker in that domain. **(f)**. Spatial expression for the top-1 markers of the identified domains. **(g)**. The Gene Ontology (GO) enrichment analysis of the domain-specific genes (161 genes) for the domain 3 in panel b, SpaceFlow. Both the color and the length of bars represent the enrichment of GO terms using -log10(p-value) metric from topGO analysis. P-values were obtained using the one-sided Fisher's exact test without multiple-testing correction. P values ¡ 0.001 were considered significant.

To study how the low-dimensional embeddings from different methods encode spatial information, we show the same UMAP embeddings colored by the spatial distances between the spot and the origin (Figure 3.2d), so that embeddings which preserve the spatial structure will maintain this color gradient. Seurat embeddings exhibit a high level of mixing in this color, indicating significant deformities in both local and global structure, whereas DGI shows local color gradients with a minor color mixture and shows no global gradient structure. In contrast, SpaceFlow and stLearn both exhibit a clear global gradient structure with a clear coloring trend in each annotation layer. This indicates the learned embeddings from SpaceFlow and stLearn encode transcriptional information while also preserving the local and global spatial structure of the data.

To check whether the identified domains from SpaceFlow are biologically meaningful, we performed a domain-specific expression analysis. We found spatial specific expression patterns for the identified domain-specific genes. For instance, the top domain-specific gene for

domain 3 (orange) in SpaceFlow is SAA1. Within domain 3, it is expressed in 90% of cells with a mean expression of 0.96 (scaled from 0 to 1), whereas outside this domain it is expressed in less than 30% of cells with a mean expression of approximately 0.13 (Figure 3.2e). This spatial expression specificity is also clear in the spatial expression heatmap showing top-1 marker genes for each domain (Figure 3.2f). Among these genes, we found that PCP4 was previously reported as the marker for layer 6 in prefrontal cortex [108]. The other genes have clear layer correlations although not previously reported, suggesting new experiments are needed for validating the potential new marker genes. We carried out a Gene Ontology (GO) analysis for the domain-specific genes whose p-value is less than 0.01 in domain 3 (Figure 3.2g). We observed GO terms associated with regulation of cAMP-dependent protein kinase activity, ionotropic glutamate receptor signaling pathway, regulation of long-term synaptic potentiation regulation of synaptic plasticity, etc. This suggests that the spatially specific expression in the identified domain 3 may be related to long-term synaptic activity, which is consistent with the observation that several top domain-3 specific expression patterns such as MALAT1 (p-value$< 5.5 \times 10^{-33}$) [109], CAMK2A (p-value$< 7.4 \times 10^{-23}$) [110], PPP3CA (p-value$< 4 \times 10^{-16}$) [111] are involved in long-term synaptic potentiation. The fact that this gene expression is clearly related to neural activity indicates a meaningful subdivision of Layer 3 despite a lack of annotations for this region. This suggests that the expert annotations, even if accurately describing the layer structure, may not paint a complete picture of the spatial structure within the data.

3.2.3 SpaceFlow uncovers pseudo-spatiotemporal relationships among cells

Next, we study the pseudo-Spatiotemporal Map (pSM) computed by SpaceFlow. Different from traditional pseudotime as used in scRNA-seq analysis, which only considers the similarity in expression between cells, the pSM considers both spatial and transcriptional relationships among cells simultaneously (Methods). In spatial visualizations of the pseu-

dotimes produced from Seurat, Monocle, traditional single-cell pseudotime methods that do not incorporate spatial information, we observed a lack of layered patterns as well as significant noise (Figure 3.3a). In contrast, both spatially-aware methods tested, stLearn and SpaceFlow present a layer-patterned pSM with a clear and smooth color gradient (Figure 3.3a), suggesting a pseudo-spatiotemporal ordering from White Matter (WM) to Layer 3. This ordering mirrors the correct inside-out developmental sequence of cortical layers and reflects the layered spatial organization of the tissue. However, stLearn shows less consistency with the annotation in the White Matter (WM) region when compared to SpaceFlow. Similar patterns can also be observed in samples 151507 and 151673 (Supplementary Fig. 1b, d). We also run SpaceFlow on more 10x Visium ST datasets, the results can be found in Supplementary Information (Supplementary Figure 3.3).

To test the capability of SpaceFlow on single-cell resolution ST data with a large number of cells, we evaluated SpaceFlow on a Stereo-seq dataset from mouse olfactory bulb tissue, capturing 28243 genes across 18197 cells [80], comparing with traditional pseudotime methods Seurat, Monocle, and Slingshot. We observed that Seurat shows little variation in pseudotime across the tissue except for the outer rings, which show slightly higher pseudotime values than in the inside. Monocle is much noisier than Seurat and shows no clear patterns. Slingshot is similar to Seurat and exhibits outer-ring patterns. By contrast, SpaceFlow presents a clear layered pattern mirroring the annotated layers of the olfactory bulb tissue (Figure 3.3b). The pSM value (red) is lowest in the external plexiform layer (EPL) and then increases when moving away in both directions. This ordering in the pSM is consistent with the developmental sequence of these layers, where starting from the central EPL, development proceeds bilaterally outwards, leading to the mitral cell layer (MCL) and glomerular layer (GL), olfactory nerve layer (ONL), and the granule cell layer (GCL) develops last [112]. The highest values are observed on the inner side, with the peak in the granule cell layer (GCL) and the rostral migratory stream (RMS). This shows that the pSM computed by SpaceFlow is not only more clearly spatially organized than non-spatial pseudotime, but that

these results also more accurately represent the temporal and developmental relationships between cells.

We next identify marker genes from the pSM. By calculating the top genes by correlation with the pSM values, we found genes that are predominantly expressed in layers of the olfactory bulb tissue (Figure 3.3c). One of the top marker genes, NRGN, shows clear expression patterns localized in the granule cell layer (GCL), and previous experiments have shown that NRGN is usually expressed in granule-like structures in pyramidal cells of the hippocampus and cortex [113]. This shows how the pSM can be used to facilitate biomarker identification for tissues.

Next, we compared the domain segmentation performance of SpaceFlow against Seurat, running without incorporating spatial data, as well as the spatial methods MERINGUE and DGI on the Stereo-seq data. We show a global and a zoomed view of the identified domains for each method (Figure 3.3d). The Seurat segmentation is characterized by two large regions – all inner layers except the olfactory nerve layer (ONL) are mainly combined into one region (domain 0 in dark blue) and the ONL is segmented as another (domain 1 in orange). However, even in the inner layers, there are many cells classified as domain 1 (orange), lacking a clear separation between domains. To control for the effect of the resolution parameter, we considered values of 0.3 to 0.8, 1.0 and 2.0, resulting in 13, 15 and 30 clusters respectively (Supplementary Fig. 2a). However, the spatial consistency of clusters does not improve with a higher resolution parameter. MERINGUE identified three major layers, with one additional compared to Seurat, which corresponds to the external plexiform layer (EPL) in the annotation in Figure 3.3b; however, there is significant spatial noise and it is difficult to see boundaries between tissue layers even in the zoomed view. With the DGI method, we observed a layered structure of domains, but significant mixing of domains is still visible in the zoomed view. In SpaceFlow, the eight-layer structure is much clearer, as nearly no mixture between occurs the corresponding ONL (domain 2 in green) and GL (domain 5 in silver gray) regions, whereas there are clear mixtures of labels by the

other methods across all regions.

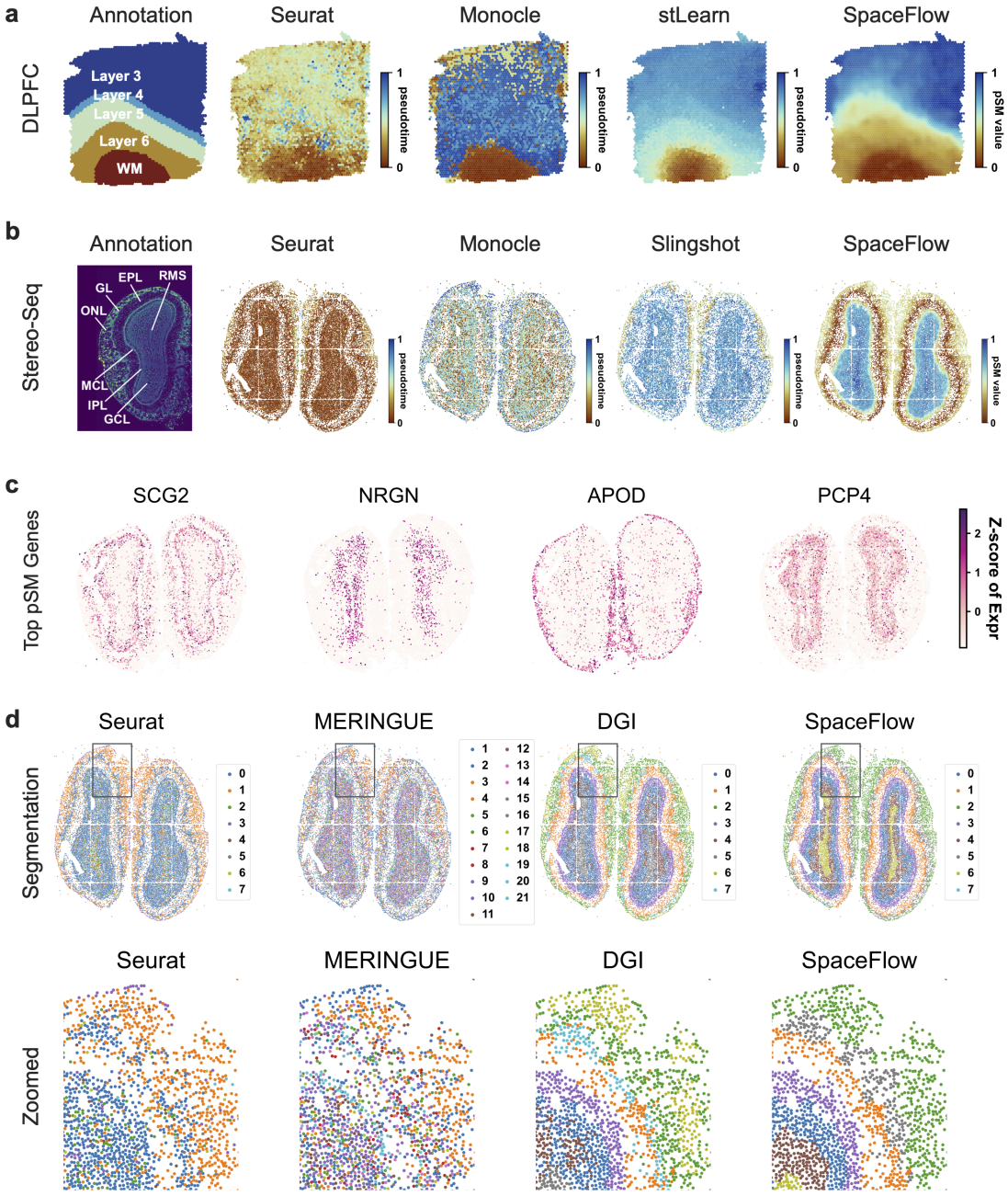

Figure 3.3: **SpaceFlow generates pseudo-Spatiotemporal Map for ST data and uncovered pseudo-spatiotemporal relationship between cells in both spots-resolution and single-cell resolution ST data. (a)**. Spatial visualization of pseudotime calculated by Seurat, Monocle, stLearn, and the pSM generated by SpaceFlow on DLPFC data (same dataset as in Figure 3.2, spot resolution) and **(b)**. Spatial visualization of pseudotime calculated by Seurat, Monocle, Slingshot, and the pSM generated by SpaceFlow on Stereo-seq data (single-cell resolution). **(c)**. Spatial expression of genes exhibiting the highest correlation between expression and the pSM value of corresponding spots/cells. The color represents the z-score of expression level. **(d)**. Domain segmentations of Stereo-seq ST data given by Seurat, MERINGUE, DGI, and SpaceFlow. Top row: full views of the domain segmentations from different methods, bottom row: zoomed views of regions boxed in each panel.

3.2.4 SpaceFlow reveals evolving cell lineage structures in chicken heart development ST data

To study how the pSM may be used to uncover spatial expression dynamics in embryonic development, we retrieved and utilized an ST dataset on the chicken heart [2] at four key Hamburger-Hamilton ventricular developmental stages. The dataset contains 12 tissue sections in total, and is sequenced at day 4 (5 sections), day 7 (4 sections), day 10 (2 sections) and day 14 (1 section). To build a baseline for comparison, we first visualized spot annotation from the original study (Figure 3.4a). Then, we computed the domain segmentation from SpaceFlow for each time point (Figure 3.4b) and labeled the domains based on their top marker genes as compared with the literature (Details in Supplementary Data 1).

We first found an evolving lineage structure, annotated as Valve in Figure 3.4b. This newly identified structure is evident from Day 7 (D7) with a layered structure and consistent shape during heart development. The identified structure is consistent with the anatomical regions of the chicken heart at the sequenced stages [2, 114]. We also characterized the transition dynamics of the myocardium from the immature to the mature state across the period from D4 to D14 (immature myocardium annotated by orange/yellow change into cardiomyocytes annotated in red/pink). Moreover, we identified that the epicardium structure (annotated

in green) on the outer ring of immature myocardium transformed into cardiomyocytes from D7 to D14.

To better understand the spatiotemporal organization of the chicken heart during development, we computed the pSM for each time point separately (Figure 3.4c), considering that pseudo-time across tissues with different time points may not be comparable. Similar to the domain segmentation, the identified valve structures are clear from D7 to D14 in the pSM. In addition, the myocardium in the ventricles is more homogeneous in the pSM (blue) than in the domain segmentation. This suggests the difference in the myocardium of ventricles might be much more subtle than regions showing different pSM values. We also found the annotated myocardium in ventricles to consistently show higher pSM values (blue) than other regions, which indicates the pseudo-spatiotemporal ordering of the myocardium in the ventricles is later than other regions in the same stage. By contrast, the identified valve structures show yellow color in the pSM from D7 to D14, suggesting the ordering is relatively late compared with the regions colored in red or orange. By plotting pSM values of spots against the first component of the UMAP embedding (Figure 3.4d), similar patterns can be observed, where the cells with valve annotations colored in blue shows intermediate pSM values (y-axis) and lies in the middle of the trajectories in Figure 3.4d. These spatiotemporal patterns revealed in the pSM are consistent with previous observations in chicken cardiac development [114]. Through a hierarchical clustering for domains across all four stages based on the expression of top domain-specific marker genes, we found expression programs specific to evolving structures (Figure 3.4e). We observed the valves of D7 and D10 to be similar to each other in expression, with genes that regulate cell growth and proliferation such as S100A11, S100A6, and CNMD, as well as genes associated with cell-collagen interaction such as TGFBI, found as the top marker genes for these populations. We also performed GO analysis to study the function of identified valve structures (Figure 3.4f) and found enrichment of GO terms associated with negative regulation of BMP signaling pathway and negative regulation of epithelial-mesenchymal transformation (EMT). Previous studies found that EMT mediated

by BMP2 is required for signaling from the myocardium to the underlying endothelium to form endocardial cushion (EC), which ultimately gives rise to the mature heart valves and septa [115]. We also observed enrichment of positive regulation of canonical Wnt signaling pathway, previously shown as a regulator of endocardial cushion maturation as well as valve leaflet stratification, homeostasis, and pathogenesis [116].

To investigate cell-cell communication between the identified valve structures and other tissue regions, we performed space-constrained CellChat analysis [117] using the domain labels from SpaceFlow as groupings. The top two identified pathways for the valve structures are midkine (MK) and pleiotrophin (PTN), which belong to the subfamily of heparin-binding growth factors. We observed strong signaling in MDK-SDC2, MDK-NCL, PTN-SDC2, and PTN-NCL ligand-receptor pairs from valve tissue to nearby immature cardiomyocytes and atrium cardiomyocytes (Figure 3.4g). These interactions have various functions, such as angiogenesis, oncogenesis, stem cell self-renewal, and play important roles in the regeneration of tissues, such as the myocardium, cartilage, neuron, muscle, and bone52. Studies have shown that midkine impedes the calcification of aortic valve interstitial cells through cell-cell communications [118]. In addition, SDC2 is found required for migration of the bilateral heart fields towards the mid-line in zebrafish model [119]. Pleiotrophin (PTN) is usually considered a cytokine and growth factor that promotes angiogenesis [120]. Together, the observed cell-cell communication based on the structures identified by SpaceFlow suggests anti-calcification and pro-angiogenesis processes are important during the maturation of valve tissue.

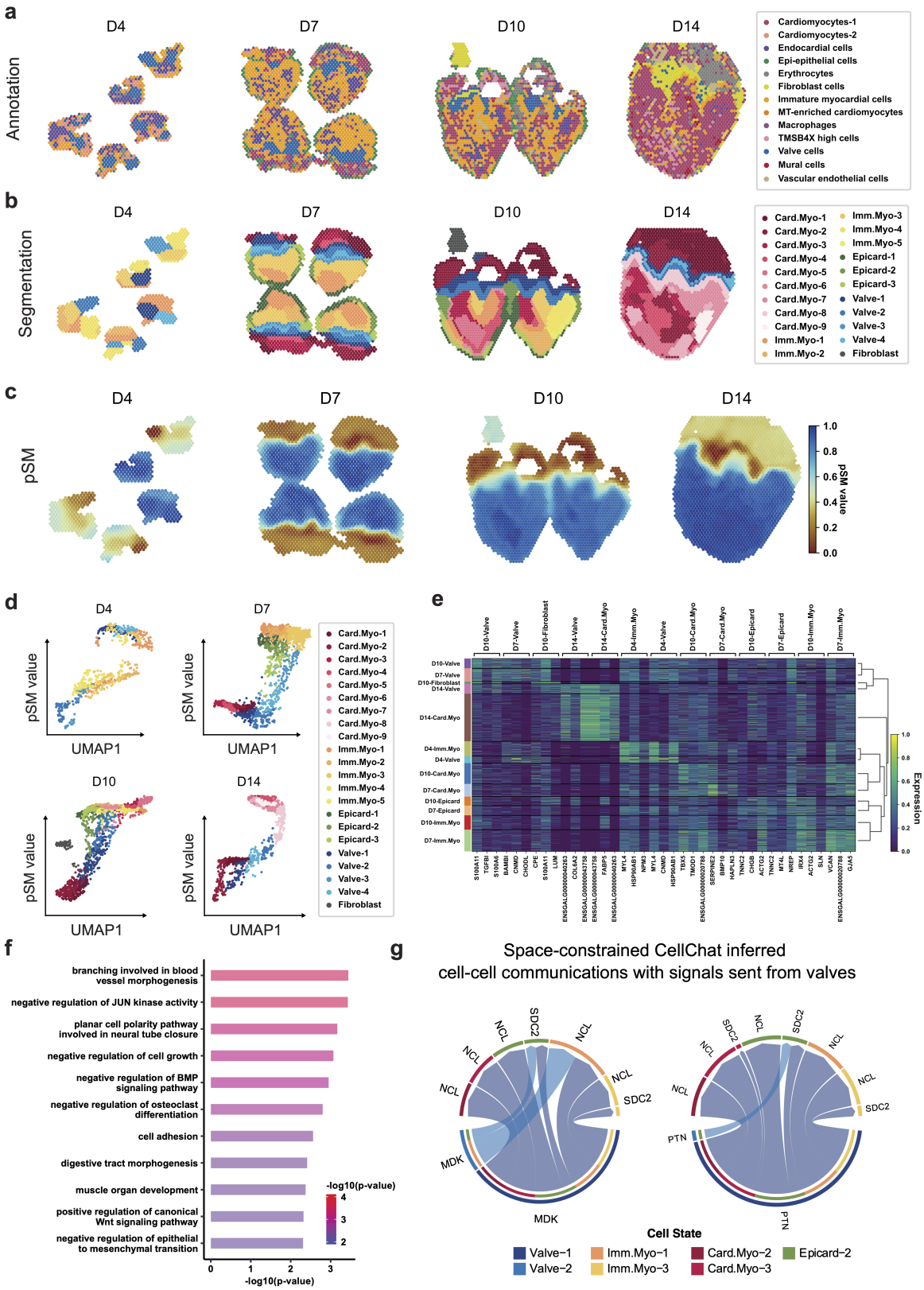
a
Annotation
D4
D7
D10
D14
Cardiomyocytes-1
Cardiomyocytes-2
Endocardial cells
Epi-epithelial cells
Erythrocytes
Fibroblast cells
Immature myocardial cells
MT-enriched cardiomyocytes
Macrophages
TMSB4X high cells
Valve cells
Mural cells
Vascular endothelial cells
b
Segmentation
Card.Myo-1
Card.Myo-2
Card.Myo-3
Card.Myo-4
Card.Myo-5
Card.Myo-6
Card.Myo-7
Card.Myo-8
Card.Myo-9
Imm.Myo-1
Imm.Myo-2
Imm.Myo-3
Imm.Myo-4
Imm.Myo-5
Epicard-1
Epicard-2
Epicard-3
Valve-1
Valve-2
Valve-3
Valve-4
Fibroblast
c
pSM
pSM value
d
pSM value
UMAP1
e
Expression
f
branching involved in blood vessel morphogenesis
negative regulation of JUN kinase activity
planar cell polarity pathway involved in neural tube closure
negative regulation of cell growth
negative regulation of BMP signaling pathway
negative regulation of osteoclast differentiation
cell adhesion
digestive tract morphogenesis
muscle organ development
positive regulation of canonical Wnt signaling pathway
negative regulation of epithelial to mesenchymal transition
-log10(p-value)
g
Space-constrained CellChat inferred cell-cell communications with signals sent from valves
NCL
SDC2
MDK
PTN
Cell State
Valve−1
Valve−2
Imm.Myo−1
Imm.Myo−3
Card.Myo−2
Card.Myo−3
Epicard−2

Figure 3.4: **SpaceFlow reveals evolving cell lineage structures in chicken heart development ST data. (a)**. Annotation of ST spots from the original study [2], where cell types are predicted by mapping scRNA-seq data to ST data. **(b)**. Annotations from SpaceFlow, with identified valve structures colored in blue. **(c)**. The pSM generated by SpaceFlow. **(d)**. The pSM value versus UMAP component 1 from low-dimensional embeddings colored by annotations from SpaceFlow. **(e)**. Hierarchically clustered heatmap of top-3 domain-specific genes for spots in all time points. **(f)**. Gene Ontology (GO) enrichment for the top domain-specific genes (32 genes) in the identified valve structures. Both the color and the length of bars represent the enrichment of GO terms using -log10(p-value) metric from topGO analysis. P-values were obtained using the one-sided Fisher's exact test without multiple-testing correction. P-values < 0.005 were considered significant. **(g)**. Space-constrained CellChat inferred cell-cell communication of MDK (left) and PTN (right) pathways with signals sent from spots in the valve regions.

3.2.5 SpaceFlow identifies tumor-immune microenvironment in human breast cancer ST data

To study the cancer microenvironment interaction and tumor progression, we applied SpaceFlow to human breast cancer ST data [121]. We show here results for sample G, consistent with the original paper. Results for other samples can be found in the Supplementary Information (Supplementary Fig. 6). First, we performed domain segmentation and compared it to the expert annotation (Figure 3.5b). The obtained domains were labeled based on their marker genes (Details in Supplementary Data 2). The regions in-situ cancer-1/2, APC, B,T-1/2, and invasive-1/2 identified in the SpaceFlow segmentation agreed with the annotations Immune rich, Immune:B/plasma, and Cancer 1 respectively from the original study. However, we also identified three tumor-immune interface regions, labeled as Tu.Imm.Itfc-1/2/3, which were labeled as mixtures of other cell types in the original study.

To reveal the pseudo-spatiotemporal relationship between spots in tissues, we generated the pSM and compared it with the spatially visualized pseudotime calculated by two alternatives: Monocle, which does not use spatial information, and applying DPT to spatially aware embeddings from stLearn. In the Monocle pseudotime, we observed regional patterns consistent with the Cancer: immune rich and Cancer 1 annotations from the original study (Figure

3.5c). However, the spatial noise in the Monocle pseudotime makes visualizing the overall structure of cancer development difficult. In the pseudotime from stLearn, we can only observe two major types of regions, the Cancer: immune rich regions with larger pseudotime values, and other regions that are more homogeneous but noisy in pseudotime (Figure 3.5d). In the SpaceFlow pSM, we see a much clearer representation of the spatiotemporal structure of the cancer cells and the patterns is highly consistent with the expert annotation in Cancer: Immune rich, Cancer 1, and Immune: B/plasma regions (Figure 3.5e). The in-situ cancer-1 regions show the lowest pSM values, whereas the Invasive-1 and Invasive-2 regions present the highest pSM values, which indicates the in-situ cancer-1 developmentally preceded than invasive regions. This trajectory can be seen clearly when we plot pSM values against the UMAP component 1 of the embeddings (Figure 3.5f). A smooth progression is shown starting from the left bottom corner with the in-situ cancer-1 and branching into APC,B,T-1/2, and in-situ-cancer-2, which then merge into tumor-immune interface populations and end in the invasive-1/2 population. This suggests that in-situ-cancer-2 may be metastasized from in-situ cancer-1.

To study the characteristics of the tumor microenvironment, we identified marker genes for each domain (Figure 3.5g). We found invasive-1, invasive-2, and Invasive-Connective (Inva-Conn) share strong expressions of the genes MMP11 and MMP14. Matrix metalloproteinase (MMP) family genes are involved in the breakdown of the extracellular matrix in processes such as metastasis [122]. In the in-situ cancer-1 population, we observed region-specific Interferon-induced expressions, such as IFI27, IFI6, which are associated with cancer growth inhibition and apoptosis promotion [123]. The in-situ cancer-2 population shows a strong and specific expression of TMEM59 and SOX4, which both can promote apoptosis. In tumor-immune interfaces, we found both pro-tumor and anti-tumor gene expressions. For instance, In tumor-immune interface-3, pro-tumor expression markers are TIMP1, a member of MMPs involved in the degradation of the extracellular matrix, whereas IGFBP4, PFDN5, CD63 repress tumor progression [124, 125]. We visualized these dual activities of pro-tumor

and anti-tumor expression and annotated with pro-tumor or anti-tumor labels to confirm our observations (Figure 3.5h). These dual activities are also confirmed in GO analysis (Figure 3.5i). The enrichment of the marker genes of tumor-immune interface-1 show pro-tumor GO terms such as: negative regulation of intrinsic apoptotic pathway in response to DNA damage by p53 class mediator, negative regulation of plasmacytoid dendritic cell cytokine production (reduce type I interferon production). Anti-tumor enrichment is also found, such as positive regulation of T cell mediated cytotoxicity (promotes the killing of cancer cells), antigen processing and presentation via MHC class I B (enhances antigen presentation).

To study the cell-cell communication between the invasive (or in-situ cancer) regions and the tumor-immune interfaces, we inferred cell-cell communication through Space-constrained CellChat analysis [117]. We found strong cell-cell communication between the invasive tissue region and the nearby tumor microenvironment through the collagen pathway, which facilities EMT transition and multiple processes associated with cancer progression and metastasis. Similar cell-cell communication is observed in in-situ cancer, where MDK-SDC1 and APP-CD74 signaling are observed to promote the progression and metastasis (Figure 3.5j-k).

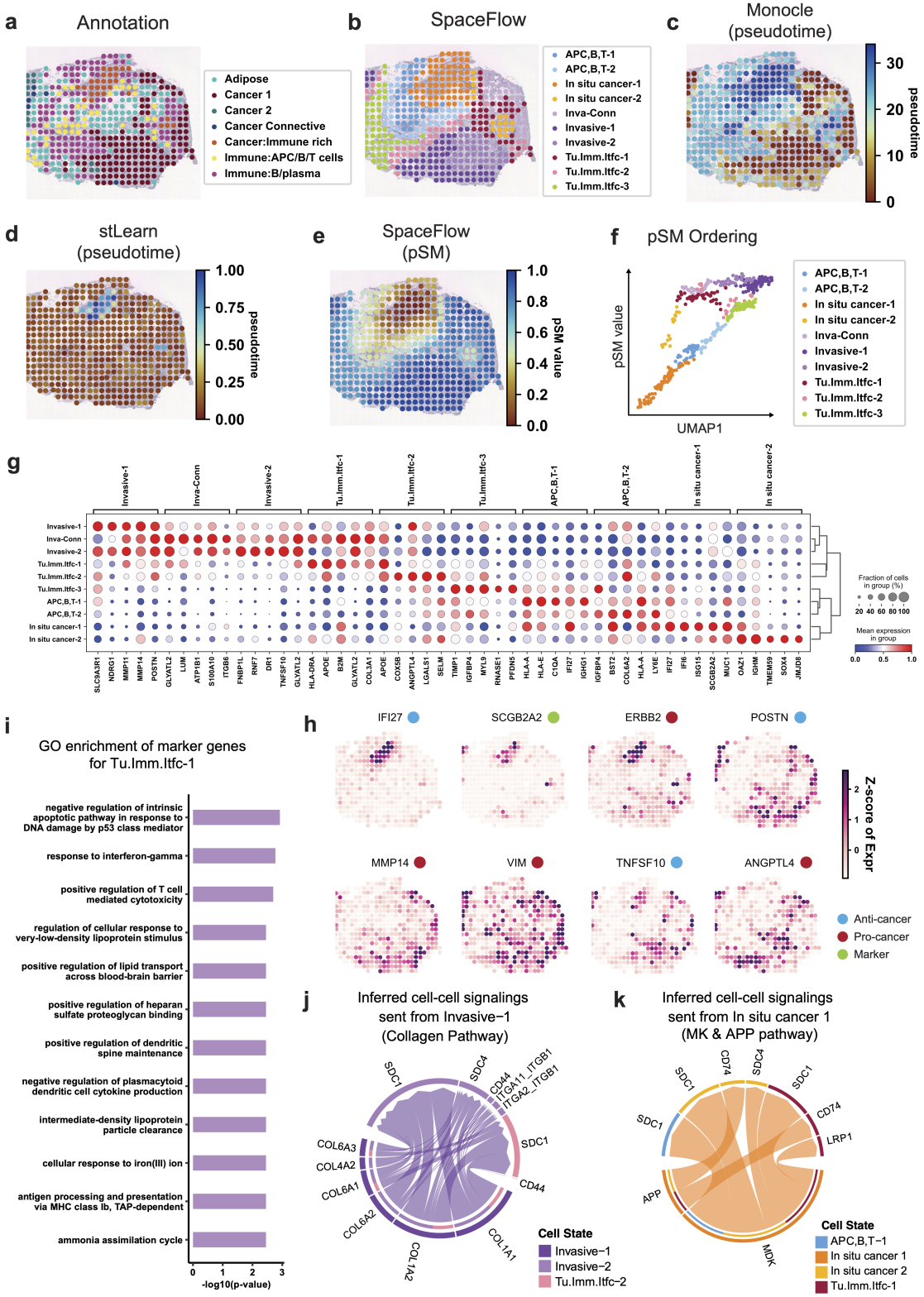
a
Annotation
Adipose
Cancer 1
Cancer 2
Cancer Connective
Cancer:Immune rich
Immune:APC/B/T cells
Immune:B/plasma
b
SpaceFlow
APC,B,T-1
APC,B,T-2
In situ cancer-1
In situ cancer-2
Inva-Conn
Invasive-1
Invasive-2
Tu.Imm.Itfc-1
Tu.Imm.Itfc-2
Tu.Imm.Itfc-3
c
Monocle
(pseudotime)
pseudotime
d
stLearn
(pseudotime)
e
SpaceFlow
(pSM)
pSM value
f
pSM Ordering
pSM value
UMAP1
g
h
IFI27
SCGB2A2
ERBB2
POSTN
MMP14
VIM
TNFSF10
ANGPTL4
Z-score of Expr
Anti-cancer
Pro-cancer
Marker
i
GO enrichment of marker genes
for Tu.Imm.Itfc-1
negative regulation of intrinsic apoptotic pathway in response to DNA damage by p53 class mediator
response to interferon-gamma
positive regulation of T cell mediated cytotoxicity
regulation of cellular response to very-low-density lipoprotein stimulus
positive regulation of lipid transport across blood-brain barrier
positive regulation of heparan sulfate proteoglycan binding
positive regulation of dendritic spine maintenance
negative regulation of plasmacytoid dendritic cell cytokine production
intermediate-density lipoprotein particle clearance
cellular response to iron(III) ion
antigen processing and presentation via MHC class Ib, TAP-dependent
ammonia assimilation cycle
-log10(p-value)
j
Inferred cell-cell signalings
sent from Invasive−1
(Collagen Pathway)
Cell State
Invasive−1
Invasive−2
Tu.Imm.Itfc−2
k
Inferred cell-cell signalings
sent from In situ cancer 1
(MK & APP pathway)
Cell State
APC,B,T−1
In situ cancer 1
In situ cancer 2
Tu.Imm.Itfc-1

Figure 3.5: **SpaceFlow identifies tumor-immune cell-cell communication in human breast cancer ST data. (a)**. H&E image and annotation from the original study for the spots of sample G in human breast cancer ST data [121]. **(b)**. Domain segmentation from SpaceFlow **(c)**. Spatial visualization of pseudotime calculated by Monocle. **(d)**. Spatial visualization of pseudotime calculated by stLearn **(e)**. The pSM from SpaceFlow. **(f)**. The pSM versus UMAP component 1 from low-dimensional embeddings colored by annotations from SpaceFlow. **(g)**. Dot plot of the gene expression of domain-specific markers. The dot size represents the fraction of cells in a domain expressing the marker and the color intensity represents the average expression of the marker in that domain.**(h)**. Spatial expression of top domain marker genes with anti-cancer, pro-cancer, or dual function labels. **(i)**. Gene Ontology (GO) enrichment for the top domain-specific genes (48 genes) in the identified Tumor-Immune-Interface-1 (Tu.Imm.Itfc.1). Enriched GO terms are presented as -log10(p-value) using topGO analysis. P-values were obtained using the one-sided Fisher's exact test without multiple-testing correction. P values ¡ 0.005 were considered significant. **(j-k)**. Space-constrained CellChat inferred cell-cell communications with cell-cell communications signaling sent from Invasive-1 in Collagen pathway (**(j)**) and from In-situ cancer 1 in MK and APP pathways (k).

3.3 Discussion

In this work, we presented SpaceFlow, which (1) encodes the ST data into low-dimensional embeddings reflecting both expression similarity and the spatial proximity of cells in ST data, (2) incorporates spatiotemporal relationships of cells or spots in ST data through a pseudo-Spatiotemporal Map (pSM) derived from the embeddings, and (3) identifies spatial domains with consistent expression patterns, clear boundaries, and less noise.

SpaceFlow achieves competitive segmentation performance with alternative methods when benchmarked against expert annotations. Furthermore, the pSM utilizes the spatially consistent embeddings to reveal pseudo-spatiotemporal patterns in tissue. In DLPFC and Stereo-seq data, the pSM shows layered patterns that are consistent with the developmental sequences of the human cortex and mouse olfactory bulb respectively, which is not visible from non-spatial pseudotime. Applied to chicken heart developmental data, the pSM reveals evolving lineage structures and uncovers the dynamics in the spatiotemporal relationships of cells across different developmental stages, helping to understand the changes of func-

tional and structural organization in tissue development. Studying human breast cancer ST data using SpaceFlow, we demonstrate its potential to identify tumor-immune interfaces and dynamics of cancer progression, providing tools to study tumor evolution and interactions between tumor and the tumor microenvironment.

Though similarity in gene expression and spatial proximity are related in many cases [126], this relationship is not absolute. Pseudotime methods developed for scRNA-seq data, such as Monocle [127] and Slingshot [128] can produce developmental trajectories that are not spatially organized. The pSM developed here can generate spatially contiguous trajectories based on the integrative usage of gene expression and spatial information. Specifically, the spatial regularization in SpaceFlow constrains the low-dimensional embedding spatially so that the embedding is continuous both in space and time. The low-dimensional spatial constraint also reduces noise in the high-dimensional gene expression data resulting in smoother domain segmentation boundaries and spatiotemporal maps.

In practice, the training time of SpaceFlow on ST data with fewer than 10,000 cells is usually less than 5 minutes on a GPU. The computational cost of training largely depends on the calculation of spatial regularization loss for model optimization, which is quadratic to the number of cells or spots. To accelerate model training, we compute this regularization loss over a random subset of cell-cell pairs (Details in Methods). With a fixed number of cell pairs in the subset, the training can scale linearly with the number of cells or spots, and it has been shown not affecting the outcome (Supplementary Fig. 2b-e). In the current implementation, the training will automatically be switched to the approximated regularization strategy when detecting a cell population larger than 10,000. With this strategy, training time varies from 30 seconds to 3 minutes for numbers of cells/spots ranging from 3,000 to 50,000 on GeForce RTX 2080 Ti GPU. Future work could explore possible alternatives to selecting random subsets such as density-based subsampling [99, 100, 129], which may be more accurate for estimating the regularization loss.

The spatial regularization used in this work reflect the a priori assumption that nearby cells

with similar gene expression are more closely related than spatially distant cells with the same level of transcriptional similarity. In connected tissues with low geometric complexity examined in this work, the current spatial regularization with Euclidean distance has good performance. However, it may not cover the complexity of the spatial distribution patterns and dependencies that may vary among different locations of a tissue. Extension of regularization for disjoint tissues like lymph nodes or tissues with high geometric complexity can be developed by location adaptive spatial regularizations. The general framework proposed in SpaceFlow can also be easily extended by combining the latent space regularization with other choices of embedding algorithms, which may offer various tradeoffs in terms of expressive ability and computational efficiency. However, we expect that the general principle that explicit regularization for spatial structure improves performance on ST to hold for a variety of different embedding architectures.

In addition to spatial regularization, SpaceFlow is a flexible framework able to incorporate auxiliary features about connectivity among cells in spatial or single-cell omics data. For example, it can be directly applied to 3D ST data with spatial graph input based on 3D coordinates. Future improvement could be achieved by adapting the framework for spatially resolved Epigenetic data with proper preprocessing steps, such as peak calling on spatially resolved chromatin modification data [130]. Other non-genomic data modalities, such as the local texture features from histological images or expert domain annotation priors could be used to improve the robustness of the SpaceFlow embeddings. Under the SpaceFlow framework, different regularization terms reflect different prior knowledge about the tissue organization and their integration might enhance the performance of the result. In addition, the directed connectivity matrix inferred by RNA velocity [131] could be used as a constraint to derive low-dimensional embeddings consistent with RNA velocity which may improve the representation of developmental trajectory. Overall, SpaceFlow provides a robust framework and an effective tool to incorporate prior knowledge or spatial constraints to ST data analysis for inference of spatiotemporal patterns of cells in tissues.

3.4 Methods

3.4.1 Data preprocessing

The raw count expression matrix of ST data is preprocessed as the following. First, genes with expression in fewer than 3 cells and cells with expression of fewer than 100 genes are removed. Next, normalization is performed, where the expression of each gene is divided by total expression in that cell, so that every cell has the same total count after normalization. Then, the normalized expression is multiplied by a scale factor (10,000 by default) and log-transformed with a pseudo-count one. The log-transformed expression matrix of the top 3,000 highly variable genes (HVGs) is then selected as the input for constructing the spatial expression graph. We adopted a dispersion-based method to select highly expressed genes67. The genes are put into 20 bins based on their mean expression, and then the normalized dispersion is computed as the absolute difference between dispersion (variance/mean) and median dispersion of the expression mean, normalized by the median absolute deviation of each bin. Genes with high dispersion in each bin are then selected.

3.4.2 Construction of Spatial Expression Graph

We next convert the log-transformed expression matrix of highly expressed genes into a Spatial Expression Graph (SEG) as the input of our deep graph network. The Spatial Expression Graph is built based on the spatial proximity of cells, with nodes representing cells with expression profiles attached, while edges characterizing the spatial neighborhood of cells. Similarly, in spot-resolution ST data, we use a node to represent a spot in the graph. The SEG is characterized by two matrices, expression matrix $\mathbf{X} = \{x_1, x_2, .., x_n\}$ and spatial adjacency matrix $\mathbf{A} \in \mathbb{R}^{(N \times N)}$. Here, x_i represents the expression features of the cell or spot i, while the element $\mathbf{A}_{i,j}$ in adjacency matrix is equal to 1 if there is an edge between cell/spot i and j, otherwise, $\mathbf{A}_(i, j) = 0$.

We provide two methods for constructing the SEG, namely, alpha-complex-based and k-

nearest-neighbor-based. The alpha-complex-based method is used by default, where a Voronoi cell is first created for each cell or spot located at r as:

$$V(r) = \{x \in \mathbb{R}^2 | \|x - r\| \leq \|x - r'\|, \forall r' \in \mathbb{C}\} \tag{3.1}$$

where $\mathbb{C}$ is the set of coordinates for all the cells or spots, and$\|\cdot\|$ is the Euclidean distance. Next, the 1-skeleton of the alpha complex [132] is used to determine the neighborhood edges $\mathbf{E}$ of the spots, which can be formulated as follows:

$$\mathbf{E} = \{(i,j) | \cap_{k \in \{i,j\}} (V(r_k \cap B(r_k, \delta)))\} \tag{3.2}$$

Where $B(r_k, \delta)$ is a circle area in $\mathbb{R}^2$ centered at x with a radius δ. The radius δ is estimated by the mean distance of k nearest neighbors of the spot. In k-nearest-neighbor-based method, the edges of SEG are built based on the top k nearest neighbors of cells.

3.4.3 Spatially-regularized Deep Graph Infomax

To encode the ST data into low-dimensional embeddings of cells or spots, we use the Deep Graph Infomax (DGI) [105], an unsupervised graph network, as the framework of our model. DGI has the advantage of capturing not only the cell expression patterns but also the cell neighborhood microenvironment, as well as high-level patterns, such as global or regional patterns. Specifically, a two-layer Graph Convolutional Network (GCN) is used as the encoder of DGI with SEG as input. The GCN generates node embeddings $\varepsilon(\mathbf{X}, \mathbf{A}) = \mathbf{H} = \{h_1, h_2, .., h_n\}$ for each cell or spot.

DGI adopts a contrastive learning strategy [133] to learn the encoder, where features are learned through teaching which data points from an unlabeled dataset are similar or distinct. Similar data points are constructed by pairing cell embedding h_i with a global summary vector $\boldsymbol{s}$, whereas the distinct data points are represented by the pairs of the summary vector $\boldsymbol{s}$ and embeddings from a constructed Expression Permuted Graph (EPG). The summary

vector s reflects global patterns of SEG, and it is implemented by a sigmoid of the mean of all cell embeddings. EPG is a graph built by random permutating the node features $\mathbf{X}$ in SEG, with the adjacency $\mathbf{A}$ keeping the same. Mathematically, this learning process is achieved by maximizing the following objective function:

$$L_{\text{DGI}} = \frac{1}{2N}(\sum_{i=1}^{N} \mathbb{E}_{(\mathbf{X},\mathbf{A})}[\log D(h_i, \boldsymbol{s})] + \sum_{j=1}^{N} \mathbb{E}_{(\tilde{\mathbf{X}},\tilde{\mathbf{A}})}[1 - \log D(\tilde{h_j}, \boldsymbol{s})])\} \tag{3.3}$$

where h_i is the embedding of node i from the SEG, $\tilde{h_j}$ is the embedding of node j from the EPG. $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{A}}$ are the permuted node features and corresponding adjacency matrix of EPG. The D is the discriminator, which is defined by $D(h_i, \boldsymbol{s}) = Sigmoid(h_i^T \Theta s)$, where $\Theta \in \mathbb{R}^{(N_F \times N_F)}$ is trainable weight. Through this contrastive learning strategy, the encoder is forced to learn global patterns and neglect random spatial expression patterns in the embeddings.

To enforce the spatial consistency in the embeddings, so that the closeness between embeddings not only reflects the expression similarity but also their spatial proximity, we add a spatial regularization to the objective function in DGI. Mathematically, the revised objective function can be expressed as follows:

$$L_{\text{Total}} = L_{\text{DGI}} + \gamma * \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\mathbf{D}_{i,j}^{(s)} * (1 - \mathbf{D}_{i,j}^{(z)})}{N * N} \tag{3.4}$$

, where $\mathbf{D}^{(s)}$, and $\mathbf{D}^{(z)}$ are the pair-wised distance matrices of cells in physical space and embedding space.

3.4.4 Domain segmentation and pseudo-Spatiotemporal Map

The domain segmentations are obtained by running Leiden clustering [134] with the low-dimensional embeddings from SpaceFlow as input. By default, the parameter for the lo-

cal neighborhood size is set to 50 for to produce a smoother segmentation. The pseudo-Spatiotemporal Map (pSM) is calculated by running the diffusion pseudotime (DPT) [98] using the low-dimensional embeddings output from SpaceFlow. The DPT is an algorithm using diffusion-like random walks to estimate the ordering and transitions between cells. Using the embeddings from SpaceFlow that encoded both spatial and expression information of cells as input, DPT can output a spatiotemporal order which is consistent in both space and pseudotime. The root cell for pSM can be specified with prior knowledge, otherwise, in default, the cell that with the largest sum distance to others in embedding space is assigned as the root cell in our strategy.

3.4.5 Parameters of the model

The deep graph network is built and trained based on PyTorch. To construct SEG, the default number of nearest neighbors k of a cell or spot for adding edges is set to 15; A larger k will lead to a bigger spatial neighborhood. The DeepGraphInfomax model in PyTorch Geometric library is used for implementing DGI. The default latent dimension size for low-dimensional embeddings is set to 50. A two-layer Graph Convolutional Network (GCN) is utilized as the encoder for SEG with Parametric ReLUs (PReLU) as the activation functions. The number of neurons for both layers is set equal to the low-dimensional embedding size.

3.4.6 Training procedure

The optimizer used for training DGI is Adam with a default learning rate lr=0.001 applied [135]. The maximum number of epochs for training is set to 1000, with an early stopping strategy applied to avoid overfitting. Specifically, the minimum epoch for early stopping is set to 100, and the patience of epochs with no loss decrease is set to 50. A GeForce RTX 2080 Ti GPU is used for training the DGI model. The training time varies from 30 seconds to 3 minutes numbers of cells/spots ranging from 3,000 to 50,000, and the subsampling strategy stated below needed to be applied when the number is greater than 10,000.

3.4.7 Accelerating the computation of spatial regularization loss

Because the computational cost of training largely depends on the calculation of spatial regularization loss, which is quadratic to the number of cells or spots, we designed a strategy as follows to accelerating the training. The spatial regularization loss is used in model optimization, which involves calculating the weighted average of an inner product of a spatial distance matrix and an embedding distance matrix. It has $O(M \times N^2)$ computational complexity and memory cost, where N^2 is the number of edges in a fully-connected spatial graph with N cells, M is the size of the latent dimension. However, during each training step, we compute the spatial regularization loss over a random fixed-size subset of edges, which reduces the computational complexity of regularization loss from quadratic to constant. When tested on the slideseqv2 dataset with 41,876 cells, the computational and memory cost dropped from over 5 hours and 18GB to less than 3 minutes and 4GB (Supplementary Fig. 4). We additionally found significant improvements in performance applying SpaceFlow to a seqFISH mouse embryogenesis dataset [136] (Supplementary Fig. 5).

3.4.8 Benchmarking

Segmentation Benchmarking

To benchmark domain segmentation performance, we compare SpaceFlow against five methods, Seurat 4 [106], Giotto [86], stLearn [89], MERINGUE [91], BayesSpace [87] using the LIBD human dorsolateral prefrontal cortex (DLPFC) ST data [107]. To make the domains comparable between benchmarking methods, we set the target number of clusters equal to the number of clusters in annotation for all benchmarking methods. The adjusted Rand index (ARI) is used to quantify the similarity between the clustering result and the annotation.

With Seurat, the RNA transcript counts are used for the input, with genes expressed in fewer than 3 cells filtered, and cells expressing fewer than 100 genes removed. Then, the

SCTransform function in Seurat R package is applied to normalize the UMI count data using regularized negative binomial regression. Next, the RunUMAP, FindNeighbors, FindClusters methods are performed on the normalized count data sequentially with the latent dimension size of 50 and default cluster resolution of 0.4.

When benchmarking with stLearn, the count matrix and the spot positions were used as input, which is downloaded directly from the data sources (see Data Availability). The count matrix input was read via Read10X function in the stLearn package. Next, filter_genes, normalize_total, log1p, run_pca functions were applied sequentially to preprocess data, with the minimal number of genes for filtering set to 3. Next, the histological image of the tissue is preprocessed by using the tiling and extract_features functions. Then, the SME_normalize function is used with the parameter setting of use_data="raw" and weights="physical_distance". Finally, the scale and run_pca are performed on the normalized data with number of principal components of 50. The principal components from normalized data will then be used for segmentation or pseudotime analysis via Leiden and DPT, respectively.

With Giotto, we input the count matrix and the spot positions, and then applied the normalizeGiotto, addStatistics, calculateHVG to preprocess data and identify highly variable genes (HVG). HVGs expressed in at least 3 cells and with a mean normalized expression greater than 0.4 are then feed into runPCA function for the principal components. The spatial network was then created through the createSpatialNetwork function with the parameter for the kNN method method set to k=5 and a maximum distance of 400 in kNN. Finally, the doHMRF method is used for clustering with the parameter beta set to 40.

For BayesSpace benchmarking, we input expression matrix and the spot positions through the getRDS("2020_maynard_prefrontal-cortex") method. Next, the modelGeneVar and getTopHVGs methods in scran method are used to model the variance of log-expression profile of each gene and extract the top 2000 highly variable genes. Then, the runPCA function in the scater package is used for principal components. The BayesSpace clustering method

`spatialCluster` is applied with 15 principal components, with 50,000 MCMC iterations and `gamma=3` for smoothing.

For MERINGUE benchmarking, we input the spatial locations and the top 50 principal components from the expression matrix of ST data. Next, the spatial adjacency weight matrix is constructed using the `getSpatialNeighbors` function in the R package of MERINGUE, with a setting of `filterDist=2`. Then, `getSpatiallyInformedClusters` is performed to get spatially informed clusters by weighting graph-based clustering with spatial information, with a setting of `k=20, alpha=1, beta=1`.

Pseudo-Spatiotemporal Map Benchmarking

The pSM is compared to the spatial embedding method stLearn [89], and three non-spatial pseudotime methods, Seurat 4 [106], Monocle [127], and Slingshot [128]. In stLearn, because the histological image of the tissue is required for spatial-aware embedding, we only made comparisons when the histological was available. We calculated pseudotime from stLearn by running the diffusion pseudotime (DPT) [98] using the stLearn embedding. In Seurat 4, the DPT is run using the principal components of the expression data, whereas in Monocle and Slingshot, the recommended workflows with the default parameters are performed.

3.4.9 Downstream analysis

Marker genes identification

To identify marker genes that can best characterize specific expressions for domains output from SpaceFlow, the `rank_genes_groups` method in the Scanpy package (v1.8.2) is used. When performing this method, the Wilcoxon rank-sum test with a Benjamini-Hochberg p-value correction is applied. The cutoff of the adjusted p-value for domain-specific marker genes is set to 0.01.

Domain Annotation

The domains identified by SpaceFlow are annotated based on the literature report of the domain-specific marker genes.

Gene Ontology Enrichment Analysis

The Gene Ontology (GO) Enrichment Analysis in the GO Consortium website is carried out to identify the enriched GO terms for domain-specific maker genes with adjusted p-value ¡ 0.01.

Space-constrained CellChat Analysis

CellChat analysis is performed on ST data using the domain labels from SpaceFlow as groupings. Inferred CellChat communications between domains are further scrutinized such that the communication links are only allowed between spatially adjacent domains. The CellChat v1.1.3 is used under a R v4.1.2 environment.

3.4.10 Data availability

All data analyzed in this paper can be downloaded in raw form from the original publication. Specifically, the DLPFC data is available in the spatialLIBD package. The processed Stereo-seq data from mouse olfactory bulb tissue is accessible at SEDR analyses. The chicken heart ST data is retrieved from GEO database under accession code GSE149457. The human breast cancer ST data can be obtained from the Zenodo dataset 4751624. The sample we used is the same as the one demonstrated in the original paper (patient G-sample 1). Both chicken heart ST data and breast cancer ST data were sequenced by 10x Visium platform. The Slide-seq V2 can be accessed in Squidpy package [136] or downloaded from Broad Institute database. The seqFISH data can be accessed at the Spatial Mouse Atlas. The Gene Ontology Consortium database can be accessed via Gene Ontology Consortium.

Chapter 4

Identifying Cell-Cell Communication Patterns in Spatial Transcriptome

4.1 Introduction

4.1.1 Cell-cell communication

Cell-cell communication is an essential process in cellular biology that facilitates the interaction and coordination of activities among cells. It is a critical element in various biological processes, including development, tissue homeostasis, and disease. The process encompasses the exchange of information between cells through different mechanisms, such as signaling molecules, gap junctions, and cell surface receptors. This ability enables cells to adapt to changing conditions, respond to their environment, and perform specialized functions.

There are several types of cell-cell communication, which include paracrine signaling, juxtacrine signaling, and endocrine signaling [1]. In paracrine signaling, a cell releases signaling molecules that have a local effect on neighboring cells. Juxtacrine signaling involves the interaction between two adjacent cells through direct contact, such as binding of cell surface receptors. Endocrine signaling occurs when signaling molecules released into the bloodstream act on distant target cells.

To directly measure the proteins involved in cell-cell communication (CCC), specialized biochemical techniques and a thorough understanding of protein domains are necessary. Standard assays for studying the protein-protein interactions (PPIs) underlying CCC include yeast two-hybrid screening, co-immunoprecipitation, proximity labelling proteomics, fluorescence resonance energy transfer imaging, and X-ray crystallography [1].

Cell-cell communication plays critical roles in lots of important biological process. For instance, organism development relys on accurate cell communication in both temporal and spatial, therefore unraveling CCC dynamics helps to understand the cell fate decision mechanism. As another example, although the immune system receives signals from various tissues, only certain signals enable it to coordinate a properly functioning immune response [1]. Cell-cell communication can be regulated at multiple levels, including the expression of signaling molecules and receptors, the activation of signaling pathways, and the localization of signaling complexes. Dysregulation of cell-cell communication can lead to various diseases, such as cancer, autoimmune disorders, and developmental disorders.

4.1.2 Spatial transcriptome

Spatial Transcriptomics (ST) is a technique that involves the spatially resolved capture of RNA molecules from tissue sections, which can be sequenced and analyzed to reconstruct the gene expression profiles of individual cells *in situ*. It provides a high-resolution map of the gene expression patterns in the context of tissue architecture, allowing the identification of cell types, cell-cell communication networks, and spatially regulated gene expression [3].

Spatial transcriptomics technologies can be classified into two main categories. The first category is next-generation sequencing (NGS)-based, which involves encoding positional information onto transcripts before sequencing. Representative technologies includes: Visium [137], Slide-seq [138, 139], etc. The second category is imaging-based, which includes *in situ* sequencing (ISS)-based methods, where transcripts are amplified and sequenced directly in the tissue, and ISH-based methods, which involve sequentially hybridizing imaging probes

in the tissue [140]. Popular fluorescence *in situ* hybridization (FISH) based techniques are MERFISH [141] and seqFish+ [142].

4.1.3 Cell-cell communication inference

To infer Cell-Cell Communication (CCC), existing techniques developed assume the gene expression can represents protein abundance, and the protein abundance is sufficient to infer PPI strength. The estimation of the CCC likelihood is typically based on gene co-expression from transcriptomic data, where a curated gene list from literature is used as the potential interacting proteins [1]. Existing CCC inference methods can be categorized into 1. differential combination based, 2. expression permutation based, 3. graph or network based. In Differential combination based method, cell-cell communication is assigned when both ligand and receptor are differentially expressed in a given pair of cells. Representative methods are : CellTalker [143], iTALK [144], PyMINEr [145]. On the other hand, the expression permutation-based methods determine a communication score for each ligand-receptor pair and then assess its significance using methods such as permutation of cluster labels, non-parametric tests to compare against a null model, or empirical approaches [1]. This category of method is most widely used across the three, popular methods are CellChat [117], CellPhoneDB [146], Giotto [147], NeuronChat [148], etc. Network based based methods usually utilize the gene gene interaction information to build a network of ligand–receptor relationships. For instance, SoptSC [149] and NicheNet [150] both integrates the downstream signalling expression in the model.

The above approaches for CCC inference all target for scRNA-seq dataset, however, a important factor in cell-cell communication, i.e. spatial, is neglected. This could results in false-positives in the identified interactions because paracrine signaling can only work in a limited spatial range [1]. To better utilize the spatial transcriptome data and improve the confidence of CCC inference, serveral spatial CCC inference methods have been developed. For example, SpaOTsc builds an optimal transport model to infer CCC from signal senders

to target signal receivers in space [151]. SpaTalk utilize graph network and knowledge graph techniques, and combines the concepts of ligand-receptor proximity and ligand-receptor-target (LRT) co-expression consideration, to model and score the patially resolved cell-cell communication [152]. stMLnet infers spatial cell-cell communication by quantifying diffusion based LR signaling and mass action models [153].

4.1.4 Research gap

Although several spatial intercellular communication inference tools have been developed, there are still some challenges remained for researchers, which impedes their analysis process and hinders the study of cell-cell communication. First, although existing tools can infer the likelihood of spatial cell communication, the output usually contains the communications likelihoods of thousands of ligand-receptor pairs. Moreover, different ligand-receptor pair may have similar or totally different spatial CCC patterns, which can indicate their specific functionalities in biological process. However, currently, there is no available tools for summarizing the major spatial patterns and revealing the representative LRs for each pattern. Instead, researches have to investigate the spatial CCC patterns of thousands of LRs one by one, which can take huge amount of time and work. As an initial trial, Cellchat provides pattern recognition based tool to discover dominant cell communication patterns. However, this tool can only identify the patterns at cell type resolution, and does not provide any spatial context, which can be a critical issue for understanding spatial cell communications in tissue. Therefore, based on the existing spatial CCC inference tool, we aims at developing methods which can summarize the major spatial CCC patterns with biological interpretation, such as the associated LRs for each pattern. This will significantly reduce the burden of analysis for researchers so that they can focus on discover the biological insights and potential applications behind these patterns. To aid that, we will develop downstream analysis approach for the identified patterns to understand them through its pathway composition, and protein interaction network, etc.

4.2 Methods

We investigated two approaches for analyzing and summarizing patterns of spatial cell-cell communication. Both approaches rely on Graph Neural Networks (GNN), with one being unsupervised and the other being supervised. The supervised GNN method consists of two main steps: pre-hierarchical clustering and followed by cluster-regularized graph neural network.

4.2.1 Unsupervised GNN approach

The unsupervised GNN approach does not utilize the clustering as a guidance, but learn the spatial CCC patterns through a unsupervised Encoder-Decoder with specifically designed regularizations (Figure 4.1).

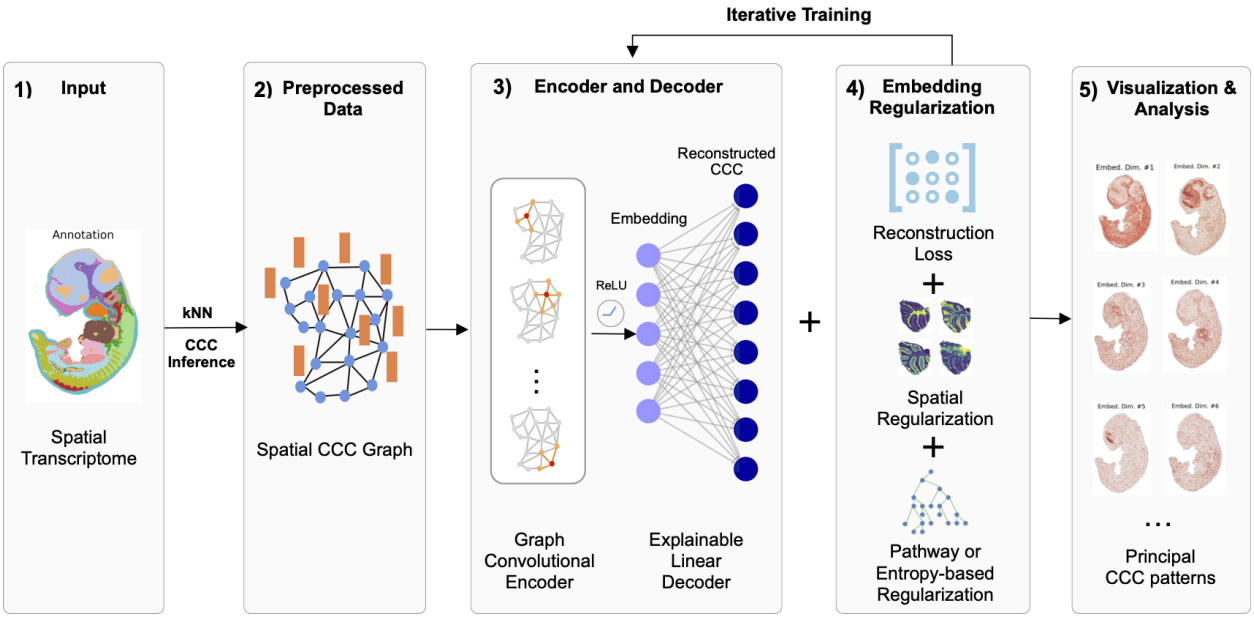

Figure 4.1: **The workflow of the unsupervised GNN approach.**

Given a spatial transcriptome data with a expression matrix ($n\_cell \times n\_cell$) and corresponding spatial coordinates ($n\_cell \times 2$), we first infer the spatial cell-cell communication through COMMOT's `spatial_communication` function [154], with CellChat [117] database used as default. To limit the spatial range of CCC, we set the parameter `dis_thr` (distance

threshold) by the maximium distance in the k-nearest neighbor(kNN) built on spatial coordinates, where k is set by 15 based on our experience. The inferred CCC matrix and the kNN graph together form a spatial CCC graph (Figure 4.1 (2)). Next, a Encoder-Decoder framework is employed to transform the Spatial CCC graph into low-dimensional embedding. Specifically, the Encoder includes one Graph Convolutional layer, which can aggregate spatial neighborhood CCC information. The Decoder, on the other hand, is a one layer fully connected network (linear condensed layer), which is used to reconstruct the spatial CCC as close to input as possible. During the training, specific regularization techniques are applied to the Encoder-Decoder to better learn the spatial CCC patterns. First, a Mean Square Error(MSE) metric is used as reconstruction loss, which takes the input CCC feature matrix (dimension: $n\_cell \times n\_LRs$) and the constructed CCC matrix as input. In addition, a spatial regularization is applied to constrain the embedding space, such that the embeddings close to each other also close in physical space (as been demonstrated in our previous work [155]). The mathematical expression of the spatial regularization is:

$$L_{\text{spatial}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\mathbf{D}_{i,j}^{(s)} * (1 - \mathbf{D}_{i,j}^{(z)})}{N * N} \tag{4.1}$$

where $\mathbf{D}^{(s)}$, and $\mathbf{D}^{(z)}$ are the pair-wised distance matrices of cells in physical space and embedding space. N represent the number of cells/spots.

Exlusive Sparsity regularization

To restrict the connections in weight matrix, we used a Exlusive Sparsity regularization so that each LR in input only contribute to mutual exclusive embeddings and contributed as few embedding as possible. This Exlusive Sparsity regularization is first proposed in [156], which here can be used to promotes each embedding to compete a few meaningful LRs from input. This is implemented through a (1,2)-norm, which enforce the weights of the network

fit to mutually exclusive sets of input features. Mathematically, this is:

$$\Omega(W) = \frac{1}{2} \sum_{g=1}^{n\_embed} (\sum_{i=1}^{n\_LRs} |W_{g,i}|)^2 \tag{4.2}$$

where g is embedding index, i is the index of LRs, W is the weight matrix in GCN.

Pathway-based regularization

To learn pathway specific spatial CCC patterns, alternatively, we can replace the Exclusive Sparsity Regularization with a Pathway regularization, which is a one-hot matrix initialization to weight matrix, similar to the initialization by cluster labels. Mathematically, the pathway based weights matrix is initialized by:

$$W_{i,j} = \begin{cases} 1, & \text{if } LR_i \text{ is from pathway } j. \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

where we retrieve the pathway membership of LRs through CellChat database [117]. The GNN will further refine the contribution of each LR to its pathway CCC patterns through training.

Visualization

We can visualize each embedding in spatial using the cell/spots locations (Figure 4.2(7)) by coloring the cells by embedding value. Each embedding represents one principal spatial CCC pattern summarized by the trained GNN, and can be used to analyze the enrichment of CCC in spatial and the spatial interactions between LRs through the downstream analysis.

4.2.2 Supervised GNN method

Pre-hierarchical Clustering

The workflow for the pre-hierarchical clustering is illustrated in Figure 4.2a. Given a spatial transcriptome data with a expression matrix ($n\_cell \times n\_cell$) and corresponding spatial coordinates ($n\_cell \times 2$), we first infer the spatial cell-cell communication through COMMOT. The output of COMMOT is a Cell-Cell Communication (CCC) matrix ($n\_cell \times n\_cell \times n\_LRs$), where n_LRs is the number of ligand-receptor pairs in CellChat database. To reduce computational burden and simplify the visualization of spatial CCC, we split each CCC entries in the obtained CCC matrix equally between the sender cell and receiver cell. This results in a distributed CCC matrix with $n\_cell \times n\_LRs$ dimension, which represents the aggregated CCC likelihood for each cell with respect to each ligand-receptor (Figure 4.2b).

In order to quantify the similarity of the spatial CCC patterns between LR pairs, we tested a lot of similarity metrics, such as the Mean Squared Error(MSE), Pearson Correlation, Structural Similarity Index (SSIM) [157], Jaccard index, etc., we found that the Jaccard index generated the best results. Moreover, we found that if we mesh the spatial locations of the cells into grids ($n\_grid \times n\_grid$), and aggregate the CCCs for cells in the same grid by sum, we can generate a better performance in clustering similar spatial CCC patterns. For instance, we calculated the similarity matrix and visualize the spatial ccc patterns in both raw and gridded format (Figure 4.2b-e), we can observe that the pattern in *Fgf8-Fgfr2* share similar parts with *Fgf7-Fgfr2*. This similarity is clear in the similarity matrix from the gridded ccc, but is not clear in the raw spatial ccc form, where the similarity between *Fgf8-Fgfr2* and *Fgf7-Fgfr2* is just 0.04. The generated gridded CCC matrix has the dimension of $n\_row \times n\_col \times n\_LRs$, where n_row and n_col are both set to 100 in default (Figure 4.2d). Next, we calculate the Jaccard Similarity Matrix ($n\_LRs \times n\_LRs$) to quantify the pair-wised similarity of spatial CCC patterns between LRs using their corresponding gridded

CCC matrices (Figure 4.2e). We then perform hierarchical clustering to group LRs with similar spatial CCC patterns (Figure 4.2f). The cluster labels assigned to LRs will be used as the regularization for the Graph Neural Network.

Cluster Supervised Graph Neural Network

As being illustrated in (Figure 4.2g), the Graph Neural Network (GNN) we constructed use Spatial Transcriptome (ST) data as input, which consists of the expression matrix of cells and their spatial locations. The ST data is preprocessed before feeding into the GNN. In preprocessing, we applied the k-nearest neighbor algorithm on cell/spot spatial locations to get an adjacent matrix A, and performed the CCC inference using COMMOT[154] to get a CCC feature matrix X. The preprocessed data next is input into a Graph Convolutional layer, whose weights are initialized by the one-hot cluster labels obtained from the pre-hiearchical clustering step. The shape of the weight matrix in GNN is $n\_LRs \times n\_embed$, which corresponds to the size of the input and output of the GNN. Specifically, the input shape of GNN is a $n\_cell \times n\_LRs$, while the output (i.e., embeddings) has a size of $n\_cell \times n\_embed$. Therefore, the weight matrix represents a mapping from LRs to embedding, which is initialized by the cluster labels. Mathematically, the weights matrix is initialized by:

$$W_{i,j} = \begin{cases} 1, & \text{if } LR_i \text{ is from clustering } j. \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Spatial Regularization

The output embedding and the weight matrix of GNN is refined during the iterative training and regularized by both spatial regularization (same as the previous) and Jaccard similarity regularization. An effect of the spatial regularization is smoothing and denoising the embeddings.

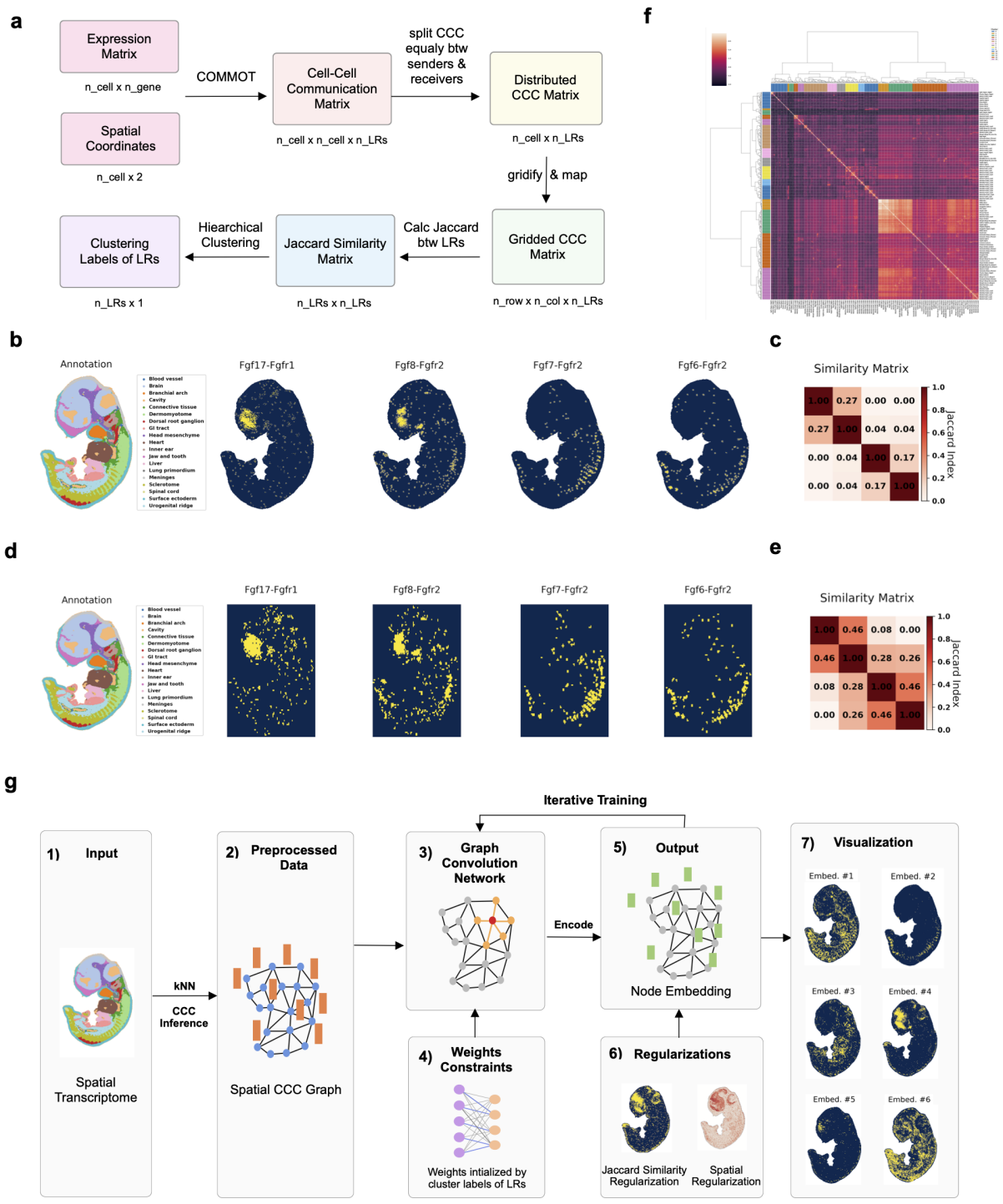

Figure 4.2: **The workflow and output of the pre-hierarchical clustering.** **(a)**. The workflow the pre-hierarchical clustering. **(b)**. The annotation of the cells in Mouse Embryogenesis ST data(left) [80], and four representative spatial CCC patterns inferred using COMMOT[154]. **(c)**. The jaccard similarity matrix for the four spatial CCC patterns. **(d)**-**(e)**, similar to **b, c**, but on the gridded spatial CCC patterns. **(f)**. The hiearchical clustering on the Jaccard similarity matrix calculated on the gridded spatial CCC patterns of all ligand-receptors. **(g)**. The architecture of the clustering supervised Graph Neural Network.

Jaccard Similarity Regularization

Another regularization added to GNN is Jaccard similarity regularization, which aims at constraint the embedding similar to the cluster-average spatial CCCs from pre-clustering. Mathematically, this is:

$$L_{\text{Jaccard}} = \frac{1}{M}\sum_{i=1}^{M} J(Z_i, \frac{1}{N_{Ci}}\sum_{j=1}^{N_{Ci}} X_{:,j}) \tag{4.5}$$

where M is the number of clusters, N_{Ci} number of LRs in cluster i, Z_i is the embedding i, $X_{:,j}$ is the CCCs of cells for LR_j, J is the Jaccard Index function.

Therefore, each trained embedding represents a specific spatial CCC pattern which summarizes the CCC patterns of a collection of LRs with similar spatial patterns.

Identification of potential spatial interacting LRs

One of the downstream analysis the output embedding can be utilized for is to identify the potential spatial interacting LRs. To achieve this, in the unsupervised method, we rank the LRs based on their trained GCN weights with respect to each embedding. For each embedding, we choose the top contributed LRs as the explanation of the embedding and the potential collections of LRs that are spatially interacting with each other.

On the other hand, in the supervised method, for each cluster, we rank the LRs based on their total CCC levels (sum over all cells). The reason for choosing sum instead of the top Jaccard index is that we find top Jaccard index tends to prioritize the LRs with less CCC signals. The LRs with more obvious patterns (in terms of CCC intensity) and high pattern similarity to embedding will be put to the last, when Jaccard index is used to rank the LRs for each cluster.

4.3 Results

4.3.1 Spatial regularization enables the identification of spatial patterns

To investigate whether the raw GNN based Encoder-Decoder framework can identify the Spatial CCC patterns in the input data, we applied the unsupervised GNN model (see Methods) on the mouse embryogenesis ST data [80]. The model includes a Graph Convolutional layer as the Encoder, followed by a fully connected layer as the decoder. In the raw form, only the Mean Square Error loss is added. After training the model, we visualize the output embeddings (Figure 4.3a). We can observe only top 11 embeddings output signals and the embeddings from #12 to #14 is just empty. In the top 11 embeddings, the output patterns are very noisy and similar to each other. The differences mainly come from the overall pattern intensity. For instance, it can be observed that embedding #5 shows the most visible pattern as the color is denser in specific regions. Embedding #3, #8 show the intermediate pattern intensity. The patterns in embedding #1 and #10 is just barely visible. To check whether the model can summarize spatial CCC patterns into the embeddings, we choose the embedding #6 and plots the spatial CCCs from the top 9 contributed LRs to this embedding according to the weights from GCN model. The reason for choosing embedding #6 is that, in the region annotated as heart in first panel of Figure 4.3a, the pattern is more obvious than the patterns in other regions. From the spatial CCCs of the top contributed LRs, we can hardly observe any similarity between the spatial distribution of the CCC patterns. Moreover, the CCCs in the heart region in the embedding #6 is not universally enriched across the CCCs in top contributed LRs.

Next, we replace the MSE loss by spatial regularization to test whether it can helps to identify the spatial CCC patterns that shared across a subset of LRs. Indeed, we observed the output embeddings present clearer and more diverse spatial patterns (Figure 4.4a). The better visibility of patterns can be seen through a greater contrast in color intensity between regions.

Figure 4.3: **Embeddings of the raw unsupervised GNN model: GCN Encoder + FC Deocoder with MSE loss only. (a)**. The first panel: the cell type annotation for the E11.5S1 sample from the MOSTA dataset [80], rest panels: the visualization of the embeddings output from the model. The intensity of red represents the embedding value. **(b)**. The embedding #6 and the visualization of the CCCs for the top 9 contributed LRs for this embedding ranked by the weights in GCN model.

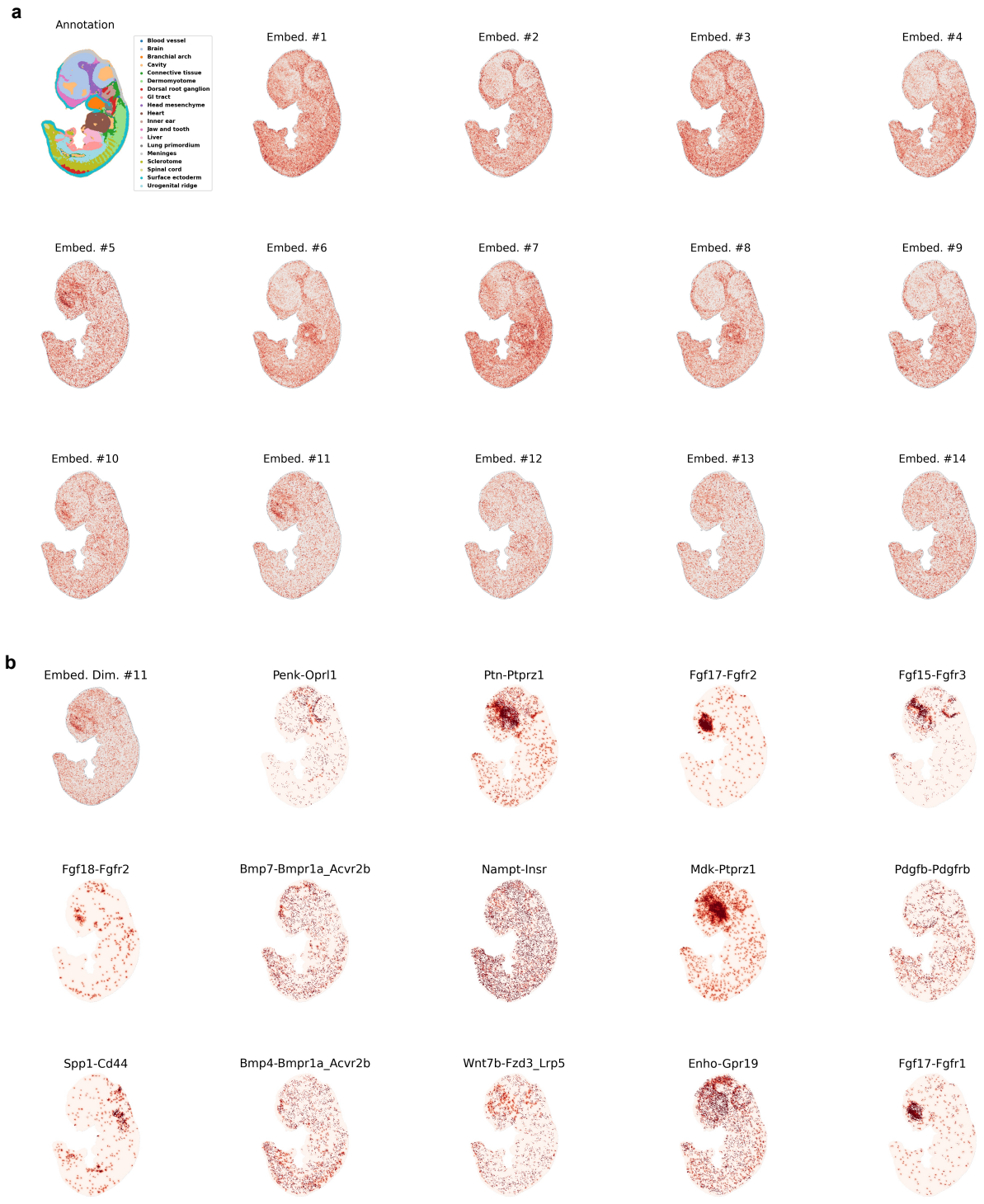

Figure 4.4: **Embeddings of the raw unsupervised GNN model: GCN Encoder + FC Deocoder with spatial regularization only. (a)**. The first panel: the cell type annotation for the E11.5S1 sample from the MOSTA dataset [80], rest panels: the visualization of the embeddings output from the model. **(b)**. The embedding #11 and the visualization of the CCCs for the top 14 contributed LRs for this embedding ranked by the weights in GCN model.

The pattern variation is shown by differences in the distribution of colors in the region with the highest intensity. Furthermore, we investigate whether the spatial regularization can helps to aggregate similar patterns into the same embedding. To check this, similar to the MSE model, we select embedding #11 and draw its top 14 contributed LRs. Unsurprisingly, we find more spatial overlap between the embedding and its top contributed LRs. Specifically, the patterns are enriched in the brain region in embedding, this regional pattern enrichness is observed in the following LRs: `Ptn-Ptprz1`, `Fgf17-Fgfr2`, `Fgf15-Fgfr3`, `Fgf18-Fgfr2`, `Mdk-Ptprz1`, `Wnt7b-Fzd3_Lrp5`, `Fgf17-Fgfr1`, etc. Overall, the results suggests that the spatial regularization can helps to identify the spatial patterns in the input.

4.3.2 Regularizations for reducing the pattern redundancy in embedding

Although the spatial regularization can better identify the spatial patterns, the issue of redundant patterns in embedding still remains. We did some exploration on how to reduce them through direct regularization on the GCN weights. Multiple techniques were tried to regularize the GCN weights, such as L1 regularization, entropy based (which aims to limit the number of embeddings each LR contributes through entropy based loss), and Exclusive Sparsity regularization [156]. We will only introduce the result of the Exclusive Sparsity regularization, since it can promotes the embeddings to compete and select a subset of the input LRs so that the subsets are mutually exclusive (Figure 4.5a). We visualized the heatmap of the trained GCN weights in Figure 4.5b to verify its effect. Indeed, we find the LRs for each embeddings are not overlapped and each LR only contributed to a limited number of embedding (usually one). By contrast, in the heatmap of the entropy based weight regularization, we observed each LR ususally contributes to over 2 embeddings, and sometimes over 90% of the embeddings (see SI Figure). However, although the weights regularization is effective, the trained embedding still present severe redundancy issue (Figure 4.5c). In addition, the correlation of the spatial patterns between the embedding and its top

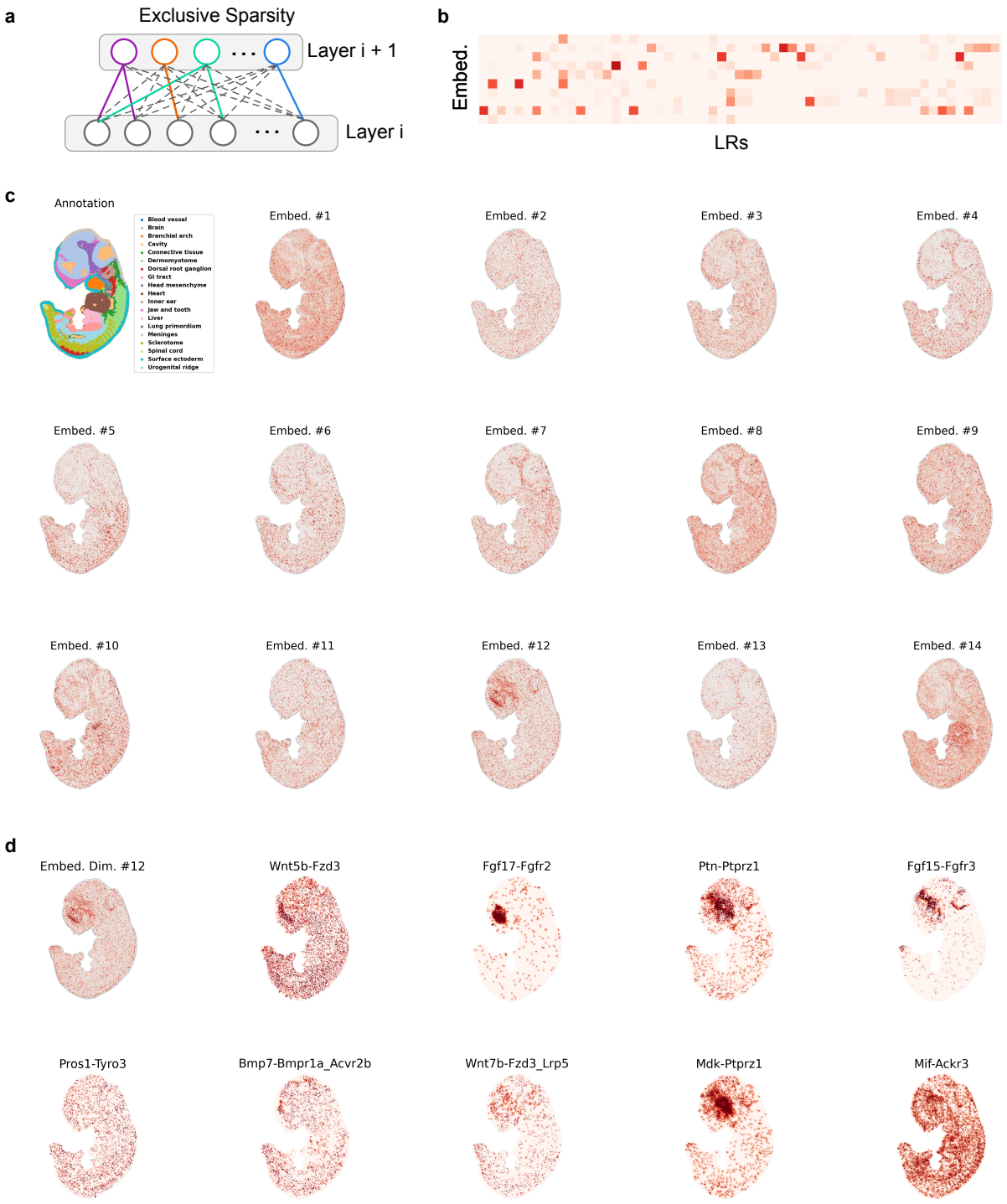

Figure 4.5: **Embeddings of the unsupervised GNN model: GCN Encoder + FC Deocoder with Exclusive Sparsity regularization [156] applied to GCN weights. (a)**. The illustration of the Exclusive Sparsity regularization, which promotes the upper layer neurons to compete and select a disjoint subset of the lower feature neurons. **(b)**. The heatmap of the GCN weights (in part) regularized by Exclusive Sparsity. **(c)**. The first panel: the cell type annotation for the E11.5S1 sample from the MOSTA dataset [80], rest panels: the visualization of the embeddings output from the model. **(d)**. The embedding #12 and the CCCs for the top 9 contributed LRs for this embedding ranked by the weights in GCN model.

contributed LRs is not improved as well.

4.3.3 Jaccard Index based hiearchical clustering can identify LRs with similar spatial patterns

A motivation of this study is to develop tools to ease the spatial CCC analysis for researchers, since there are usually thousands of LRs with different spatial CCC patterns to analyze. Therefore, the identification of a few major CCC patterns in the LRs is necessary, so that the identified patterns can be representative and diverse. However, based on the previous results, we find the top LRs contributed to the same embedding are not neccessarily similar in their spatial CCC patterns produced by GNN with Exclusive Sparsity regularization (or Entropy based Regularization). Therefore, we think the a explicit regularization, instead of the implicit ones, may be able to encourage LRs with similar spatial CCC patterns cluster into the same embedding. However, the form of this explicit regularization is hard to come up with. Thus, we try to first find a method which can cluster similar spatial CCC patterns well as a initial step.

As been mentioned in Methods, when clustering spatial patterns, a good metric to quantify their similarities is critical for the ultimate performance. An intuitive thought is to compare the difference of their CCC levels cell by cell between two LRs using Mean Square Error (MSE) or calculate the correlation with Pearson Correlation. In addition, image similarity metrics, such as Structural Similarity Index (SSIM) [157], seems also good candidates. However, upon investigation, we discovered that utilizing a metric that analyzes data at the level of individual cells typically leads to poor performance due to a high level of noise present in the data. To mitigate this issue, we mesh the cells in to a lower resolution grid (100×100), aggregate the CCCs of cells into grids, and then use these grids instead of raw data to calculate the pairwised similarity matrix. Through exploration, we find the metric of Jaccard Index can give the best performance in clustering similar spatial patterns probably because of it's a coarse grained metric and count on positive overlap instead of counting both

negative and positive (the data is extremely sparse and we only care to how much degree the positive parts are overlapped).

By hiearchical clustering the LRs using Jaccard Index, we obtained the information about which groups of LRs are similar to each other in spatial CCC patterns. To further derive the representative spatial CCC pattern for each cluster, we remodel the GNN network as the Figure 4.2g shows. Specifically, we removed the decoder and constrained the GCN weights matrix by initializing it using the one hot representation of the cluster labels. In this way, the LRs that does not belongs to the specific cluster will have zero weights for the embedding that corresponds to that cluster. In addition, we added the spatial regularization to the embedding to make the output embedding more spatially continuous. Furthermore, we added a Jaccard similarity regularization to make the embedding as similar to the cluster average CCC patterns as possible. This is more like a refinement to the cluster average CCC patterns so that it's more spatially continuous and less noisy.

Next, we visualized the trained Embeddings and annotated the cell types distribution in the cells with embedding signals (Figure 4.6a). From the figure, it can be observed that the spatial patterns are now quite diverse across embeddings. This is due to the Jaccard index based clustering, because we can find similar patterns if we calculate the average CCCs for LRs in each cluster (Figure 4.6b). By analyzing the pattern, we observed that the Embedding #2, #7, #8, #10 are enriched in brain cells (light blue annotation), but their spatial location is different, which may indicates distinct functions through CCCs. By contrast, embedding #4 shows significant enrichment in the dermomyotome and sclerotome, which might suggests its function role in somite development for the LRs associated with this embeddings. In embedding #6, we observe enrichment in heart, liver, GI tract, dorsal root ganglion, this could imply that these organs share similar CCCs for specific functions during mouse development.

Although the embedding and cluster average CCC is similar, we can still observe differences. For example, in the embedding (or cluster) #5, #13, #14, we can see the pattern in em-

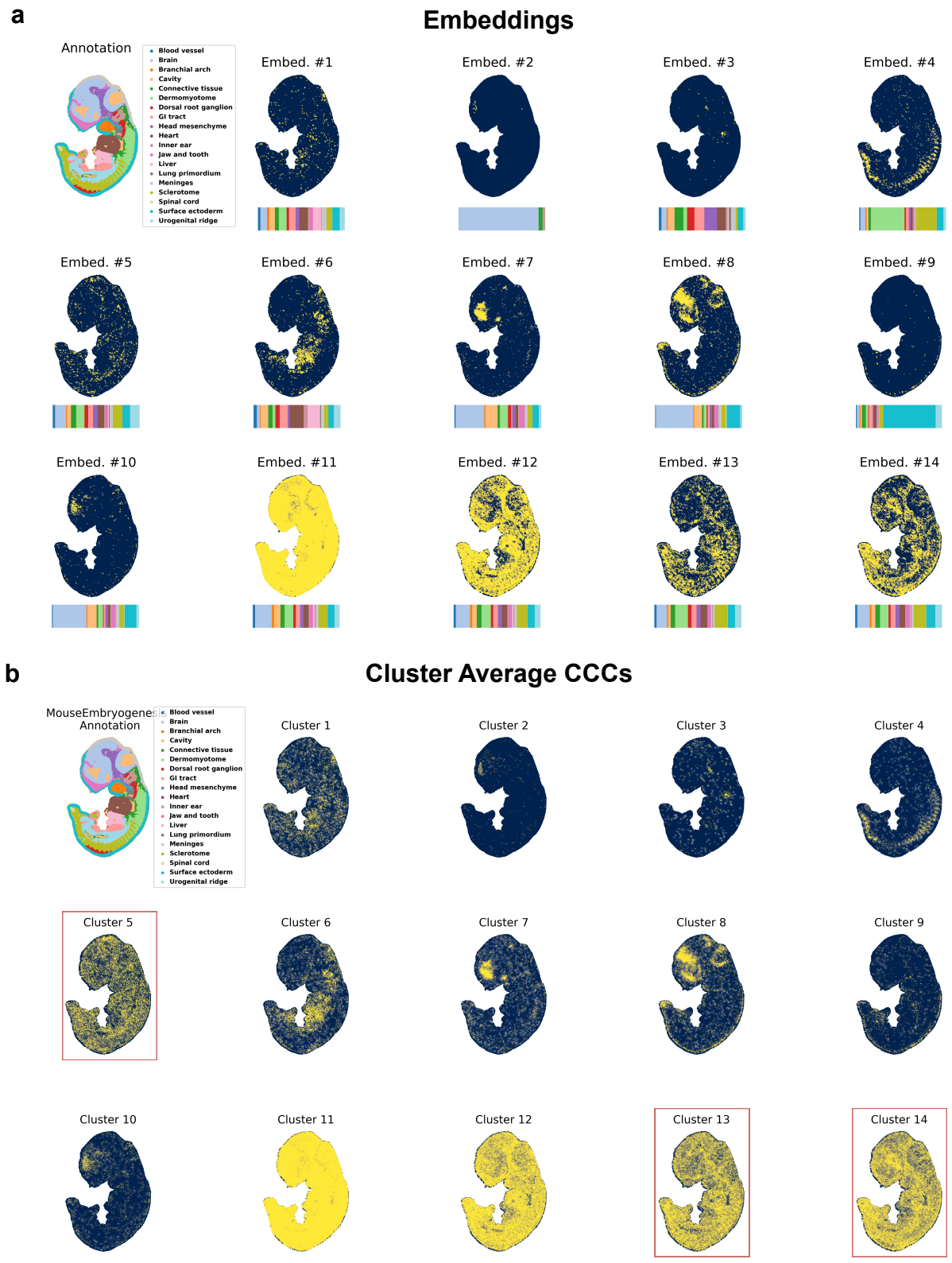

Figure 4.6: **Comparison between embeddings and cluster average CCC patterns (a)**. Visualization of Embeddings by GNN with Jaccard Index based Preclustering **(b)**. The Cluster Average CCC Patterns, where the CCCs for LRs in the same cluster is averaged.

bedding is much more clear and enriched in specific regions, whereas the CCC patterns in cluster average distribute across the whole sample.

4.3.4 Identifying potential interacting ligand-receptors (LRs) through embedding associated LRs

To identify the potential interacting LRs and understand the pattern source for each embedding, we hypothesize that interacting LRs are more likely to co-locate in spatial and therefore have similar spatial CCC patterns. Based on this hypothesis, we extracted the LRs in the same cluster as the top contributed LRs to each embedding (Figure 4.7g-l). It can be seen that the embeddings are now highly overlapped with the spatial CCC patterns for the top contributed LRs, which confirmed the effectiveness of the clustering. Moreover, we can observed that many contributed LRs within the same embedding usually share the same or similar ligand or receptor. For instance, in embedding #4, we find ligand `Fgf6`, `Fgf7`, receptor `Fgfr1`, `Fgfr2` appear multiple times in the contributed LRs (Figure 4.7g). To investigate how the embedding associated LRs may perform function collectively through interacting with each other, we inferred the interaction network with the embedding associated LRs using STRING.db (Figure 4.7m-o). Three groups of LRs can be seen for the interaction network for embedding #4. One is Fgf family, where ligands `Fgf4/5/6/7` interact with `Fgfr1/2/3/4` receptors. Gene Ontology (GO) analysis suggests `Fgf4` and `Fgf6` associate with cartilage condensation. Indeed, it's been reported that `Fgf6` has restricted expression in the myogenic lineage during the development of skeletal muscle, indicating its signaling role in somite formation [158, 159]. `Fgf7` is also suggested to play similar role [160]. These indicates the CCC patterns in embedding #4 associated with the somite formation process. In addition, GO analysis suggests `Fzd1/2/3` associate with neural tube closure.

As another example, we find although the patterns in embedding #7 and #8 are similar to each other and partially overlapped in brain region, their associated pathway composition is quite different (Figure 4.7k, l). Eembedding #7 is predominantly associated with `Bmp` and

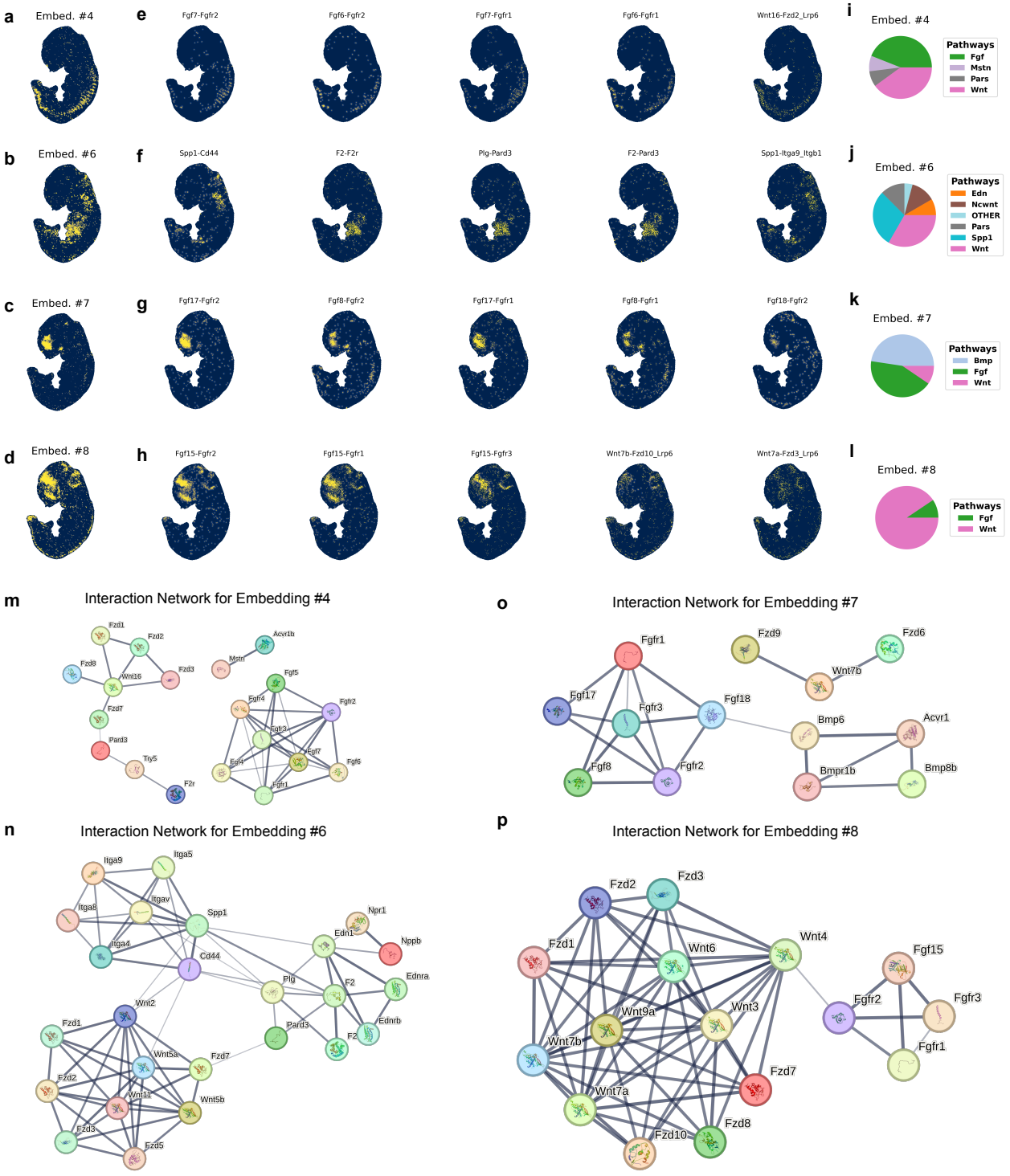

Figure 4.7: **Inferred cell-cell communication patterns in embeddings and the top contributed LRs for each embeddding (a)-(d)**. The selected representative embeddings. **(e)-(h)**. The top 5 contributed LRs corresponding to the embeddings in the first columns. **(i)-(l)**. The pathway distribution of all contributed LRs for the embeddings in the first columns. textbf(m)-**(p)**. Interaction Network for LRs Contributed to Embedding #4, #6, #7 and #8. Edges indicate both functional and physical protein associations from STRING.db. Edge thickness indicates the confidence from data support.

Fgf, whereas Wnt play a major role in Embedding #8. Interaction network of their LRs reveal this difference clearly (Figure 4.7o, p). In the interaction network of embedding #7, ligands Fgf8/17/18 interact with Fgfr1/2/3 receptors, Bmp6/8b interact with Bmp1r and Acvr1. Only one Wnt family ligand Wnt7b is in embedding #7. By contrast, Wnt3/4/6/7a/7b/9a interacts densely with Fzd1/2/3/7/8/10, and the only one Fgf family ligand Fgf15 is in embedding #8. These suggests although spatially similar, the cell cell communications in embedding #7 and #8 is distinct, and the LRs utilized across embeddings are not usually overlapped. Together, these examples demonstrates the usefulness of the embedding in CCC pattern summarization, and its potential to reveal LRs which might interact with each other and perform functions collectively.

4.3.5 More confirmative results from 10x Visium datasets

To further validate the effectiveness of the proposed method, we tested our pre-clustering GCN model on a 10x Visium ST dataset (Figure 4.8. Sample: Adult Mouse Brain Coronal Section 1). Similar to the previous analysis, we extracted the top embedding associated LRs for two representative embedding pattern (Embedding #5, #7). Again, we can see the highly overlapped patterns between the embedding and CCCs in embedding associated LRs, which confirmed the effectiveness of our method on a different sequencing technology.

4.4 Discussion

In this work, we explored the approaches for summarizing and analyzing the spatial cell-cell communication patterns. We proposed two graph neural network (GNN) based methods for achieving this, unsupervised GNN and supervised GNN. The unsupervised GNN model optimize the embedding through reconstruction loss, spatial regularization and exclusive sparsity regularization. However, the resulting embeddings have severe redundancy issue, and some embedding associated LRs do not share similar patterns with its embedding. To

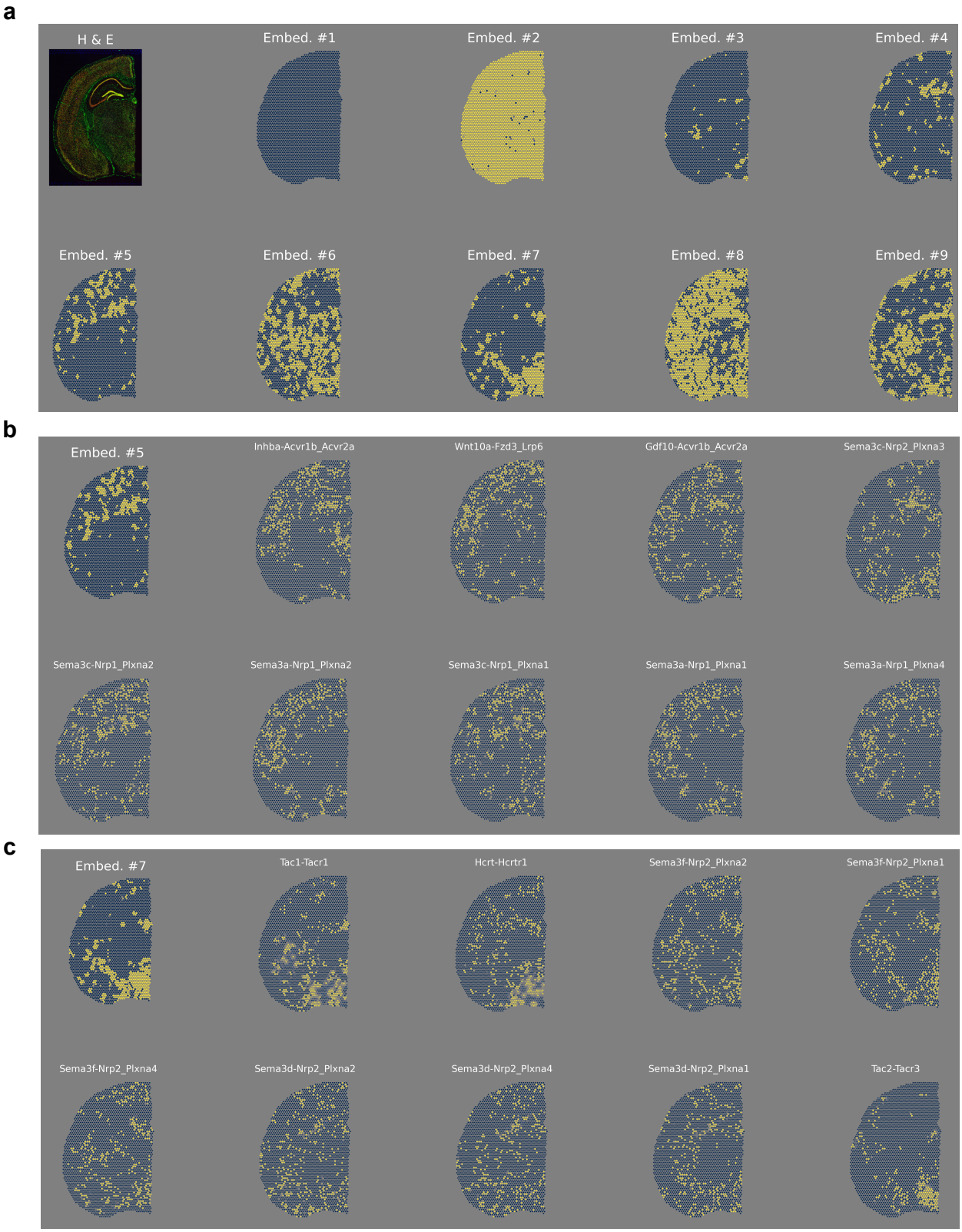

Figure 4.8: **Inferred cell-cell communication patterns in embeddings and the top contributed LRs for each embeddding (a)**. The trained embeddings from pre-hiearchical clustering GCN model. **(b)-(c)**. The top 9 embedding associated LRs corresponding to the representative embeddings: #5, and #7.

address this, we hypothesize a explicit regularization which based on a metric that can measure the similarity between spatial CCC patterns is necessary. However, the form of this explicit regularization is hard to come up with. Therefore, we proposed a pre-clustering strategy to supervise the GNN model. This strategy helps to group similar patterns into the same embedding and differentiate different patterns into separate embeddings. To clustering similar spatial CCC patterns, the choice of similarity metric is critical. We found the metric of Jaccard Index achieves the best performance among the metrics from Min Square Error, Pearson correlation, to other image similarity metrics such as SSIM. Using the clustering labels as the GCN weight initialization, the trained GNN model can identify diverse and biological meaningful spatial CCC patterns. Although the trained embeddings are similar to the cluster average CCC patterns, the patterns from embedding are much cleaner, which may help researchers to prioritize important patterns or spatial regions. To facilitate the interpretation of the patterns from embeddings, the top contributed LRs to each embedding are also generated. These LRs might represent potential interacting LRs, which can help us understand the full picture of cellular communication mechanism. We showcase the utility of the proposed methods on a mouse embryogenesis ST dataset. The analysis results demonstrated the effectiveness of the pre-clustering strategy in grouping similar spatial patterns and diversifying the embedding patterns. By exploring the protein interaction network revealed via the embedding associated LRs, we showed how this method can be used to uncover new biological insights.

Despite the usefulness of the proposed method in aiding researchers to analyze the biological significance of numerous LR patterns, there are still some unresolved issues. First, the pre-clustering based GNN model seems too ad-hoc, and the GNN model seems unnecessary compared with the results from cluster average CCCs. To improve this, an end-to-end deep learning approach with flexibility to serve multiple purpose for spatial CCC analysis would be ideal. However, based on the results from the unsupervised GNN, implicit regularization strategies seems not working well. Moreover, learned from the fact that Jaccard Index is

a great metric for clustering similar spatial patterns, it is reasonable to construct regularizations based on Jaccard similarity matrix between CCCs from LRs. One thing we tried with this is to regularize the GCN weights so that the connections to embeddings from LRs that have less Jaccard similarity would be penalized. However, the results seems not good. Another aspect the current work can be improved is providing the strategy to determine the best number of major patterns. Currently, we leave it as a free parameter to user in hierarchical clustering (i.e. `n_cluster`: # of cluster). A better approach could be founded on the principle of information gain, where the new information might be quantified by the similarities between the new embedding to all existing ones. If the maximum similarities are less than a threshold, adding the new embedding may provide richer pattern information. Finally, this study only investigated the approach to summarize spatial cell-cell communication patterns using a single sample of spatial transcriptomic data. There are numerous possibilities for extending this research, which could be both interesting and valuable. For example, making the embeddings comparable between different conditions, would be a interesting perspective to explore, such as comparison between healthy and disease samples, samples with different temporal points. Similarly, integrating multi-slice samples would also be valuable to study since it is the trend of ST in the future. One strategy to integrating multi-slice samples that works great and has been tested in SpaceFlow is integrating the embeddings from multiple slices by methods such as Harmony [161]. Moreover, we can further extend the current potential interacting LR network identified to their downstream targets to form a larger network for a larger picture. In addition, the downstream target of the potential interacting LR network may also have different CCC patterns or expression patterns. Investigating the expression and downstream impact of the CCCs in LR interacting network, which produce multiple heterogeneous response, might be interesting to study. Finally, we hope that this work may pave the way for better analyzing the spatial CCC patterns and contribute to the discovery of novel insights into cell-cell communication.

Chapter 5

Future Work

In this dissertation, we explored the spatial and temporal regulation of cells from three distinct perspectives, and revealed the complicated connections between transcriptome, epigenome, and intercellular communication. With the development of multi-omics, spatial transcriptomics, and live cell imaging, and the recent advances in AI, it is possible to develop approaches to consider these factors and their connections together. For example, we may soon be able to sequence multi-omics data in a spatially resolved manner [162], just like the spatial CITE-seq published recently, which can detect both proteins and whole transcriptome simultaneously. Based on our exploration, one natural extension for DNA methylation maintenance modeling is to build a multiple cell and across cell cycle version, which can helps us understand the spatialtemporal dynamics of the DNA methylation maintenance mechanism. Additionally, we can extend SpaceFlow to deal with spatial multi-omic data (such as spatial epigenome–transcriptome co-profiling [163]) for a more fine grained picture of the spatiotemporal state and dynamics of cells. Lastly, for the spatial CCC analysis study, it might be useful to utilize epigenome–transcriptome co-profiling data to identify the sub-communication state, similar to the metastable state concept in transcriptome or primed state in epigenome. This may refine our understanding on the mechnism of the CCC process.

Appendix A

Supplement to Locally-correlated kinetics of post-replication DNA methylation reveals processivity and region-specificity in DNA methylation maintenance

A.1 Extended Methods

A.1.1 Maximum Likelihood Estimation of Remethylation Rates

The Repli-BS dataset[35] contains read-depths that vary widely both across CpGs and across measured timepoints. The data can be expressed as $\{N_{ij}^0, N_{ij}^1\}$, or the number of unmethylated (0) and methylated (1) reads observed at site i at timepoint j. We assume that the

probability of finding a methylated read at site i and at time t_j is

$$p(1|k_i, f_i, t_j) = f_i(1 - \exp(-k_i t_j)) \tag{A.1}$$

where k_i is the rate of accumulation of methylation post-replication and f_i is the steady-state fraction of cells in the population methylated at site i. Parameters are estimated from the data by Maximum Likelihood Estimation (MLE), as in [37]. In the MLE pipeline, we use a per-site minimum read-depth threshold to increase the baseline accuracy of inferred parameters. In the current work, the threshold was set as follows: each CpG required a minimum of 5 reads (methylated or unmethylated) at both timepoint $t = 0$ and combined timepoints $t > 0$ (for a total minimum of 10 reads over all timepoints). Furthermore, each CpG required a minimum of 5 methylated reads in total (over all measured timepoints). (We performed detailed testing and uncertainty quantification for various methods of setting thresholds previously [37], finding that a read-depth-based threshold gave good balance between genomic coverage and uncertainty in estimates). For each site, we compute the Likelihood Function, $L(\theta \| x_i)$, where θ contains the parameters $\theta = \{k, f\}$ and x_i is the observed data at site i, given the Poisson process model. We observed three cases:

1. The parameters are identifiable. This occurs when the likelihood function contains a global maximum in $\theta = \{k, f\}$; the location of the maximum gives the inferred parameter values. The majority of sites in the dataset fall in this category;

2. The parameters are practically unidentifiable. This occurs when a global maximum exists, but the confidence interval is very wide. We discard sites for which this occurs, requiring that the 75% confidence interval on k (obtained by the profile likelihood method) fall within the range of values $[0.01 - 10]\mathrm{h}^{-1}$;

3. The rate k is unidentifiable, but its value can be bounded. This occurs frequently, as many sites show full methylation already at the first experimental timepoint, indicating a post-replication methylation rate that is faster than the time-resolution of the

experiment, given the one-hour pulse length. We observe this in a likelihood function $L(\theta \| x_i)$ that plateaus at high rate values. In these cases, we use a procedure to estimate a lower bound for k_i: we determine the 25% lower bound of the confidence interval on k, relative to the plateaued value of $L(\theta \| x_i)$. We choose this value for the lower bound on k, reasoning that it estimates the location where the likelihood function is no longer increasing significantly.

A.1.2 Stochastic Simulation

The models are written in the form of stochastic biochemical kinetic reaction models (or in the case of the processive mechanism, a stochastic reaction-diffusion model). The models concentrate solely on the process of maintenance methylation, i.e., the remethylation process that occurs in less than 24 hours, while ignoring other processes such as demethylation. Both models treat DNA as a 1D strand with CpG sites that can be unmethylated (u), hemimethylated (h), or methylated (m). Sites are assumed to be unmethylated or hemimethylated immediately after replication, with hemimethylated sites susceptible to remethylation by the enzyme. The distributive and processive models differ in how the enzyme moves to new hemimethylated sites after catalyzing methylation at one site. The Distributive mechanism treats individual CpG sites as fully independent. The DNMT1 will be disassociated from DNA after it methylated a hemimethylated site. By contrast, in the Processive model, after methylation DNMT1 can diffuse towards its neighbor sites either upstream or downstream. We assume the enzyme travels with 1D diffusion coefficient D and unbinds with rate koff. E denotes a free enzyme (not bound to DNA). Eh is a complex formed when E is bound to h (similar for m). The model reactions are as follows.

Distributive Model

$$\mathrm{E} + \mathrm{h} \rightleftharpoons \mathrm{Eh} \qquad k_{1f}, k_{1r}$$

$$\mathrm{Eh} \longrightarrow \mathrm{E} + \mathrm{m} \qquad k_{cat}$$

After the catalytic step, unbinding is assumed to happen rapidly, therefore the catalytic step and unbinding are assumed to occur together in one step, with rate k_{cat} (i.e., we assume that $k_{off} \ll k_{cat}$).

Processive Model

$$\mathrm{E} + \mathrm{h} \rightleftharpoons \mathrm{Eh} \qquad k_{1f}, k_{1r}$$

$$\mathrm{Eh} \longrightarrow \mathrm{Em} \qquad k_{cat}$$

$$\mathrm{Em}_i + \mathrm{h}_j \longrightarrow \mathrm{m}_i + \mathrm{Eh}_j \qquad D$$

$$\mathrm{Em} \longrightarrow \mathrm{E} + \mathrm{m} \qquad k_{off}$$

In the processive model, the diffusive step allows the enzyme to reach a hemimethylated CpG from a nearby CpG in Em state. The likelihood of this occuring before unbinding with rate k_{off} depends on the properties of diffusion (see below).

Model Parameters

Model parameters were chosen to be approximately in line with experimentally measured values for DNMT1 where possible [73], while also showing reasonable agreement with our experiment-derived methylation rates and rate correlation (in the case of the Processive model). Parameters used in simulations are shown in Table 1.

Simulation algorithms

For the distributive model, the standard Gillespie Stochastic Simulation Algorithm was used [71]. For the processive model, we combine stochastic reactions with a 1D lattice model of diffusion. The lattice spaciging is 1 basepair. We apply a First Passage Time Kinetic Monte Carlo algorithm inspired by [72]. This allows us to simulate diffusive steps efficiently, since we do not need to track each diffusive hop with basepair resolution. Rather, the algorithm requires sampling from the First Passage Time Distribution for an enzyme to reach a nearby CpG by diffusion. This allows us to simulate large stretches (tens of thousands) of CpGs.
In the simulations, the enzyme is allowed to diffuse bidirectionally along DNA. Consider an enzyme with a starting position located at a CpG in the Em state. Let there be two nearest neighbor CpGs in the h state (which are thus candidates to be subsequently targeted by the enzyme, according to the Processive model reactions above). We refer to the diffusion *Domain* as the 1D region between left and right neighbors. Let d_L be the distance (in bp) between the left neighbor and the starting position, and similarly d_R for the right neighbor. Then the next event that can occur for this Em CpG is one of: (1) the enzyme diffuses distance d_L to the left neighbor with diffusion coefficient D. The current CpG undergoes Em $\longrightarrow$ m and the left neighbor undergoes h $\longrightarrow$ Eh; (2) Similar for the right neighbor; or (3) The enzyme unbinds from DNA before reaching either left or right neighbor at rate k_{off}, so Em $\longrightarrow$ E + m. For the simulation, we must compute the waiting time distribution (or First Passage Time Distribution $f(t)$) for the enzyme to exit the domain by one of these events. We must also compute the exit probabilities, i.e. the probability of events (1),(2) or (3) occuring, which we refer to as P_L, P_R, P_{off}, and $P_L + P_R + P_{off} = 1$. The distribution $F(t)$ and the exit probabilities can be solved analytically (see Finite Domain model, below). These events, which involve diffusion, have non-exponential waiting times, and we refer to them as FPT (first passage time) events. The other reactions $(k_{1f}, k_{1r}, k_{cat})$ are assumed to proceed via standard Markovian (memoryless) kinetics, with exponential waiting times, and are referred to as KMC (kinetic Monte Carlo) events. Following [72], a sketch of the

algorithm for the processive simulation is then:

Initialize N CpGs at post-replication time $t_{p.r.} = 0$ in u or h state by sampling from input methylation landscape

while $t_{p.r.} < t_{out}$ (t_{out} is the desired readout time)

Determine which CpG undergoes next event and which reaction occurs at that site by:

* For all CpGs in h or Eh, determine next KMC event and time to event, τ_{KMC}, by the Gillespie algorithm [71].
* *Build new diffusion domains.* For all CpGs newly in Em state, determine domain length $L{=}d_R + d_L$ and $x_0 = d_L$. Sample τ^i_{FPT} from $F(t)$ and store in event queue. Compute corresponding exit probabilities for each ith site, P^i_L, P^i_R, P^i_{off}.
* *Update diffusion domains.* For all CpGs remaining in Em state since last event, determine whether left or right neighbor has changed. If so, rebuild diffusion domain and resample τ^i_{FPT} from $F(t)$ and replace in event queue.
* Select the minimum FPT time, and the corresponding site j for the next FPT event, $\tau_{FPT} = \tau^j_{FPT} = \min_i \tau^i_{FPT}$.

if $\tau_{KMC} < \tau_{FPT}$, set $\tau_{next} = \tau_{KMC}$. Update system state according to KMC event.

else $\tau_{next} = \tau_{FPT}$. Sample from the exit probabilities for the jth site to determine outcome and update system state according to FPT event.

end if

Update the system time: $t_{p.r.} = t_{p.r.} + \tau_{next}$.

Update FPT queue by $\tau^i_{FPT} = \tau^i_{FPT} - \tau_{next}$. (Remove the site which underwent diffusion previously from event queue).

end while

Record system state.

Table A.1: Table of Parameters

Parameter	Symbol	Units	Value	Source or Comment
Catalytic rate, $h \to m$	k_{cat}	hr^{-1}	40	[73]
Michaelis Constant	K_m	μM	0.8	[73]
Stochastic Michaelis Constant	$K_m^{\text{stoch.}}$	#	11500	estimated from K_m, assuming well-mixed reaction volume $\approx 10^{-15}$L
Enzyme unbinding from h-CpG	k_{1r}	hr^{-1}	5	estimated from rate correlation maximum as $d \to 0$ (Processive)
Enzyme binding to h-CpG	k_{1f}	$(\# \cdot hr)^{-1}$	0.039	from $K_m^{\text{stoch.}}$, k_{1r}
Enzymes per h initially	$E_0/h(t=0)$	unitless	1/45	estimated by assuming a median half-time to methylation of $\approx$ 0.5 h (based on Repli-BS data)
Initial hemimethylated substrate #	$S_0 = h(t=0)$	#	≈ 20000	arbitrary, balance of simulation time and region coverage
Enzyme 1-D diffusion coefficient	D	$bp^2 s^{-1}$	10^6	estimate from other DNA-binding proteins [164]
Enzyme drop-off from DNA	k_{off}	hr^{-1}		
Ratio D/k_{off}	D/k_{off}	bp^2	1300	estimated from rate correlation decay (Processive)

A.1.3 Analytic 1D Diffusion Model

Consider two target sites labeled i and j, located on a 1D strand of DNA, that are located some distance d away. Let τ_i be the time (post-replication) at which site i acquires methylation (similarly, τ_j). Assume that in a particular realization of the stochastic process, two possible events can occur:

1. Each of the sites is independently methylated, in which case τ_i and τ_j are independent samples from some common distribution, denoted $\mathcal{T}_{\text{methyl}}$. $\mathcal{T}_{\text{methyl}}$ reflects the distribution of times required for all biochemical steps in the site acquiring methylation, e.g., for DNMT1 to bind DNA from the nucleosol, form a complex with the CpG site, and successfully catalyze methylation.

2. Alternatively, one of the sites could be methylated by an enzyme that has reached it by diffusion after first acting on the other site. In this case, assuming (without loss of generality) that site i is the first, then τ_i is sampled from $\mathcal{T}_{\text{methyl}}$ and $\tau_j = \tau_i + \tau_D$, where τ_D is sampled from random variable $\mathcal{T}_{\text{diff}}$. $\mathcal{T}_{\text{diff}}$ reflects the distribution of times for DNMT1 to diffuse the distance d to reach the neighbor site, form a complex, and successfully catalyze methylation.

The likelihood of event 1 or 2 above taking place depends in part on the probability that the enzyme can remain bound to DNA and diffuse a distance d before unbinding. In the limit of $\sqrt{D/k_{off}} \ll d$, then event 1 always occurs, and

$$Corr(X, Y|d) = 0 \tag{A.2}$$

where X, Y now denotes pairs of methylation times at sites with intervening distance d. The correlation is zero because τ_i and τ_j are assumed to be i.i.d. random variables. Conversely, in the limit of $\sqrt{D/k_{off}} \gg d$, then event 2 always occurs, and

$$Corr(X, Y|d) = 1, \tag{A.3}$$

(in the limit of $\tau_D \ll \tau_i$). Assuming that the enzyme need only *reach* the next target site by diffusion, in which case the catalytic step occurs subsequently with probability 1, then the

correlation is equal to the probability of event 2 occuring, i.e.

$$Corr(X, Y|d) = Prob(\text{enzyme reaches neighbor}) \tag{A.4}$$

$$= \mathrm{e}^{-d\sqrt{k_{off}/D}} \tag{A.5}$$

where the exponential dependence on linear distance, with decay rate $\sqrt{k_{\text{off}}/D}$, is obtained from a 1D analytical diffusion model with parameters k_{off}, D, on a semi-infinite domain with an absorbing state at the target site at distance d from the starting point. Such models have been used previously for theory of transcription factor searching on DNA [52]. In the full Processive model, the enzyme does not always successfully methylate the target upon reaching it, leading to the proportional relationship of Equation 1 (Main Text). The complete analytical results are presented below.

Semi-infinite Domain

Following the model described in the main text, consider two CpG sites, linear distance d apart. Let $P(x,t)$ be the probability density of a particle (enzyme) to be at location x in 1D at time t. For simple diffusion (no unbinding) the governing equation

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2}$$

with initial condition $P(x = 0, t = 0) = 1$ and absorbing boundary condition $P(x = d, t > 0) = 0$ gives the density

$$\tilde{P}(x,t) - \tilde{P}(x - 2d, t)$$

where $\tilde{P}(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-x^2/4Dt\right)$ is the solution to the standard 1D diffusion model on an infinite domain (no absorbing boundary condition).

When unbinding is included in the model, the governing equation becomes:

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2} - k_{off}P(x,t)$$

The solution to this PDE with the same absorbing condition at distance d as above gives

$$P(x,t) = \frac{e^{-k_{off}t}}{\sqrt{4\pi Dt}}\left(e^{-x^2/4Dt} - e^{-(x-2d)^2/4Dt}\right).$$

We then have the survival probability for the protein to remain bound on DNA at time t:

$$S(t) = \int_{-\infty}^{d} P(x,t)dx = e^{-kt}\text{erf}(\frac{d}{\sqrt{4Dt}})$$

leading to an expression for the total probability flux to "exit" the DNA (either by unbinding or by absorption at $x = d$):

$$F(t) = -\frac{\partial S}{\partial t} = e^{-kt}\left(k\text{erf}(\frac{d}{\sqrt{4Dt}}) + \frac{d}{t}\frac{1}{\sqrt{4\pi Dt}}e^{-d^2/4Dt}\right)$$

The first term is the flux to exit by unbinding, the second is the flux to the absorbing state at distance d, i.e. $F(t) = F_{unbind}(t) + F_{absorb}(t)$. Total probability to reach the absorbing state at distance d by diffusion, integrating the absorbing flux,

$$R(d) = \int_0^{\infty} F_{absorb}(t)dt = e^{-d\sqrt{k/D}}$$

Thus, the probability for the enzyme to reach a site at distance d away from the initial site before unbinding is given by $e^{-d\sqrt{k/D}}$.

Finite Domain

In the simulations, we need to consider not just pairs of CpGs distance d apart, but consider that an enzyme bound to DNA may have two neighbors within reach in h state, left and right.

Let the total domain length be $L{=}d_R + d_L$, and the enzyme starting position is $x_0 = d_L$. We have a diffusion model, with unbinding, with initial condition $P(x = x_0, t = 0) = 1$ and absorbing boundary conditions at both ends, $P(x = 0, t > 0) = 0$; $P(x = L, t > 0) = 0$. The Fourier series solution is

$$
\begin{aligned}
P(x,t) &= \sum_{n=1}^{\infty} P_n(x,t) \\
P_n(x,t) &= \frac{2}{L} \sin\frac{n\pi x_0}{L} \sin\frac{n\pi x}{L} \exp\left(-(\frac{Dn^2\pi^2}{L^2} + k)t\right)
\end{aligned}
$$

The survival probability of the enzyme on DNA at location x (i.e. before either unbinding or reaching left or right neighbor) is

$$
\begin{aligned}
S(t) &= \sum_{n=1}^{\infty} S_n(t) \\
S_n(t) &= \int_0^L P_n(x,t)dx \\
&= \frac{2}{n\pi} \sin\frac{n\pi x_0}{L}(1-(-1)^n) \exp\left(-(\frac{Dn^2\pi^2}{L^2} + k)t\right)
\end{aligned}
$$

The first passage time distribution $f(t)$ can be obtained by normalizing the time-dependent probability flux $F(t) = -\frac{\partial S(t)}{\partial t}$. For purposes of the simulation, we sample from $f(t)$ making use of the C.D.F. for $f(t)$. This C.D.F. is given by:

$$C(t) = 1 - S(t).$$

To determine the exit probabilities as a function of time, (i.e., the relative likelihood of exiting left, right, or to solution), we make use of the time-dependent probability fluxes to

each end-state: $F_L(t), F_R(t), F_{solution}(t)$. These are obtained as:

$$F_L(t) = DP(x = 1, t)$$

$$F_R(t) = DP(x = L - 1, t)$$

$$F_{solution}(t) = kS(t).$$

A.1.4 Relationship between Semi-infinite domain model and Processive simulation

Using the analytical semi-infinite domain model together with the arguments in main text, the correlation between methylation times at two sites separated by distance d is given by $e^{-d\sqrt{k/D}}$. In our Processive simulation, there are a number of additional complicating factors that are not present in the idealized theoretical model. For example, after the enzyme reaches the neighbor by diffusion, there is a chance for it to unbind (rate k_{1r}) before successfully catalyzing methylation (rate k_{cat}). There is also a chance that the neighbor site will already be occupied by another copy of the enzyme. Finally, in the simulation τ_D is on the same order as τ_i, not much less as assumed in the idealized model, because it accounts for the fast diffusion time as well as the slower time for complex formation and methylation, introducing delay between subsequent methylation events. All of these factors serve to decrease the correlation at a given distance d, but can be assumed to not significantly affect the distance-dependence, thus predicting that correlation should be proportional, but not strictly equal to, $e^{-d\sqrt{k/D}}$.

Note the distinction between the correlation of methylation times at individual sites (on which our analytical theory is based is based) and methylation rates, which we extract from experiments and simulations. Exact methylation times at individual sites are not currently measureable in genome-wide pulse-chase experiments, but methylation rates k_i are estimated from Repli-BS-Seq experiments, where, for a given CpG sites i, the experiment-inferred

methylation rate constant k_i is by definition taken to be equal to $1/ < \tau_i >$, i.e., the inverse of the average over methylation times at site i across different cells. The inverse relationship between rates and times could in principle affect the linear Pearson correlation, however, from simulations we found that the Pearson correlation function was not sensitive to the distinction, and the rate correlation decay was well described by $e^{-d\sqrt{k_{off}/D}}$, in accordance with the analytical theory (see Figure A.2).

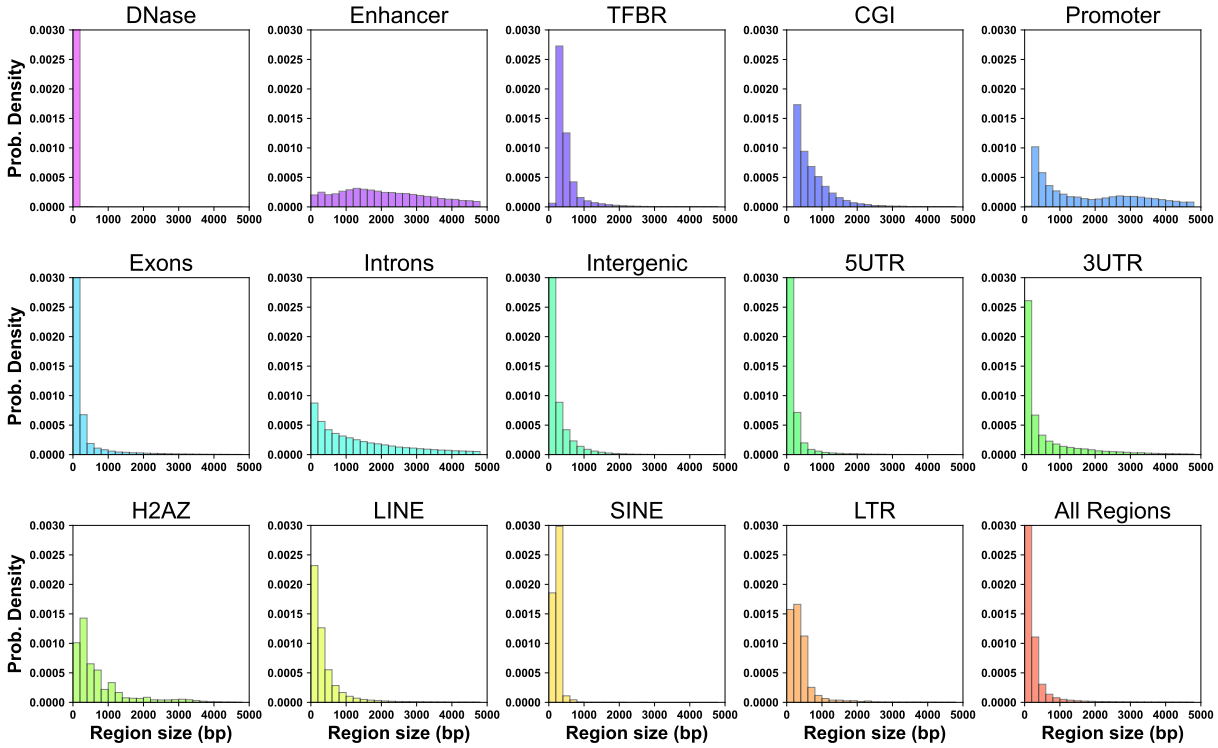

Figure A.1: **Histogram of the sizes of contiguous elements in different genomic regional contexts.** y-axis is the probability density (occurence) and x-axis is the region size in basepairs. The studied regions vary in size.

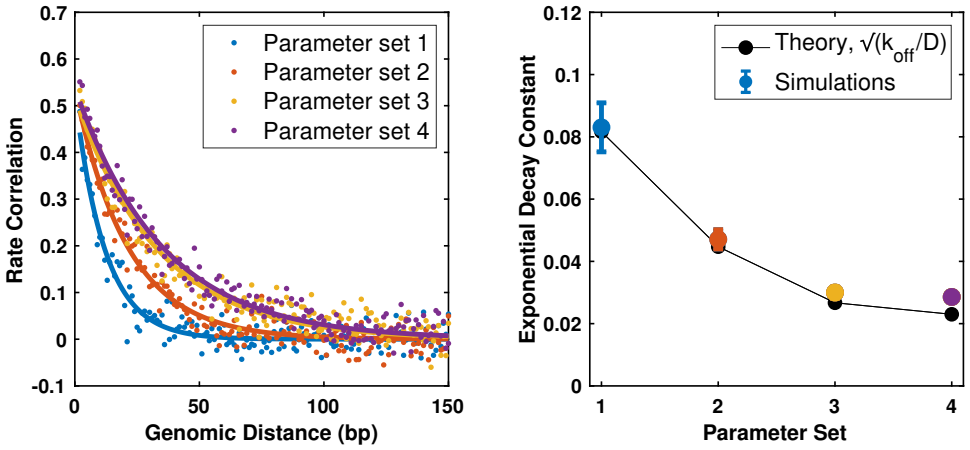

Figure A.2: **Comparison of analytical theory of enzyme processivity to stochastic simulations** (Left) Processive simulations were performed choosing four different values for the parameter ratio D/koff. (Values for sets $\{1, 2, 3, 4\}$ were $\{150, 500, 1402, 1890\}bp^2$. All other parameters were held constant.) After inferring remethylation rates from the synthetic data by the MLE pipeline, ensuing rate correlations were fitted to a single exponential decay function in an unbiased manner. (Right) Fitted decay constants from simulated data (colored dots) overlaying the theoretically predicted decay constant from the model parameters ($\sqrt{koff/D}$), showing close agreement of theory and simulation, despite the increased complexity of the methylation dynamics simulation, as compared to the 1D analytical model.

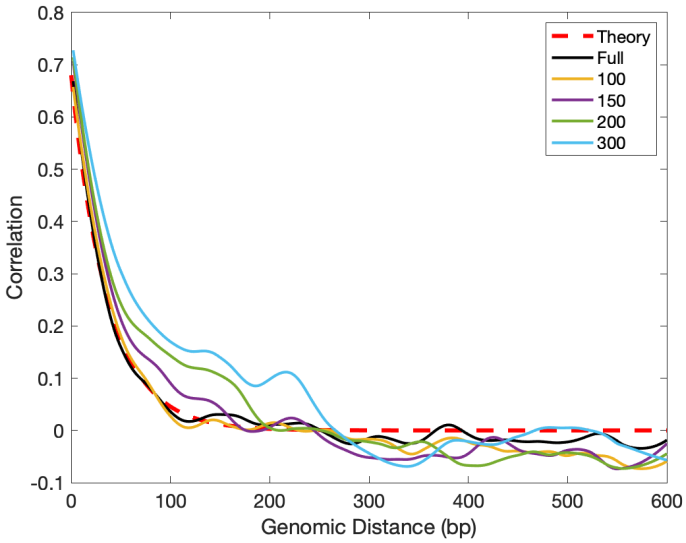

Figure A.3: **Rate Correlations from Simulations of Read Length Effect.** Simulations were performed to study the effect of experimental sequencing read-length on kinetic correlation using read-lengths of 100, 150, 200, or 300 bps. "Full" refers to a read-length that spans the full set of simulated CpGs, corresponding to a read-length of $\approx 10^5$ bps (black curve). Theory (red dashed) is from Equation 1. Read-length simulations were performed as follows: 3000 simulations were performed corresponding to each experimental timepoint. Replicates contained at least 1500 CpGs spanning on the order of 10^5 bps. From these simulated replicates, reads of a given length were randomly sampled from the available sites, i.e., only those sampled CpG states were recorded. The shorter the read-length, the higher the number of independent replicates were sampled, in order that the total per-site read numbers reached the minimum required from the experiments (approximately 5-10 independent reads per timepoint per site). The Repli-BS experiments used 100-bp paired-end sequencing on fragments with an average length of 300 bp. Overall the simulations indicate that, although read-length can introduce spurious correlation, the short read-lengths used in the experiments do not strongly interfere with the measurement of the exponential decay length resulting from enzyme processivity.

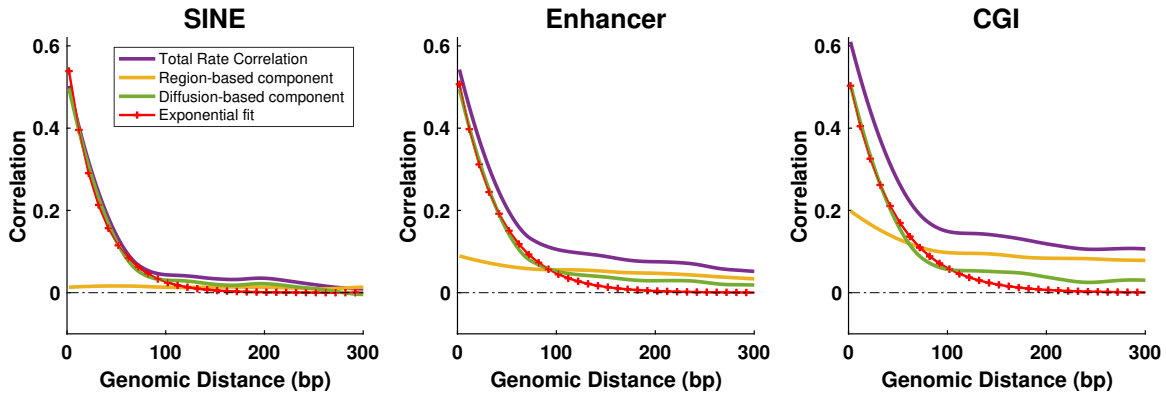

Figure A.4: **Decomposition of experiment-derived rate correlation functions by region into region-based and diffusion-based components.** Simultaneous plotting of correlation components and exponential fits for the three representative regions (same data as Figure 6b in main text, but providing a more detailed view). SINE shows a nearly-zero region-based component (yellow), indicating that the total rate-correlation is already reasonably well-fit by an exponential decay. In contrast, Enhancer and CGI both show small but persistent region-based correlation that is the dominant contributor to the total correlation beyond 100 bp (Enhancer) and beyond 50 bp (CGI). Removal of the region-based component according to the model in Figure 6 results in remaining component θ (green), which is reasonably well fit by a single exponential decay. However, some residual nonzero correlation is still apparent at longer range in the green curves. The exponential fits shown correspond to a 100 bp fitting window to fit the short-range decay while minimizing contribution from residual correlation at longer range.

Appendix B

Supplement to Identifying Multicellular Spatiotemporal Organization of Cells with SpaceFlow

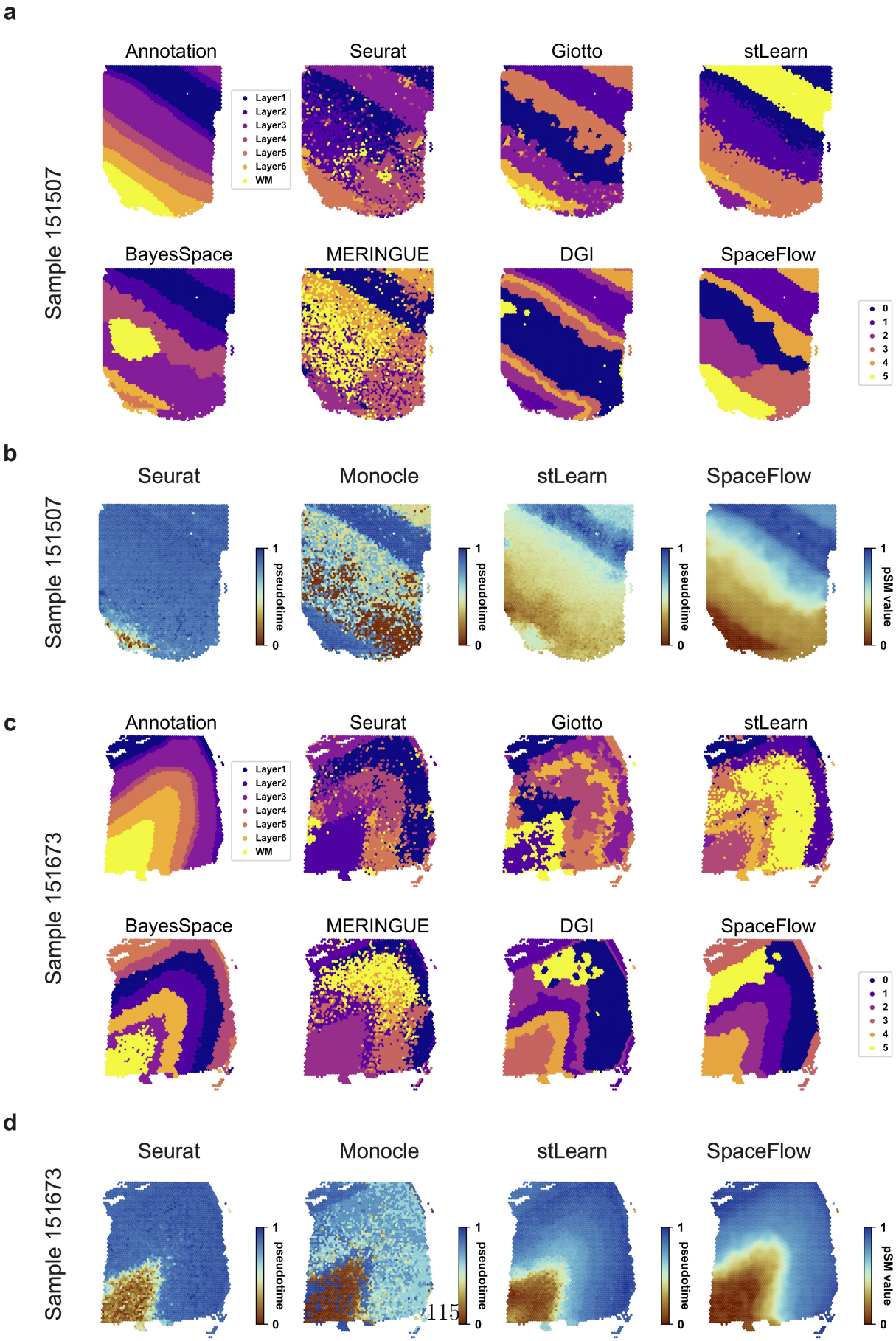
a
Annotation
Seurat
Giotto
stLearn
Sample 151507
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
BayesSpace
MERINGUE
DGI
SpaceFlow
0
1
2
3
4
5
b
Seurat
Monocle
stLearn
SpaceFlow
Sample 151507
1
pseudotime
0
1
pseudotime
0
1
pseudotime
0
1
pSM value
0
c
Annotation
Seurat
Giotto
stLearn
Sample 151673
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
BayesSpace
MERINGUE
DGI
SpaceFlow
0
1
2
3
4
5
d
Seurat
Monocle
stLearn
SpaceFlow
Sample 151673
1
pseudotime
0
1
pseudotime
0
1
pseudotime
0
1
pSM value
0

Figure B.1: **Benchmarking on sections 151507 and 151673 using LIBD human dorsolateral prefrontal cortex (DLPFC) ST data.** a. Domain annotation (top left panel) and segmentations given by different methods (other panels) using section 151507 of DLPFC data. b. The spatial visualization of pseudotimes calculated by Seurat, Monocle, stLearn, and the pseudoSpatiotemporal Map (pSM) generated by SpaceFlow on section 151507 of DLPFC data. c. Domain annotation (top left panel) and segmentations given by different methods (other panels) using section 151673 of DLPFC data. d. The spatial visualization of pseudotimes calculated by Seurat, Monocle, stLearn, and the pSM generated by SpaceFlow on section 151673 of DLPFC data.

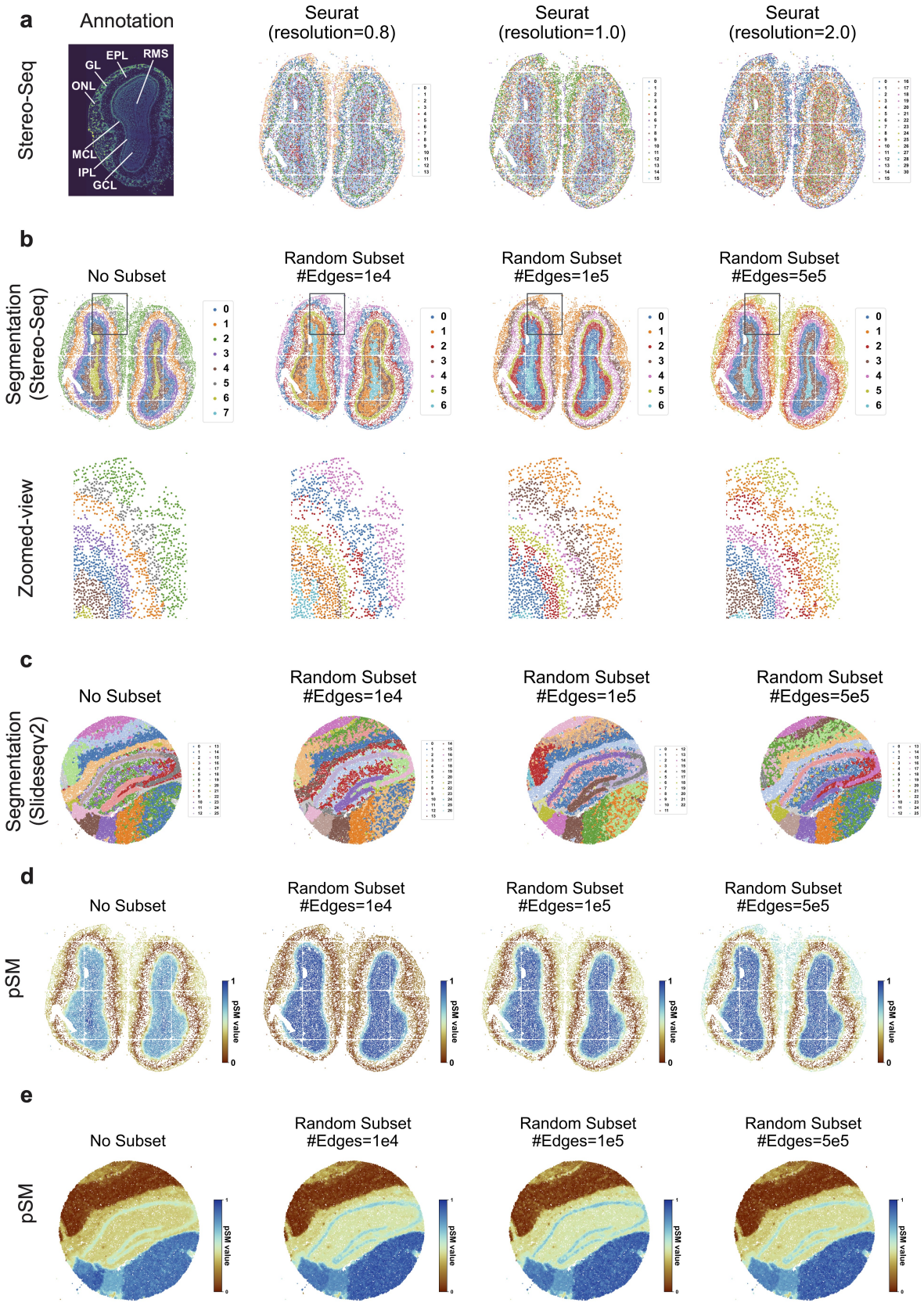
a
Annotation
Seurat (resolution=0.8)
Seurat (resolution=1.0)
Seurat (resolution=2.0)
Stereo-Seq
EPL
RMS
GL
ONL
MCL
IPL
GCL
b
Segmentation (Stereo-Seq)
No Subset
Random Subset #Edges=1e4
Random Subset #Edges=1e5
Random Subset #Edges=5e5
Zoomed-view
c
Segmentation (Slideseqv2)
No Subset
Random Subset #Edges=1e4
Random Subset #Edges=1e5
Random Subset #Edges=5e5
d
pSM
No Subset
Random Subset #Edges=1e4
Random Subset #Edges=1e5
Random Subset #Edges=5e5
pSM value
e
pSM
No Subset
Random Subset #Edges=1e4
Random Subset #Edges=1e5
Random Subset #Edges=5e5
pSM value

Figure B.2: **Domain segmentation and pSM on Stereo-seq data and SlideseqV2 data.** a. Domain annotation (left panel) and segmentations given by Seurat with different resolution settings (right three panels) on Stereo-seq data. b-c. Domain segmentations of (b) Stereo-seq data and (c) Slide-seqv2 data given by SpaceFlow computing regularization over different subsets of cells, showing lack of meaningful variation even when regularization is computed over only a subset. d-e. The pseudo-Spatiotemporal Map (pSM) of (d) Stereo-seq data and (e) Slide-seqv2 data given by SpaceFlow computing regularization over different subsets of cells, again showing qualitatively matching results in each case.

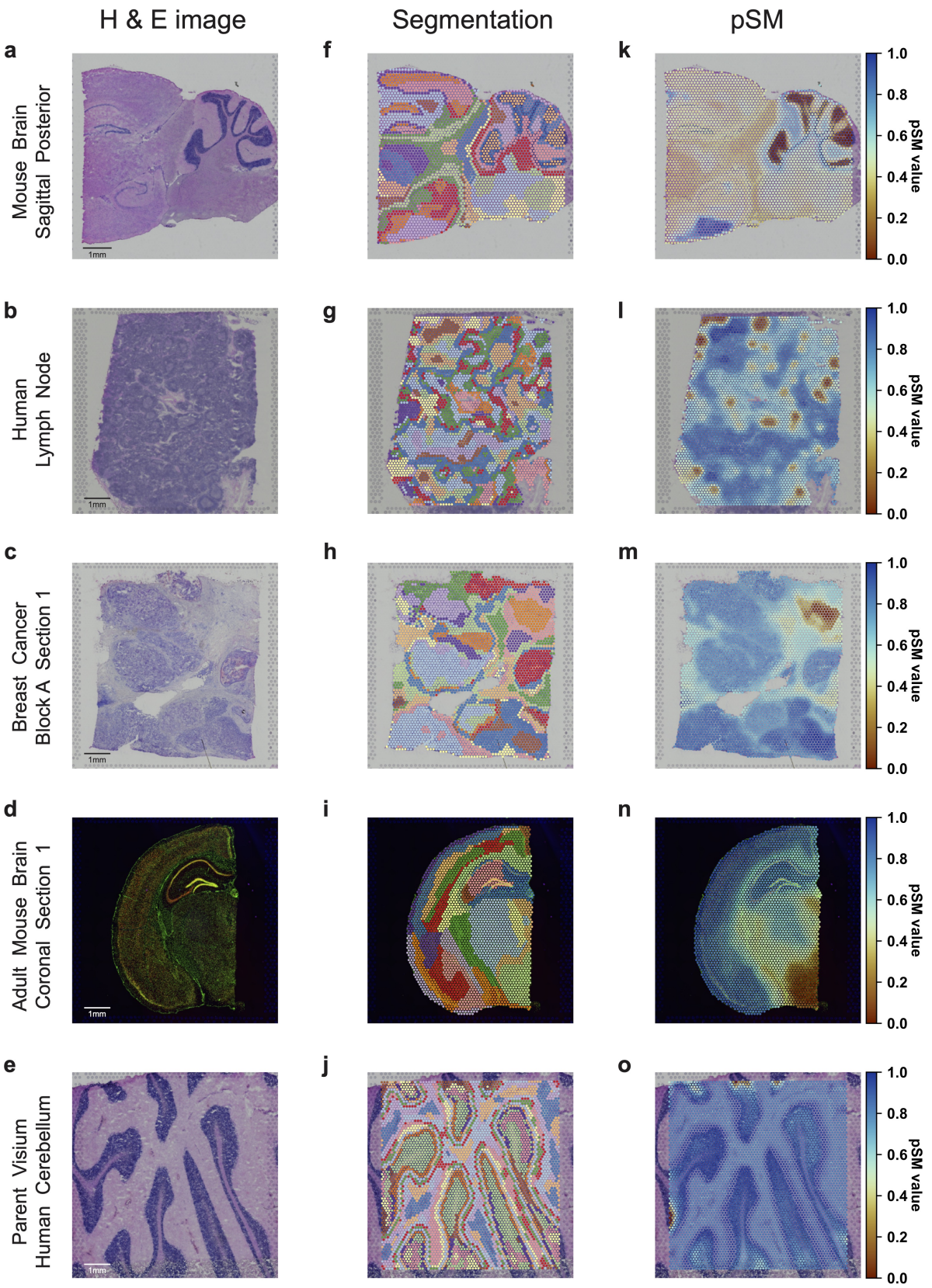
H & E image
Segmentation
pSM
a
Mouse Brain Sagittal Posterior
b
Human Lymph Node
c
Breast Cancer Block A Section 1
d
Adult Mouse Brain Coronal Section 1
e
Parent Visium Human Cerebellum
f
g
h
i
j
k
l
m
n
o
1mm
pSM value
1.0
0.8
0.6
0.4
0.2
0.0

Figure B.3: **Domain segmentation and pSM on five 10x Genomics official Visium datasets.** a-e. H & E images of five Visium datasets (a. V1 Mouse Brain Sagittal Posterior, b. V1 Human Lymph Node, c. V1 Breast Cancer Block A Section 1, d. V1 Adult Mouse Brain Coronal Section 1, e. Parent Visium Human Cerebellum). f-j. Domain segmentations of the corresponding datasets given by SpaceFlow. k-o. pSM of the five datasets given by SpaceFlow.

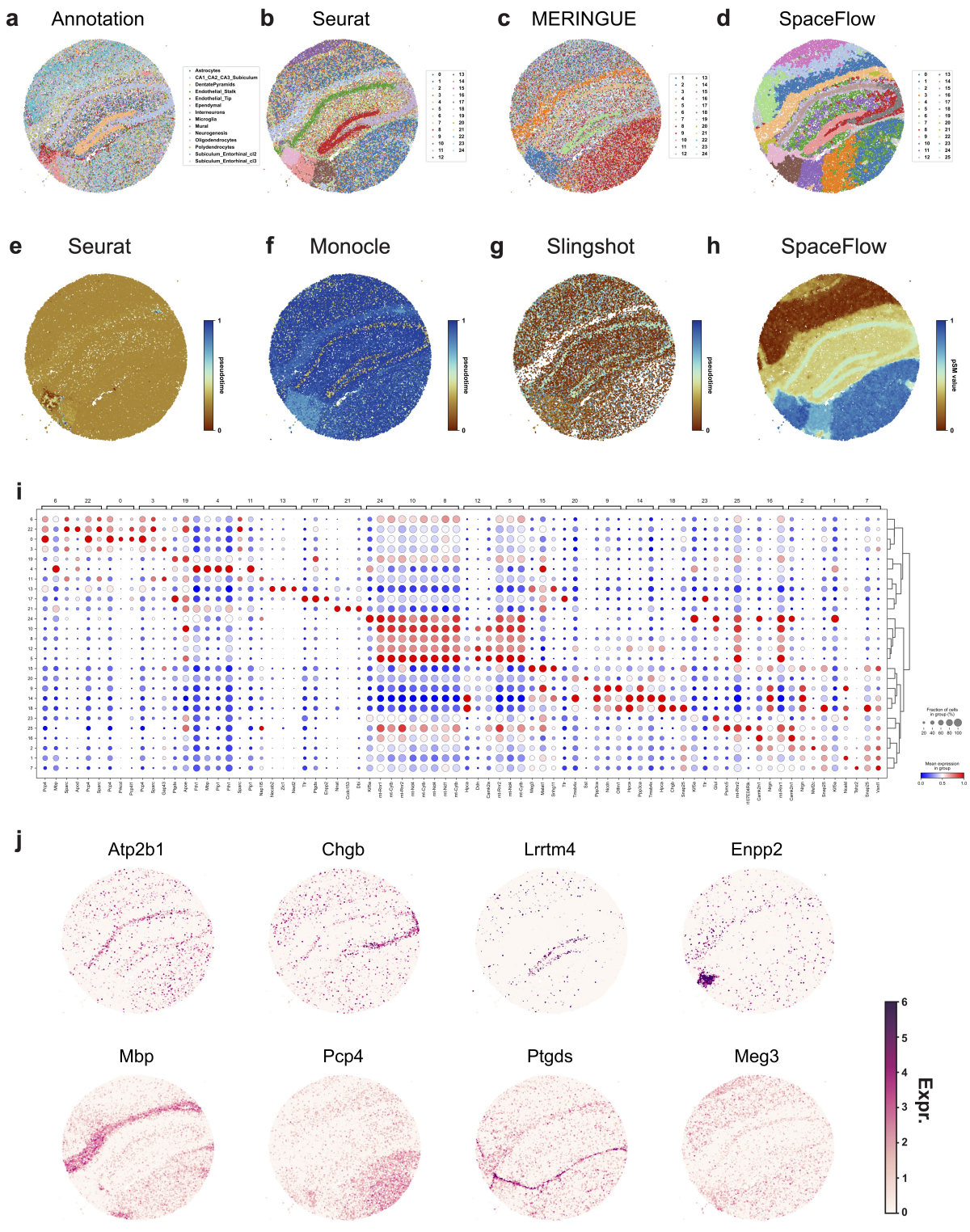
a
Annotation
b
Seurat
c
MERINGUE
d
SpaceFlow
e
Seurat
f
Monocle
g
Slingshot
h
SpaceFlow
i
j
Atp2b1
Chgb
Lrrtm4
Enpp2
Mbp
Pcp4
Ptgds
Meg3
Expr.

Figure B.4: **SpaceFlow analysis on slideseq v2 dataset.** a. Annotation b-d. Domain segmentations produced by Seurat (b), MERINGUE (c) and SpaceFlow (d). e-h. The spatial visualization of pseudotimes calculated by Seurat (e), Monocle (f), Slingshot (g), and the pSM generated by SpaceFlow (h). i. Dot plot of the gene expression of domain-specific markers. The dot size represents the fraction of cells in a domain expressing the marker and the color intensity represents the average expression of the marker in that domain. j. Spatial expression for representative markers of the identified domains.

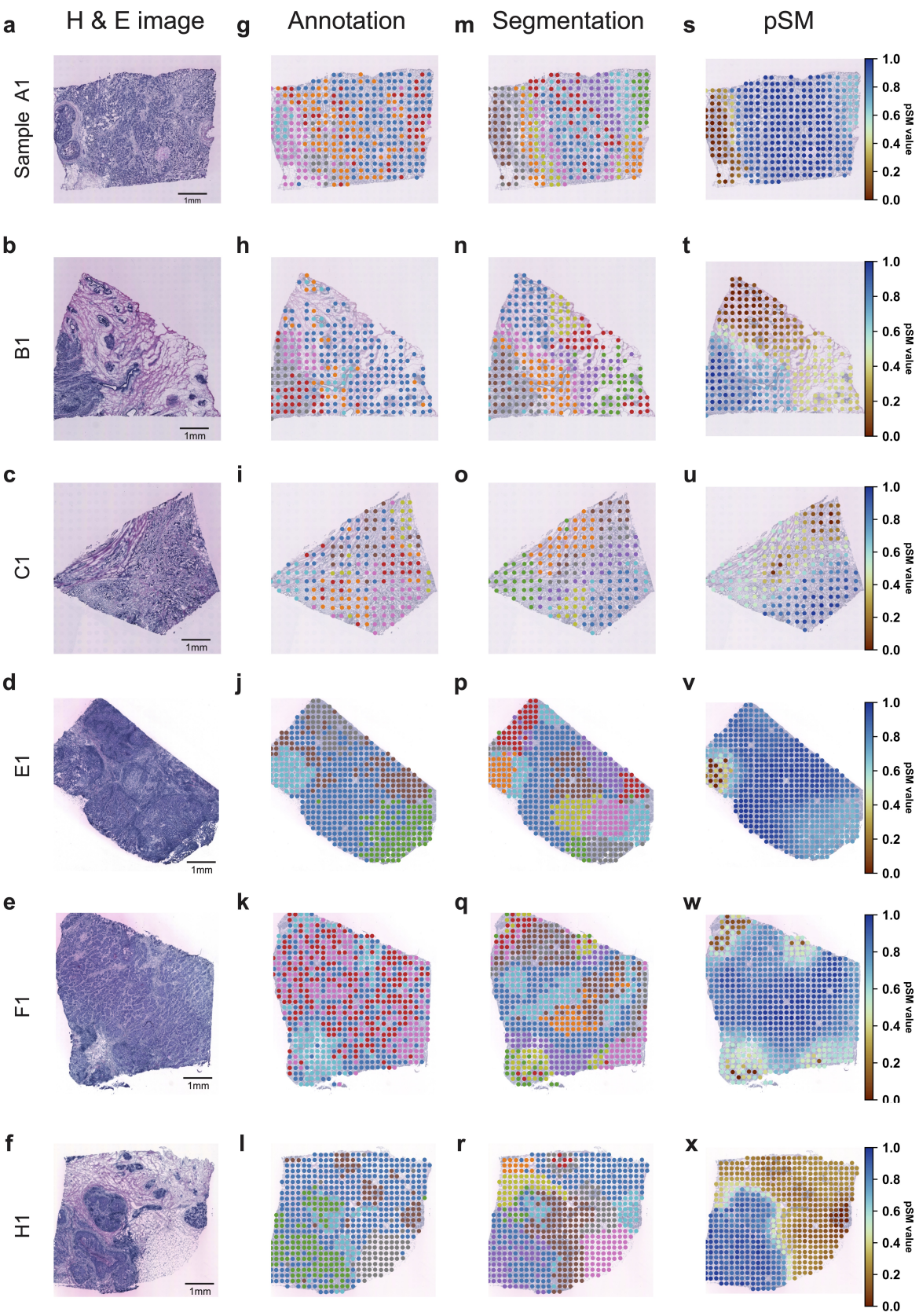
a
H & E image
g
Annotation
m
Segmentation
s
pSM
Sample A1
B1
C1
E1
F1
H1
1mm
pSM value
1.0
0.8
0.6
0.4
0.2
0.0

Figure B.5: **Domain segmentation and pSM on six samples of Human Breast Cancer datasets.** a-f. H & E images of the six samples (Sample A1, B1, C1, E1, F1, H1, respectively). g-l. Annotations of the datasets given by the original paper. m-r. Domain segmentations of the datasets given by SpaceFlow. s-x. pSM of the datasets given by SpaceFlow.

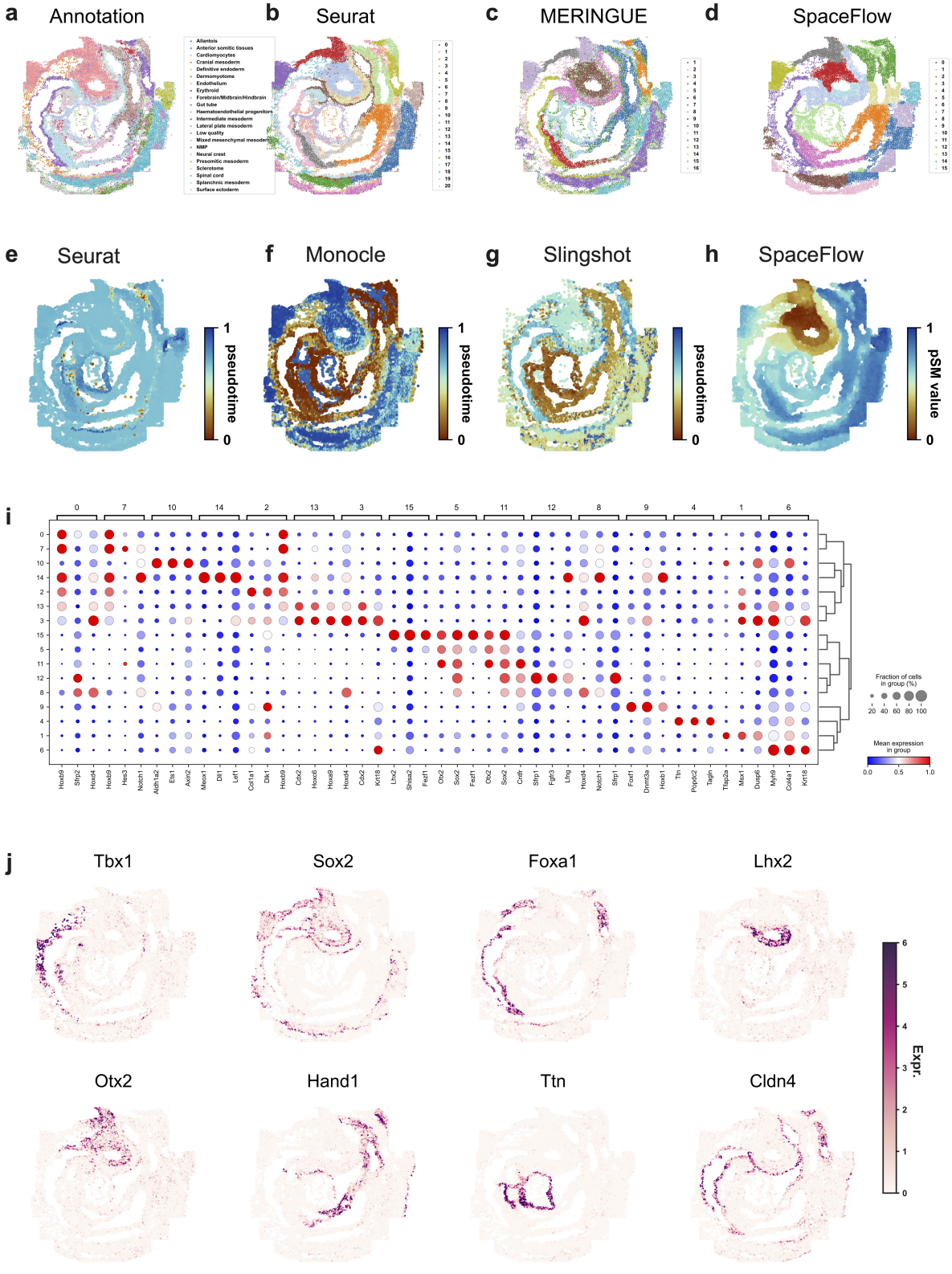
a
Annotation
Allantois
Anterior somitic tissues
Cardiomyocytes
Cranial mesoderm
Definitive endoderm
Dermomyotome
Endothelium
Erythroid
Forebrain/Midbrain/Hindbrain
Gut tube
Haematoendothelial progenitors
Intermediate mesoderm
Lateral plate mesoderm
Low quality
Mixed mesenchymal mesoderm
NMP
Neural crest
Presomitic mesoderm
Sclerotome
Spinal cord
Splanchnic mesoderm
Surface ectoderm
b
Seurat
c
MERINGUE
d
SpaceFlow
e
Seurat
f
Monocle
g
Slingshot
h
SpaceFlow
1
pseudotime
0
1
pseudotime
0
pseudotime
1
pSM value
0
i
0
7
10
14
2
13
3
15
5
11
12
8
9
4
1
6
Hoxb9
Sfrp2
Hoxd4
Hoxb9
Hes3
Notch1
Aldh1a2
Efs1
Axin2
Meox1
Dll1
Lef1
Col1a1
Dlk1
Hoxb9
Cdx2
Hoxc6
Hoxa9
Hoxd4
Cdx2
Krt18
Lhx2
Shisa2
Fezf1
Otx2
Sox2
Fezf1
Otx2
Sox2
Cntfr
Sfrp1
Fgfr3
Lfng
Hoxd4
Notch1
Sfrp1
Foxf1
Dnmt3a
Hoxb1
Ttn
Popdc2
Tagln
Tbx2a
Msx1
Dusp6
Myh9
Col4a1
Krt18
Fraction of cells in group (%)
20
40
60
80
100
Mean expression in group
0.0
0.5
1.0
j
Tbx1
Sox2
Foxa1
Lhx2
Otx2
Hand1
Ttn
Cldn4
Expr.
6
5
4
3
2
1
0

Figure B.6: **SpaceFlow analysis on seqFISH mouse embryogenesis dataset.** a.Annotation b-d. Domain segmentations produced by Seurat (b), MERINGUE (c) and SpaceFlow (d). e-h. The spatial visualization of pseudotimes calculated by Seurat (e), Monocle (f), Slingshot (g), and the pSM generated by SpaceFlow (h). i. Dot plot of the gene expression of domain-specific markers. The dot size represents the fraction of cells in a domain expressing the marker and the color intensity represents the average expression of the marker in that domain. j. Spatial expression for representative markers of the identified domains.

Bibliography

[1] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis, "Deciphering cell–cell interactions and communication from gene expression," *Nature Reviews Genetics*, vol. 22, no. 2, pp. 71–88, 2021.

[2] M. Mantri, G. J. Scuderi, R. Abedini-Nassab, M. F. Z. Wang, D. McKellar, H. Shi, B. Grodner, J. T. Butcher, and I. De Vlaminck, "Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis," *Nat. Commun.*, vol. 12, p. 1771, Mar. 2021.

[3] B. L. Walker, Z. Cang, H. Ren, E. Bourgain-Chang, and Q. Nie, "Deciphering tissue structure and function using spatial transcriptomics," *Communications biology*, vol. 5, no. 1, p. 220, 2022.

[4] S. Feng, S. E. Jacobsen, and W. Reik, "Epigenetic reprogramming in plant and animal development," *Science*, vol. 330, no. 6004, pp. 622–627, 2010.

[5] Z. D. Smith and A. Meissner, "Dna methylation: roles in mammalian development," *Nature Reviews Genetics*, vol. 14, no. 3, pp. 204–220, 2013.

[6] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & development*, vol. 16, no. 1, pp. 6–21, 2002.

[7] A. Jeltsch, "Beyond watson and crick: Dna methylation and molecular enzymology of dna methyltransferases," *Chembiochem*, vol. 3, no. 4, p. 274, 2002.

[8] R. J. Klose and A. P. Bird, "Genomic dna methylation: the mark and its mediators," *Trends in biochemical sciences*, vol. 31, no. 2, pp. 89–97, 2006.

[9] M. Curradi, A. Izzo, G. Badaracco, and N. Landsberger, "Molecular mechanisms of gene silencing mediated by dna methylation," *Molecular and cellular biology*, vol. 22, no. 9, pp. 3157–3173, 2002.

[10] V. Nanavaty, E. W. Abrash, C. Hong, S. Park, E. E. Fink, Z. Li, T. J. Sweet, J. M. Bhasin, S. Singuri, B. H. Lee, *et al.*, "Dna methylation regulates alternative polyadenylation via ctcf and the cohesin complex," *Molecular cell*, vol. 78, no. 4, pp. 752–764, 2020.

[11] P. A. Jones and S. M. Taylor, "Cellular differentiation, cytidine analogs and dna methylation," *Cell*, vol. 20, no. 1, pp. 85–93, 1980.

[12] A. D. Riggs, "X inactivation, differentiation, and dna methylation," *Cytogenetic and Genome Research*, vol. 14, no. 1, pp. 9–25, 1975.

[13] P. A. Jones and G. Liang, "Rethinking how dna methylation patterns are maintained," *Nature Reviews Genetics*, vol. 10, no. 11, pp. 805–811, 2009.

[14] E. Li, T. H. Bestor, and R. Jaenisch, "Targeted mutation of the dna methyltransferase gene results in embryonic lethality," *Cell*, vol. 69, no. 6, pp. 915–926, 1992.

[15] R. Holliday and J. E. Pugh, "Dna modification mechanisms and gene activity during development," *Science*, vol. 187, no. 4173, pp. 226–232, 1975.

[16] T. H. Bestor, "The dna methyltransferases of mammals," *Human molecular genetics*, vol. 9, no. 16, pp. 2395–2402, 2000.

[17] A. Jeltsch and R. Z. Jurkowska, "New concepts in dna methylation," *Trends in biochemical sciences*, vol. 39, no. 7, pp. 310–318, 2014.

[18] M. C. Lorincz, D. Schübeler, S. R. Hutchinson, D. R. Dickerson, and M. Groudine, "Dna methylation density influences the stability of an epigenetic imprint and dnmt3a/b-independent de novo methylation," *Molecular and cellular biology*, vol. 22, no. 21, pp. 7572–7580, 2002.

[19] Y. Zhang, C. Rohde, S. Tierling, T. P. Jurkowski, C. Bock, D. Santacruz, S. Ragozin, R. Reinhardt, M. Groth, J. Walter, *et al.*, "Dna methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution," *PLoS genetics*, vol. 5, no. 3, p. e1000438, 2009.

[20] G. Landan, N. M. Cohen, Z. Mukamel, A. Bar, A. Molchadsky, R. Brosh, S. Horn-Saban, D. A. Zalcenstein, N. Goldfinger, A. Zundelevich, *et al.*, "Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues," *Nature genetics*, vol. 44, no. 11, pp. 1207–1214, 2012.

[21] J. Arand, D. Spieler, T. Karius, M. R. Branco, D. Meilinger, A. Meissner, T. Jenuwein, G. Xu, H. Leonhardt, V. Wolf, *et al.*, "In vivo control of cpg and non-cpg dna methylation by dna methyltransferases," *PLoS genetics*, vol. 8, no. 6, p. e1002750, 2012.

[22] J. O. Haerter, C. Lövkvist, I. B. Dodd, and K. Sneppen, "Collaboration between cpg sites is needed for stable somatic inheritance of dna methylation states," *Nucleic acids research*, vol. 42, no. 4, pp. 2235–2244, 2014.

[23] Y. Mishima, L. Brueckner, S. Takahashi, T. Kawakami, J. Otani, A. Shinohara, K. Takeshita, R. G. Garvilles, M. Watanabe, N. Sakai, *et al.*, "Enhanced processivity of dnmt1 by monoubiquitinated histone h3," *Genes to cells*, vol. 25, no. 1, pp. 22–32, 2020.

[24] X. Ming, Z. Zhang, Z. Zou, C. Lv, Q. Dong, Q. He, Y. Yi, Y. Li, H. Wang, and B. Zhu, "Kinetics and mechanisms of mitotic inheritance of dna methylation and their roles in aging-associated methylome deterioration," *Cell Research*, vol. 30, no. 11, pp. 980–996, 2020.

[25] A. Hermann, R. Goyal, and A. Jeltsch, "The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites," *Journal of Biological Chemistry*, vol. 279, no. 46, pp. 48350–48359, 2004.

[26] G. Vilkaitis, I. Suetake, S. Klimašauskas, and S. Tajima, "Processive methylation of hemimethylated cpg sites by mouse dnmt1 dna methyltransferase," *Journal of Biological Chemistry*, vol. 280, no. 1, pp. 64–72, 2005.

[27] Ž. M. Svedružić and N. O. Reich, "Mechanism of allosteric regulation of dnmt1's processivity," *Biochemistry*, vol. 44, no. 45, pp. 14977–14988, 2005.

[28] M. Bostick, J. K. Kim, P.-O. Estève, A. Clark, S. Pradhan, and S. E. Jacobsen, "Uhrf1 plays a role in maintaining dna methylation in mammalian cells," *Science*, vol. 317, no. 5845, pp. 1760–1764, 2007.

[29] J. Sharif, M. Muto, S.-i. Takebayashi, I. Suetake, A. Iwamatsu, T. A. Endo, J. Shinga, Y. Mizutani-Koseki, T. Toyoda, K. Okamura, *et al.*, "The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna," *Nature*, vol. 450, no. 7171, pp. 908–912, 2007.

[30] L. B. Sontag, M. C. Lorincz, and E. G. Luebeck, "Dynamics, stability and inheritance of somatic dna methylation imprints," *Journal of theoretical biology*, vol. 242, no. 4, pp. 890–899, 2006.

[31] R. Goyal, R. Reinhardt, and A. Jeltsch, "Accuracy of dna methylation pattern preservation by the dnmt1 methyltransferase," *Nucleic acids research*, vol. 34, no. 4, pp. 1182–1188, 2006.

[32] S. C. Zheng, M. Widschwendter, and A. E. Teschendorff, "Epigenetic drift, epigenetic clocks and cancer risk," *Epigenomics*, vol. 8, no. 5, pp. 705–719, 2016.

[33] Y. Song, H. Ren, and J. Lei, "Collaborations between cpg sites in dna methylation," *International Journal of Modern Physics B*, vol. 31, no. 20, p. 1750243, 2017.

[34] L. Zagkos, M. Mc Auley, J. Roberts, and N. I. Kavallaris, "Mathematical models of dna methylation dynamics: Implications for health and ageing," *Journal of theoretical biology*, vol. 462, pp. 184–193, 2019.

[35] J. Charlton, T. L. Downing, Z. D. Smith, H. Gu, K. Clement, R. Pop, V. Akopian, S. Klages, D. P. Santos, A. M. Tsankov, *et al.*, "Global delay in nascent strand dna methylation," *Nature structural & molecular biology*, vol. 25, no. 4, pp. 327–332, 2018.

[36] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, *et al.*, "The nih roadmap epigenomics mapping consortium," *Nature biotechnology*, vol. 28, no. 10, pp. 1045–1048, 2010.

[37] L. Busto-Moner, J. Morival, H. Ren, A. Fahim, Z. Reitz, T. L. Downing, and E. L. Read, "Stochastic modeling reveals kinetic heterogeneity in post-replication dna methylation," *PLOS Computational Biology*, vol. 16, no. 4, p. e1007195, 2020.

[38] S. Guo, D. Diep, N. Plongthongkum, H.-L. Fung, K. Zhang, and K. Zhang, "Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna," *Nature genetics*, vol. 49, no. 4, pp. 635–642, 2017.

[39] W. G. Jacoby, "Loess:: a nonparametric, graphical tool for depicting relationships between variables," *Electoral Studies*, vol. 19, no. 4, pp. 577–613, 2000.

[40] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[41] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos, "Sequencing newly replicated dna reveals widespread plasticity in human replication timing," *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 139–144, 2010.

[42] T. Ryba, D. Battaglia, B. D. Pope, I. Hiratani, and D. M. Gilbert, "Genome-scale analysis of replication timing: from bench to bioinformatics," *Nature protocols*, vol. 6, no. 6, pp. 870–895, 2011.

[43] C. Marchal, T. Sasaki, D. Vera, K. Wilson, J. Sima, J. C. Rivera-Mulia, C. Trevilla-García, C. Nogues, E. Nafie, and D. M. Gilbert, "Genome-wide analysis of replication timing by next-generation sequencing with e/l repli-seq," *Nature protocols*, vol. 13, no. 5, pp. 819–839, 2018.

[44] Q. Du, S. A. Bert, N. J. Armstrong, C. E. Caldon, J. Z. Song, S. S. Nair, C. M. Gould, P.-L. Luu, T. Peters, A. Khoury, *et al.*, "Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer," *Nature communications*, vol. 10, no. 1, pp. 1–15, 2019.

[45] C. Xu and V. G. Corces, "Genome-wide mapping of protein-dna interactions on nascent chromatin," *Methods in molecular biology (Clifton, NJ)*, vol. 1766, p. 231, 2018.

[46] C. Alabert, T. K. Barth, N. Reverón-Gómez, S. Sidoli, A. Schmidt, O. N. Jensen, A. Imhof, and A. Groth, "Two distinct modes for propagation of histone ptms across the cell cycle," *Genes & development*, vol. 29, no. 6, pp. 585–590, 2015.

[47] N. Reverón-Gómez, C. González-Aguilera, K. R. Stewart-Morgan, N. Petryk, V. Flury, S. Graziano, J. V. Johansen, J. S. Jakobsen, C. Alabert, and A. Groth, "Accurate recycling of parental histones reproduces the histone modification landscape during dna replication," *Molecular cell*, vol. 72, no. 2, pp. 239–249, 2018.

[48] P. Vasseur, S. Tonazzini, R. Ziane, A. Camasses, O. J. Rando, and M. Radman-Livaja, "Dynamics of nucleosome positioning maturation following genomic replication," *Cell reports*, vol. 16, no. 10, pp. 2651–2665, 2016.

[49] M. P. Gutiérrez, H. K. MacAlpine, and D. M. MacAlpine, "Nascent chromatin occupancy profiling reveals locus-and factor-specific chromatin maturation dynamics behind the dna replication fork," *Genome research*, vol. 29, no. 7, pp. 1123–1133, 2019.

[50] C. Xu and V. G. Corces, "Nascent dna methylome mapping reveals inheritance of hemimethylation at ctcf/cohesin sites," *Science*, vol. 359, no. 6380, pp. 1166–1170, 2018.

[51] J. Gorman and E. C. Greene, "Visualizing one-dimensional diffusion of proteins along dna," *Nature structural & molecular biology*, vol. 15, no. 8, pp. 768–774, 2008.

[52] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, "How a protein searches for its site on dna: the mechanism of facilitated diffusion," *Journal of Physics A: Mathematical and Theoretical*, vol. 42, no. 43, p. 434013, 2009.

[53] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, "The lac repressor displays facilitated diffusion in living cells," *Science*, vol. 336, no. 6088, pp. 1595–1598, 2012.

[54] K. Kamagata, Y. Itoh, C. Tan, E. Mano, Y. Wu, S. Mandali, S. Takada, and R. C. Johnson, "Testing mechanisms of dna sliding by architectural dna-binding proteins: dynamics of single wild-type and mutant protein molecules in vitro and in vivo," *Nucleic acids research*, vol. 49, no. 15, pp. 8642–8664, 2021.

[55] T. H. Bestor and V. M. Ingram, "Two dna methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with dna," *Proceedings of the National Academy of Sciences*, vol. 80, no. 18, pp. 5559–5563, 1983.

[56] S. Adam, H. Anteneh, M. Hornisch, V. Wagner, J. Lu, N. E. Radde, P. Bashtrykov, J. Song, and A. Jeltsch, "Dna sequence-dependent activity and base flipping mechanisms of dnmt1 regulate genome-wide dna methylation," *Nature communications*, vol. 11, no. 1, pp. 1–15, 2020.

[57] K. Kamagata, E. Mano, K. Ouchi, S. Kanbayashi, and R. C. Johnson, "High free-energy barrier of 1d diffusion along dna by architectural dna-binding proteins," *Journal of molecular biology*, vol. 430, no. 5, pp. 655–667, 2018.

[58] A. B. Kochaniak, S. Habuchi, J. J. Loparo, D. J. Chang, K. A. Cimprich, J. C. Walter, and A. M. van Oijen, "Proliferating cell nuclear antigen uses two distinct modes to move along dna," *Journal of Biological Chemistry*, vol. 284, no. 26, pp. 17700–17710, 2009.

[59] D. M. Suter, "Transcription factors and dna play hide and seek," *Trends in cell biology*, vol. 30, no. 6, pp. 491–500, 2020.

[60] L. S.-H. Chuang, H.-I. Ian, T.-W. Koh, H.-H. Ng, G. Xu, and B. F. Li, "Human dna-(cytosine-5) methyltransferase-pcna complex as a target for p21waf1," *Science*, vol. 277, no. 5334, pp. 1996–2000, 1997.

[61] X. Liu, Q. Gao, P. Li, Q. Zhao, J. Zhang, J. Li, H. Koseki, and J. Wong, "Uhrf1 targets dnmt1 for dna methylation through cooperative binding of hemi-methylated dna and methylated h3k9," *Nature communications*, vol. 4, no. 1, pp. 1–13, 2013.

[62] R. M. Vaughan, B. M. Dickson, M. F. Whelihan, A. L. Johnstone, E. M. Cornett, M. A. Cheek, C. A. Ausherman, M. W. Cowles, Z.-W. Sun, and S. B. Rothbart, "Chromatin structure and its chemical modifications regulate the ubiquitin ligase substrate selectivity of uhrf1," *Proceedings of the National Academy of Sciences*, vol. 115, no. 35, pp. 8775–8780, 2018.

[63] P. A. Ginno, D. Gaidatzis, A. Feldmann, L. Hoerner, D. Imanci, L. Burger, F. Zilbermann, A. H. Peters, F. Edenhofer, S. A. Smallwood, *et al.*, "A genome-scale map of dna methylation turnover identifies site-specific dependencies of dnmt and tet activity," *Nature communications*, vol. 11, no. 1, pp. 1–16, 2020.

[64] S. P. Otto and V. Walbot, "Dna methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle.," *Genetics*, vol. 124, no. 2, pp. 429–437, 1990.

[65] A. A. Johnson, K. Akman, S. R. Calimport, D. Wuttke, A. Stolzing, and J. P. De Magalhaes, "The role of dna methylation in aging, rejuvenation, and age-related disease," *Rejuvenation research*, vol. 15, no. 5, pp. 483–494, 2012.

[66] A. Nishiyama and M. Nakanishi, "Navigating the dna methylation landscape of cancer," *Trends in Genetics*, 2021.

[67] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, *et al.*, "Human dna methylomes at base resolution show widespread epigenomic differences," *nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[68] Z. D. Smith, M. M. Chan, K. C. Humm, R. Karnik, S. Mekhoubad, A. Regev, K. Eggan, and A. Meissner, "Dna methylation dynamics of the human preimplantation embryo," *Nature*, vol. 511, no. 7511, pp. 611–615, 2014.

[69] J. Liao, R. Karnik, H. Gu, M. J. Ziller, K. Clement, A. M. Tsankov, V. Akopian, C. A. Gifford, J. Donaghey, C. Galonska, *et al.*, "Targeted disruption of dnmt1, dnmt3a and dnmt3b in human embryonic stem cells," *Nature genetics*, vol. 47, no. 5, pp. 469–478, 2015.

[70] I. Akerman, B. Kasaai, A. Bazarova, P. B. Sang, I. Peiffer, M. Artufel, R. Derelle, G. Smith, M. Rodriguez-Martinez, M. Romano, *et al.*, "A predictable conserved dna base composition signature defines human core dna replication origins," *Nature communications*, vol. 11, no. 1, pp. 1–15, 2020.

[71] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The journal of physical chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[72] A. Bezzola, B. B. Bales, R. C. Alkire, and L. R. Petzold, "An exact and efficient first passage time algorithm for reaction–diffusion processes on a 2d-lattice," *Journal of Computational Physics*, vol. 256, pp. 183–197, 2014.

[73] I. Hemeon, J. A. Gutierrez, M.-C. Ho, and V. L. Schramm, "Characterizing dna methyltransferases with an ultrasensitive luciferase-linked continuous assay," *Analytical chemistry*, vol. 83, no. 12, pp. 4996–5004, 2011.

[74] M. Asp, J. Bergenstråhle, and J. Lundeberg, "Spatially resolved transcriptomes-next generation tools for tissue exploration," *Bioessays*, vol. 42, p. e1900221, Oct. 2020.

[75] A. Rao, D. Barkley, G. S. França, and I. Yanai, "Exploring tissue architecture using spatial transcriptomics," *Nature*, vol. 596, pp. 211–220, Aug. 2021.

[76] G. Palla, D. S. Fischer, A. Regev, and F. J. Theis, "Spatial components of molecular tissue biology," *Nat. Biotechnol.*, pp. 1–11, Feb. 2022.

[77] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang, "RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells," *Science*, vol. 348, p. aaa6090, Apr. 2015.

[78] S. Shah, E. Lubeck, W. Zhou, and L. Cai, "In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus," *Neuron*, vol. 92, pp. 342–357, Oct. 2016.

[79] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko, "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution," *Science*, vol. 363, pp. 1463–1467, Mar. 2019.

[80] A. Chen, S. Liao, M. Cheng, K. Ma, L. Wu, Y. Lai, X. Qiu, J. Yang, J. Xu, S. Hao, *et al.*, "Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays," *Cell*, vol. 185, no. 10, pp. 1777–1792, 2022.

[81] J. W. Bohland, H. Bokil, S. D. Pathak, C.-K. Lee, L. Ng, C. Lau, C. Kuan, M. Hawrylycz, and P. P. Mitra, "Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy," *Methods*, vol. 50, pp. 105–112, Feb. 2010.

[82] S. M. H. Huisman, B. van Lew, A. Mahfouz, N. Pezzotti, T. Höllt, L. Michielsen, A. Vilanova, M. J. T. Reinders, and B. P. F. Lelieveldt, "BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome," *Nucleic Acids Res.*, p. gkx046, Jan. 2017.

[83] V. Svensson, S. A. Teichmann, and O. Stegle, "SpatialDE: identification of spatially variable genes," *Nat. Methods*, vol. 15, pp. 343–346, May 2018.

[84] D. Edsgärd, P. Johnsson, and R. Sandberg, "Identification of spatial expression trends in single-cell gene expression data," *Nat. Methods*, vol. 15, pp. 339–342, May 2018.

[85] B. L. Walker, Z. Cang, H. Ren, E. Bourgain-Chang, and Q. Nie, "Deciphering tissue structure and function using spatial transcriptomics," *Commun. Biol.*, vol. 5, pp. 1–10, Dec. 2022.

[86] R. Dries, Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, R. E. George, N. Pierson, L. Cai, and G.-C. Yuan, "Giotto: a toolbox for integrative analysis and visualization of spatial expression data," *Genome Biol.*, vol. 22, p. 78, Mar. 2021.

[87] E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. B. Taylor, P. Nghiem, J. H. Bielas, and R. Gottardo, "Spatial transcriptomics at subspot resolution with BayesSpace," *Nat. Biotechnol.*, vol. 39, pp. 1375–1384, June 2021.

[88] Y. Yang, X. Shi, W. Liu, Q. Zhou, M. Chan Lau, J. Chun Tatt Lim, L. Sun, C. C. Y. Ng, J. Yeong, and J. Liu, "SC-MEB: spatial clustering with hidden markov random field using empirical bayes," *Brief. Bioinform.*, vol. 23, Jan. 2022.

[89] D. Pham, X. Tan, J. Xu, L. F. Grice, P. Y. Lam, A. Raghubar, J. Vukovic, M. J. Ruitenberg, and Q. Nguyen, "stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues." May 2020.

[90] J. Moehlin, B. Mollet, B. M. Colombo, and M. A. Mendoza-Parra, "Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer," *Cell Syst.*, vol. 12, pp. 694–705.e3, July 2021.

[91] B. F. Miller, D. Bambah-Mukku, C. Dulac, X. Zhuang, and J. Fan, "Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities," *Genome Res.*, vol. 31, pp. 1843–1855, Oct. 2021.

[92] J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li, "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network," *Nat. Methods*, vol. 18, pp. 1342–1351, Oct. 2021.

[93] H. Fu, H. Xu, K. Chong, M. Li, K. S. Ang, H. K. Lee, J. Ling, A. Chen, L. Shao, L. Liu, and J. Chen, "Unsupervised spatially embedded deep representation of spatial transcriptomics." July 2021.

[94] Z. Cang, X. Ning, A. Nie, M. Xu, and J. Zhang, "Scan-IT: Domain segmentation of spatial transcriptomics images by graph neural network," in *British Machine Vision Conference*, 2021.

[95] C. Zuo, Y. Zhang, C. Cao, J. Feng, M. Jiao, and L. Chen, "Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning." Feb. 2022.

[96] K. Dong and S. Zhang, "Deciphering spatial domains from spatially resolved transcriptomics with adaptive graph attention auto-encoder." Aug. 2021.

[97] A. Misra, C. D. Baker, E. M. Pritchett, K. N. Burgos Villar, J. M. Ashton, and E. M. Small, "Characterizing neonatal heart maturation, regeneration, and scar resolution using spatial transcriptomics," *J. Cardiovasc. Dev. Dis.*, vol. 9, p. 1, Dec. 2021.

[98] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nat. Methods*, vol. 13, pp. 845–848, Oct. 2016.

[99] V. van Unen, T. Höllt, N. Pezzotti, N. Li, M. J. T. Reinders, E. Eisemann, F. Koning, A. Vilanova, and B. P. F. Lelieveldt, "Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types," *Nat. Commun.*, vol. 8, p. 1740, Nov. 2017.

[100] W. E. Marcílio-Jr, D. M. Eler, F. V. Paulovich, and R. M. Martins, "HUMAP: Hierarchical uniform manifold approximation and projection," June 2021.

[101] N. Pezzotti, J.-D. Fekete, T. Höllt, B. P. F. Lelieveldt, E. Eisemann, and A. Vilanova, "Multiscale visualization and exploration of large bipartite graphs," *Comput. Graph. Forum*, vol. 37, pp. 549–560, June 2018.

[102] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, (New York, New York, USA), ACM Press, 2014.

[103] T. N. Kipf and M. Welling, "Variational graph Auto-Encoders," Nov. 2016.

[104] A. Bojchevski and S. Günnemann, "Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking," July 2017.

[105] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," Sept. 2018.

[106] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, 3rd, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, pp. 3573–3587.e29, June 2021.

[107] K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, 2nd, M. N. Tran, Z. Besich, M. Tippani, J. Chew, Y. Yin, J. E. Kleinman, T. M. Hyde, N. Rao, S. C. Hicks, K. Martinowich, and A. E. Jaffe, "Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex," *Nat. Neurosci.*, vol. 24, pp. 425–436, Mar. 2021.

[108] A. Watakabe, J. Hirokawa, N. Ichinohe, S. Ohsawa, T. Kaneko, K. S. Rockland, and T. Yamamori, "Area-specific substratification of deep layer neurons in the rat cortex," *J. Comp. Neurol.*, vol. 520, pp. 3553–3573, Nov. 2012.

[109] D. Bernard, K. V. Prasanth, V. Tripathi, S. Colasse, T. Nakamura, Z. Xuan, M. Q. Zhang, F. Sedel, L. Jourdren, F. Coulpier, A. Triller, D. L. Spector, and A. Bessis, "A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression," *EMBO J.*, vol. 29, pp. 3082–3093, Sept. 2010.

[110] J. Lisman, R. Yasuda, and S. Raghavachari, "Mechanisms of CaMKII action in long-term potentiation," *Nat. Rev. Neurosci.*, vol. 13, pp. 169–182, Feb. 2012.

[111] J. Li, B. Wilkinson, V. A. Clementel, J. Hou, T. J. O'Dell, and M. P. Coba, "Long-term potentiation modulates synaptic phosphorylation networks and reshapes the structure of the postsynaptic interactome," *Sci. Signal.*, vol. 9, p. rs8, Aug. 2016.

[112] E. Martín-López, R. Corona, and L. López-Mascaraque, "Postnatal characterization of cells in the accessory olfactory bulb of wild type and reeler mice," *Front. Neuroanat.*, vol. 6, p. 15, May 2012.

[113] Y. Xiang, J. Xin, W. Le, and Y. Yang, "Neurogranin: A potential biomarker of neurological and mental diseases," *Front. Aging Neurosci.*, vol. 12, p. 584743, Oct. 2020.

[114] B. J. Martinsen, "Reference guide to the stages of chick heart embryology," *Dev. Dyn.*, vol. 233, pp. 1217–1237, Aug. 2005.

[115] R. N. Wang, J. Green, Z. Wang, Y. Deng, M. Qiao, M. Peabody, Q. Zhang, J. Ye, Z. Yan, S. Denduluri, O. Idowu, M. Li, C. Shen, A. Hu, R. C. Haydon, R. Kang, J. Mok, M. J. Lee, H. L. Luu, and L. L. Shi, "Bone morphogenetic protein (BMP) signaling in development and human diseases," *Genes Dis.*, vol. 1, pp. 87–105, Sept. 2014.

[116] C. M. Alfieri, J. Cheek, S. Chakraborty, and K. E. Yutzey, "Wnt signaling in heart valve development and osteogenic gene induction," *Dev. Biol.*, vol. 338, pp. 127–135, Feb. 2010.

[117] C. F. Suoqin, L. Guerrero-Juarez, I. Zhang, R. Chang, C.-H. Ramos, P. Kuan, V. Maksim, and Q. Plikus, "Inference and analysis of Cell-Cell communication using CellChat," *Nature Communications*, vol. 12, no. 1, pp. 1–20, 2021.

[118] Q. Zhou, H. Cao, X. Hang, H. Liang, M. Zhu, Y. Fan, J. Shi, N. Dong, and X. He, "Midkine prevents calcification of aortic valve interstitial cells via intercellular crosstalk," *Front. Cell Dev. Biol.*, vol. 9, p. 794058, Dec. 2021.

[119] C. B. Arrington and H. J. Yost, "Extra-embryonic syndecan 2 regulates organ primordia migration and fibrillogenesis throughout the zebrafish embryo," *Development*, vol. 136, pp. 3143–3152, Sept. 2009.

[120] J. Li, H. Wei, A. Chesley, C. Moon, M. Krawczyk, M. Volkova, B. Ziman, K. B. Margulies, M. Talan, M. T. Crow, and K. R. Boheler, "The pro-angiogenic cytokine pleiotrophin potentiates cardiomyocyte apoptosis through inhibition of endogenous AKT/PKB activity," *J. Biol. Chem.*, vol. 282, pp. 34984–34993, Nov. 2007.

[121] A. Andersson, L. Larsson, L. Stenbeck, F. Salmén, A. Ehinger, S. Z. Wu, G. Al-Eryani, D. Roden, A. Swarbrick, Å. Borg, J. Frisén, C. Engblom, and J. Lundeberg, "Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions," *Nat. Commun.*, vol. 12, pp. 1–14, Oct. 2021.

[122] G. A. Cabral-Pacheco, I. Garza-Veloz, C. Castruita-De la Rosa, J. M. Ramirez-Acuña, B. A. Perez-Romero, J. F. Guerrero-Rodriguez, N. Martinez-Avila, and M. L. Martinez-Fierro, "The roles of matrix metalloproteinases and their inhibitors in human diseases," *Int. J. Mol. Sci.*, vol. 21, p. 9739, Dec. 2020.

[123] H. Wang, X. Qiu, S. Lin, X. Chen, T. Wang, and T. Liao, "Knockdown of IFI27 inhibits cell proliferation and invasion in oral squamous cell carcinoma," *World J. Surg. Oncol.*, vol. 16, p. 64, Mar. 2018.

[124] H.-C. Ma, T.-W. Lin, H. Li, S. M. M. Iguchi-Ariga, H. Ariga, Y.-L. Chuang, J.-H. Ou, and S.-Y. Lo, "Hepatitis C virus ARFP/F protein interacts with cellular MM-1 protein and enhances the gene trans-activation activity of c-myc," *J. Biomed. Sci.*, vol. 15, pp. 417–425, July 2008.

[125] H. M. Koh, B. G. Jang, and D. C. Kim, "Prognostic value of CD63 expression in solid tumors: A meta-analysis of the literature," *In Vivo*, vol. 34, pp. 2209–2215, Sept. 2020.

[126] M. Nitzan, N. Karaiskos, N. Friedman, and N. Rajewsky, "Gene expression cartography," *Nature*, vol. 576, pp. 132–137, Dec. 2019.

[127] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nat. Biotechnol.*, vol. 32, pp. 381–386, Apr. 2014.

[128] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, Dec. 2018.

[129] N. Pezzotti, T. Höllt, B. Lelieveldt, E. Eisemann, and A. Vilanova, "Hierarchical stochastic neighbor embedding," *Comput. Graph. Forum*, vol. 35, pp. 21–30, June 2016.

[130] Y. Deng, M. Bartosovic, P. Kukanja, D. Zhang, Y. Liu, G. Su, A. Enninful, Z. Bai, G. Castelo-Branco, and R. Fan, "Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level," *Science*, vol. 375, pp. 681–686, Feb. 2022.

[131] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity of single cells," *Nature*, vol. 560, pp. 494–498, Aug. 2018.

[132] C. J. A. Delfinado and H. Edelsbrunner, "An incremental algorithm for betti numbers of simplicial complexes on the 3-sphere," *Comput. Aided Geom. Des.*, vol. 12, pp. 771–784, Nov. 1995.

[133] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," Oct. 2020.

[134] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, p. 5233, Mar. 2019.

[135] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014.

[136] G. Palla, H. Spitzer, M. Klein, D. S. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg, I. Virshup, M. Lotfollahi, S. Richter, and F. J. Theis, "Squidpy: a scalable framework for spatial single cell analysis." Feb. 2021.

[137] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, *et al.*, "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics," *Science*, vol. 353, no. 6294, pp. 78–82, 2016.

[138] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko, "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution," *Science*, vol. 363, no. 6434, pp. 1463–1467, 2019.

[139] R. R. Stickels, E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, and F. Chen, "Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2," *Nature biotechnology*, vol. 39, no. 3, pp. 313–319, 2021.

[140] A. Rao, D. Barkley, G. S. França, and I. Yanai, "Exploring tissue architecture using spatial transcriptomics," *Nature*, vol. 596, no. 7871, pp. 211–220, 2021.

[141] J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, *et al.*, "Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region," *Science*, vol. 362, no. 6416, p. eaau5324, 2018.

[142] C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, *et al.*, "Transcriptome-scale super-resolved imaging in tissues by rna seqfish+," *Nature*, vol. 568, no. 7751, pp. 235–239, 2019.

[143] A. R. Cillo, C. H. Kürten, T. Tabib, Z. Qi, S. Onkar, T. Wang, A. Liu, U. Duvvuri, S. Kim, R. J. Soose, *et al.*, "Immune landscape of viral-and carcinogen-driven head and neck cancer," *Immunity*, vol. 52, no. 1, pp. 183–199, 2020.

[144] Y. Wang, R. Wang, S. Zhang, S. Song, C. Jiang, G. Han, M. Wang, J. Ajani, A. Futreal, and L. Wang, "italk: an r package to characterize and illustrate intercellular communication," *BioRxiv*, p. 507871, 2019.

[145] S. R. Tyler, P. G. Rotti, X. Sun, Y. Yi, W. Xie, M. C. Winter, M. J. Flamme-Wiese, B. A. Tucker, R. F. Mullins, A. W. Norris, *et al.*, "Pyminer finds gene and autocrine-paracrine networks from human islet scrna-seq," *Cell reports*, vol. 26, no. 7, pp. 1951–1964, 2019.

[146] M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo, "Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes," *Nature protocols*, vol. 15, no. 4, pp. 1484–1506, 2020.

[147] R. Dries, Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, *et al.*, "Giotto: a toolbox for integrative analysis and visualization of spatial expression data," *Genome biology*, vol. 22, pp. 1–31, 2021.

[148] W. Zhao, K. G. Johnston, H. Ren, X. Xu, and Q. Nie, "Inferring neuron-neuron communications from single-cell transcriptomics through neuronchat," *bioRxiv*, pp. 2023–01, 2023.

[149] S. Wang, M. Karikomi, A. L. MacLean, and Q. Nie, "Cell lineage and communication network inference via optimization for single-cell transcriptomics," *Nucleic acids research*, vol. 47, no. 11, pp. e66–e66, 2019.

[150] R. Browaeys, W. Saelens, and Y. Saeys, "Nichenet: modeling intercellular communication by linking ligands to target genes," *Nature methods*, vol. 17, no. 2, pp. 159–162, 2020.

[151] Z. Cang and Q. Nie, "Inferring spatial and signaling relationships between cells from single cell transcriptomic data," *Nature communications*, vol. 11, no. 1, p. 2084, 2020.

[152] X. Shao, C. Li, H. Yang, X. Lu, J. Liao, J. Qian, K. Wang, J. Cheng, P. Yang, H. Chen, *et al.*, "Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with spatalk," *Nature Communications*, vol. 13, no. 1, p. 4429, 2022.

[153] J. Chen, L. Yan, Q. Nie, and X. Sun, "Modeling spatial intercellular communication and multilayer signaling regulations using stmlnet," *bioRxiv*, pp. 2022–06, 2022.

[154] Z. Cang, Y. Zhao, A. A. Almet, A. Stabell, R. Ramos, M. V. Plikus, S. X. Atwood, and Q. Nie, "Screening cell–cell communication in spatial transcriptomics via collective optimal transport," *Nature Methods*, vol. 20, no. 2, pp. 218–228, 2023.

[155] H. Ren, B. L. Walker, Z. Cang, and Q. Nie, "Identifying multicellular spatiotemporal organization of cells with spaceflow," *Nature communications*, vol. 13, no. 1, p. 4076, 2022.

[156] Y. Zhou, R. Jin, and S. C.-H. Hoi, "Exclusive lasso for multi-task feature selection," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 988–995, JMLR Workshop and Conference Proceedings, 2010.

[157] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[158] O. Delapeyrière, V. Ollendorff, J. Planche, M. O. Ott, S. Pizette, F. Coulier, and D. Birnbaum, "Expression of the fgf6 gene is restricted to developing skeletal muscle in the mouse embryo," *Development*, vol. 118, no. 2, pp. 601–611, 1993.

[159] S. Grass, H.-H. Arnold, and T. Braun, "Alterations in somite patterning of myf-5-deficient mice: a possible role for fgf-4 and fgf-6," *Development*, vol. 122, no. 1, pp. 141–150, 1996.

[160] I. J. Mason, F. Fuller-Pace, R. Smith, and C. Dickson, "Fgf-7 (keratinocyte growth factor) expression during mouse development suggests roles in myogenesis, forebrain regionalisation and epithelial-mesenchymal interactions," *Mechanisms of development*, vol. 45, no. 1, pp. 15–30, 1994.

[161] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[162] K. Vandereyken, A. Sifrim, B. Thienpont, and T. Voet, "Methods and applications for single-cell and spatial multi-omics," *Nature Reviews Genetics*, pp. 1–22, 2023.

[163] D. Zhang, Y. Deng, P. Kukanja, E. Agirre, M. Bartosovic, M. Dong, C. Ma, S. Ma, G. Su, S. Bao, *et al.*, "Spatial epigenome–transcriptome co-profiling of mammalian tissues," *Nature*, pp. 1–10, 2023.

[164] A. Tafvizi, F. Huang, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, "A single-molecule characterization of p53 search on dna," *Proceedings of the National Academy of Sciences*, vol. 108, no. 2, pp. 563–568, 2011.