

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Data-Augmentation for Single-Cell Transcriptomics and Multi-Omics

Permalink

<https://escholarship.org/uc/item/1zw7z4hs>

Author

Karikomi, Matthew K

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Data-Augmentation for Single-Cell Transcriptomics and Multi-Omics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology

by

Matthew K Karikomi

Dissertation Committee:
Professor Qing Nie, Chair
Associate Professor Scott Atwood
Assistant Professor Jing Zhang

2023

DEDICATION

To my parents, for teaching me the joy of building something new.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	vi
LIST OF TABLES	viii
ACKNOWLEDGMENTS	ix
VITA	x
ABSTRACT OF THE DISSERTATION	xi
1 scRNA-seq Imputation and bulk RNA-seq Deconvolution	1
1.1 Introduction	1
1.2 Solutions to Dropout	2
1.2.1 Overview of Method Classes	2
1.2.2 Integrated-Endpoint vs Standalone Imputation Solutions	2
1.2.3 Single vs Multiple-Dataset Solutions	3
1.3 DURIAN	3
1.3.1 Imputation By Multiple Dataset Integration: Intro to DURIAN . . .	3
1.3.2 Iterative Imputation Scheme	5
1.3.3 Deconvolution	6
1.3.4 Deconvolution via Topic-Model	6
1.3.5 dsLDA Likelihood and Posterior Distribution	7
1.3.6 Deconvolution via W-NNLS	8
1.3.7 DURIAN Iterative Algorithm	9
1.3.8 Empirical Convergence Properties of DURIAN	10
1.3.9 Imputation benchmarks for down-sampled and synthetic data	11
1.3.10 Benchmarking with Down-Sampling	11
1.3.11 Benchmarking on Synthetic Data	13
1.3.12 DURIAN improves clustering quality for real data	15
1.3.13 DURIAN enhances intracellular ligand recovery in embryonic skin . .	17
1.4 Downstream Tasks and Prior-Knowledge	18
1.5 Figures and Tables	19
1.5.1 Empirical Convergence Analysis Histograms	20
1.5.2 Error Box-Plots	22
1.5.3 Cluster Quality Box-Plots	25

1.5.4	Signaling Analysis	27
1.5.5	Mean Error Tables: Aggregate by Method	28
1.5.6	Mean Error Tables: Down-Sampled	29
1.5.7	Mean-Error Tables: Simulated	31
1.5.8	Error L2 Norm Tables: Aggregate by method	32
1.5.9	Error L2 Norm Tables: Down-Sampled	33
1.5.10	Error L2 Norm Tables: Simulated	35
2	Automated Mining, Integration, and Annotation of Biological Signaling Pathways	36
2.1	Introduction	36
2.2	Pathway Knowledge Curation and Integration	37
2.2.1	Biological Knowledge Databases and Ontologies	37
2.2.2	Use of Curated Pathways in Cell-Signaling Models	38
2.2.3	Challenges to Pathway Integration	39
2.2.4	Current Approaches to Pathway Integration	40
2.2.5	Graph Databases and RDF	41
2.2.6	BioPAX	43
2.3	PCquery	43
2.3.1	Containerized RDF Endpoints for PathwayCommons and NextProt	43
2.3.2	Pathway Mining and Reactant Imputation	44
2.3.3	Pathway Integration and Annotation	44
2.4	A Unified Pipeline for Distributed Pathway Integration and Annotation	45
2.5	Figures and Tables	46
3	Single-Cell Multiomic scRNA/scATAC Integration	51
3.1	Introduction	51
3.2	Single-Cell Multiomic Profiling of Chromatin Accessibility and Transcription	52
3.2.1	Single-Cell Multiomics	52
3.2.2	Complementary Regulation of Accessibility and Transcription	52
3.3	Lineage Plasticity and Tissue Memory	53
3.3.1	Existence and Characterization of Memory in Epithelial Lineages	53
3.4	Single-Cell Multiomics Integration and the Identification of Modality-Specific Features	53
3.4.1	Paired vs Non-Paired Integration Techniques	54
3.4.2	Interpretable Integration and Modality-Separation	54
3.5	A Variational Factor Model for Detection and Interpretation of Global Decoupling	55
3.5.1	Single-Cell Embeddings and Lineage-Uncertainty	55
3.5.2	Unsupervised Learning With Deep, Generative Models	55
3.5.3	Variational Autoencoders in Single-Cell Data Analysis	55
3.5.4	VAE-based Multiomic Integration Models	56
3.5.5	Representation Learning With Variational Factor Models	56
3.5.6	Augmenting the Multiome Sequencing with Unsupervised ATAC Clustering	56

3.5.7	A Multiomic VAE-Based Factor Model	57
3.5.8	Synthetic Downsampling of PBMC scATC/RNA Multiome Data . . .	57
3.5.9	Exploratory Analysis of the Decoder Weights	59
4	Discussion	61
4.1	scRNA-seq Imputation	61
4.2	Pathway Mining and Integration	62
4.3	Conditional Factor Models for Single-Cell Multiomics	63
	Bibliography	64

LIST OF FIGURES

	Page
1.1 Schematic of DURIAN	19
1.2 Imputation and Deconvolution Error Mutually Improve as DURIAN Converges	20
1.3 Mean Error of DURIAN Imputation Benchmarks	22
1.4 L2 Error of DURIAN Imputation Benchmarks	23
1.5 Cluster-Quality of Imputed Single-Cell Data	25
1.6 Quantitative comparison of signaling network inference following imputation via DURIAN vs. DrImpute: Mouse Embryonic Skin	27
2.1 A) query BioPAX pathway repository based on organism and keyword search. B) for each annotated pathway (blue,yellow) recursively impute complex- forming reactions (RHS blue arrow) based on existing product annotations (green) and their BioPAX-annotated constituents (purple,blue). C) Integrate augmented pathways (blue, yellow) based on common molecular identifiers (green). D) Mine inter-species pathway markers based on sequence-homology between reference integrated pathway protein gene loci and the gene loci of the target species. E) Enumerate signaling trajectories for integrated pathways. Query gene regulatory interactions downstream of cell-surface interactions based on GO annotation of integrated pathways.	46
2.2 Notch1 Cytosolic Pathway (Reactome HSA-2122948)	47
2.3 Notch1 Nuclear Pathway (Reactome HSA-2122947)	48
2.4 Expansion of complexes imputes complex-forming interactions and allows GO annotation of reactants to be mined from NextProt [71]	48
2.5 Integrated signaling graph for human Notch1.	49
2.6 Mapping the gene expression of a single zebrafish neural crest cell to human orthologs in the integrated Notch1 pathway from human	50
3.1 a) Unsupervised clustering of the ATAC modality. b) Augmentation of the paired single-cell multiome ATAC/RNA sequencing with ATAC clustering. .	57

3.2	a) Input data augmented with ATAC clustering. b) Mean, variance, and scale encoders for the RNA and ATAC data. c) Single-modality embedding of the RNA. d) Annealing operator that averages the two single-modality embeddings. e) Multiomic embedding to be processed by the RNA decoder. f) ATAC clusters conditioning the RNA decoder weights. g) RNA decoder weights. h) ATAC decoder weights. i) RNA reconstruction. j) ATAC reconstruction . . .	58
3.3	a) Cells from ATAC clusters labeled via unsupervised Leiden clustering were selected for RNA downsampling. b) Box plot prediction of synthetic decoupling in down-sampled cells based on $W_C^{(RNA)}$	58
3.4	Correlation between ATAC Leiden cluster labels and the dimensions of the multiomic embedding for decoder weight sub-matrix $W_C^{(RNA)}$	59
3.5	Correlation between genes for decoder weight sub-matrix $W_C^{(RNA)}$	60

LIST OF TABLES

	Page
1.1 Stats for benchmark mean error (RMSE)	28
1.2 Adjusted p-values for paired Wilcox tests on benchmark mean error (RMSE)	29
1.3 Stats for benchmark mean error (RMSE) - Strategy:Down-Sampling	29
1.4 Adjusted p-values for paired Wilcox tests on benchmark mean error (RMSE) - Strategy:Down-Sampling	30
1.5 Stats for benchmark mean error (RMSE) - Strategy:Simulation	31
1.6 Adjusted p-values for paired Wilcox tests on benchmark mean error (RMSE) - Strategy:Simulation	31
1.7 Stats for L2 norm error	32
1.8 Adjusted p-values for paired Wilcox tests on imputation error L2 norm . . .	33
1.9 Stats for imputation error L2 norm - Strategy:Down-Sampling	33
1.10 Adjusted p-values for paired Wilcox tests on benchmark L2 norm - Strategy:Down- Sampling	34
1.11 Stats for benchmark L2 norm - Strategy:Simulation	35
1.12 Adjusted p-values for paired Wilcox tests on benchmark L2 norm - Strat- egy:Simulation	35

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation to all those who have supported me throughout this incredible journey of pursuing my Ph.D. and the completion of this dissertation. Without their guidance, encouragement, and unwavering belief in my abilities, this accomplishment would not have been possible.

First and foremost, I would like to extend my heartfelt gratitude to my advisor, Qing Nie, for his invaluable guidance, mentorship, unwavering support, and for creating a research environment where the many collaborative opportunities and exchange of ideas provided so many great opportunities for me. Within the Nie Lab, I would especially like to thank Adam MacLean, Shuxiong Wang, Peijie Zhao, Zixuan Cang, and Axel Almet for sharing their expertise, insightful feedback, and for patiently fielding questions on so many aspects of my PhD work and future career.

I would also like to thank the members of my dissertation committee: Scott Atwood and Jing Zhang, for their valuable contributions, critical insights, and constructive feedback. Their expertise in their respective fields has immensely enriched this research and broadened my understanding of the subject matter.

I am indebted to the faculty at UCI for fostering an intellectually stimulating environment. Their dedication to academic excellence and their commitment to pushing the boundaries of knowledge have been a constant source of inspiration for me.

My deepest appreciation goes to my parents for their support and belief in me. I'm also hugely grateful to the members of my MCSB cohort at UCI for their unwavering love, encouragement, and for making this journey with me both inside and outside the classroom.

Finally, I am grateful to Michelle, for teaching me to "do it scared".

In conclusion, this dissertation would not have been possible without the support, guidance, and encouragement of the individuals and institutions mentioned above. I am deeply grateful to each and every one of them for their contributions, both big and small, in making this academic endeavor a reality.

Thank you.

Matthew K Karikomi

VITA

Matthew K Karikomi

EDUCATION

Doctor of Philosophy in Mathematical, Computational, and Systems Biology	2023
University of California, Irvine	<i>Irvine, CA</i>

Bachelor of Science in Mathematics	2017
The Ohio State University	<i>Columbus, OH</i>

Bachelor of Science in Molecular Genetics	2008
The Ohio State University	<i>Columbus, OH</i>

ABSTRACT OF THE DISSERTATION

Data-Augmentation for Single-Cell Transcriptomics and Multi-Omics

By

Matthew K Karikomi

Doctor of Philosophy in Mathematical, Computational and Systems Biology

University of California, Irvine, 2023

Professor Qing Nie, Chair

With the advent of single-cell transcriptomics and epigenomics, the goals of experimental biology have shifted toward interpretation of the global response to well-defined experimental perturbations. These investigations have two descriptive components: 1) the identification of cells and 2) how the interactions between these units (both individually and collectively) shape this identity. In many cases, these problems involve the identification of unknown biological patterns at multiple scales. This hierarchy of unknowns results in a chicken-or-the-egg problem that I will address through techniques that operate directly on well-defined transformations of raw data, an approach commonly referred to as “data-augmentation”. The relationship between these contributions closely follows the above biological goals: By necessity, the fulfilment of (1) involves measurements are far noisier than traditional techniques which are both more accurate and more limited in the scope of variation they reveal. In this context, (2) introduces extra challenges: it requires accurate, low-noise input, but also benefits from prior knowledge in order to improve its prediction of physically-constrained interactions. I address the problem (1) of denoising in scRNA-seq by developing an iterative imputation algorithm and demonstrate its facilitation of the cell-cell communication (CCC) inference problem posed by (2). I also introduce a data-mining tool whose ability to mine and annotate machine-readable knowledge graphs greatly facilitates the parameterization of existing CCC inference pipelines. Finally, I explore strategies for the prediction of adaptive inflammatory response through global alterations in the coupling of transcription and accessibility.

Chapter 1

scRNA-seq Imputation and bulk RNA-seq Deconvolution

1.1 Introduction

Single-cell library-preparation and sequencing technology has brought about revolution in biology driven by the adaptation of existing measurement techniques and observation modalities [115]. This technology deliberately prioritizes breadth over depth with respect to the feature space. These opportunities and challenges that characterize all single-cell modalities were initially explored and documented for single-cell RNA sequencing (scRNA-seq) [115]. The biggest challenges include the presence of batch effects and dropout in scRNA-seq [48, 43].

Both dropout and batch effects represent platform effects whose structure encompasses non-biological variation across both the features and observations of a dataset [43]. These effects can be handled both implicitly and explicitly by computational methods allowing for varying degrees of modularity in bioinformatics pipelines [81, 7, 34].

1.2 Solutions to Dropout

1.2.1 Overview of Method Classes

There are two broad axes of comparison for existing solutions to dropout: the first axis differentiates between methods that seek to represent dropout within the mathematical framework of an existing analysis "endpoint", such as differential gene expression, vs methods which directly impute missing reads. The second axis differentiates methods whose mathematical representation encompasses single (scRNA-seq) dataset, vs those methods which utilize multiple, complementary datasets.

1.2.2 Integrated-Endpoint vs Standalone Imputation Solutions

Integrated-endpoint approaches can be classified by the stage at which dropout is addressed and we briefly introduce them in order of increasing distance from the original data, with respect to the overall processing pipeline. The pooling approach implemented **scran** R package used a hierarchical model of sparsity to address biological sources of dropout at the normalization stage [69]. The SCDE and ZINB-WAVE packages explicitly account for dropout at the differential-expression stage [112, 58]. Finally, ZIFA [94], URSM [127], CIDR [79], and scVI [81] represent dropout within a dimensionality-reduction framework: ZIFA utilizes a probabilistic version of PCA, URSM utilizes a topic-modeling approach, CIDR uses principle-coordinate analysis similar to ZIFA but which imputes zero reads while fitting the linear model, DrImpute clusters the cells and then averages their reads [34], and scVI learns a non-linear manifold.

Standalone methods for addressing dropout directly impute the missing data without completing an endpoint task such as clustering. These approaches seek to provide the same benefits for downstream analysis afforded by the above approaches, in addition to supporting the use of existing tools that themselves do not explicitly represent dropout in the raw data, such as basic network inference pipelines [49]. Standalone imputation methods can be broadly classified into those which adopt a discriminative, regression-like approach, vs those that do not. Among methods of the first

group, ScImpute [78], VIPER [18], and SAVER [47] fit a sparse regression model whose response is the observed expression value, with both ScImpute and VIPER using a non-negative least squares objective, while the latter fits a hierarchical model via empirical Bayes. A more holistic approach to the problem these methods address, namely the prediction of many genes via sparse regression, is the low-rank matrix completion approach of SCRABBLE [92].

1.2.3 Single vs Multiple-Dataset Solutions

Among the above methods, only URSM and SCRABBLE utilize multiple datasets. Both of these methods make two strong assumptions regarding the source of complementary data: first, the data must be derived from bulk sequencing on a biological sample whose tissue/organ or origin and developmental state are identical to the scRNA-seq data. Second, the bulk sequencing is assumed to be free of dropout. Fortunately, both of these conditions are easily satisfied: bulk mRNA sequencing is a mature technology with thousands of high-quality datasets over the last 15 years [104], bulk sequencing provides greater read depth and immunity to biological sources of dropout such as stochastic gene expression, due to averaging [58, 116].

1.3 DURIAN

1.3.1 Imputation By Multiple Dataset Integration: Intro to DURIAN

The utilization of bulk data in SCRABBLE takes the form of a multiple regression objective representing the difference between the average over all cells of the imputed single-cell data and the average over all samples of the normalized bulk 1.3. The utilization of bulk data in the URSM generative model takes the form of the reference distribution for simulating pre-dropout reads in the single-cell data. Based on these ideas, we introduce DURIAN (Deconvolution and mUltitask-Regression-based ImputAtion), a method which addresses two important limitations of these exist-

ing multi-dataset methods. First, we improve the efficiency of bulk data utilization in SCRABBLE by taking the cell-type profile of the bulk into consideration, turning the third term of the objective (eq.1.3) into a multivariate loss. Second, we retain the Bayesian topic model framework of URSM, while removing the fully-generative framework. This allows greater parameter efficiency, and the option to perform rapid analysis of data when full posterior distribution of the individual reads is not required.

Input for DURIAN includes a single-cell data set and a second bulk (or pseudobulk) data set (Fig. 1.1a,b). At each iteration, DURIAN first computes the deconvolution, an estimate of the percentage of each celltype in the bulk data (Fig. 1.1c). Either Gibbs sampling or non-negative regression may be used for deconvolution, depending on the research priority and available metadata. The scheme under this modular approach is outlined in alg. 1-2, following notation provided in Supplemental Methods. Regardless of algorithm selection, deconvolution of the bulk data is supervised by the original single-cell data during the first iteration, and by the current single-cell imputation estimate during all subsequent iterations. Following deconvolution, the deconvolution map is constructed (Fig. 1.1d). The deconvolution map converts the current imputation estimate to a pseudobulk reference for the observed bulk data via the deconvolution estimate ("Supplemental Methods"). Next, the imputation objective is updated based on the deconvolution map and the minimum-rank imputation estimate is computed (Fig. 1.1e). This imputation is constrained by the difference between the continuously-updated pseudobulk reference and the observed bulk data ("Supplemental Methods"). The next DURIAN iteration then begins with deconvolution of the bulk (Fig. 1.1c), supervised by the newly imputed single-cell data. Stopping criteria are satisfied when the optimized objectives of two successive DURIAN iterations are within tolerance, or a maximum number of iterations has been performed. Empirically, this scheme leads to correlated improvement in both imputation and deconvolution as DURIAN converges (Fig. 1.2). The final output includes both the final deconvolution of the bulk data and the imputed single-cell data (Fig. 1.1f,h). Direct applications of this output include inference of inter-cellular communication networks (Fig. 1.1g), based on complete data.

1.3.2 Iterative Imputation Scheme

Imputation of single-cell data in DURIAN fits a convex optimization objective utilizing a tissue-matched bulk dataset to improve accuracy. Here $\tilde{\Theta}$ represents the deconvolution of the bulk data W (eq. 1.1), which is initialized using raw single-cell data, and based on the imputation estimate $\tilde{X}$ during subsequent iterations. Once $\tilde{\Theta}$ is estimated, I update the deconvolution map $\tilde{S}$, a matrix whose individual columns represent the pseudobulk reference for the corresponding bulk sample in W (eq. 1.2). Finally, dropout reads in X are imputed based on the singular-value thresholding scheme described in [15], which is closely related to the ADMM implementation described in [92]. In contrast to SCRABBLE, I utilize a deconvolution map to pull information from celltype-specific gene expression signatures in the bulk data via the third term in eq. 1.3. In particular, the parameter β (eq. 1.3) penalizes the discrepancy between the expression levels $\tilde{X}$ estimated by imputation and observed bulk data W . I also include the standard parameters for convex matrix completion [15, 92] in eq. 1.3: the parameter α penalizes the rank of the imputed data $\tilde{X}$, and the parameter γ controls the ADMM learning rate. Mathematical details of the two deconvolution approaches I address under this scheme (including our novel topic-model), in addition to pseudocode for DURIAN under these two approaches, are provided in the Supplemental Methods.

$$\tilde{\Theta} = \text{deconv}(W, \tilde{X}) \tag{1.1}$$

$$\tilde{S} = \tilde{\Phi}\tilde{\Theta}, \quad \tilde{\Phi} = \tilde{X}\mathbf{I}^X$$

$$\mathbf{I}_{i,j}^X = \begin{cases} 1 & \text{cell } i \text{ is of type } j \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

$$\tilde{X} = \underset{X \geq 0}{\operatorname{argmin}} \left(\frac{1}{2} \|P_{\Omega}(X) - X_0\|_F^2 + \alpha \|X\|_* + \beta \|\tilde{S} - W\|_2^2 \right) \tag{1.3}$$

1.3.3 Deconvolution

DURIAN alternates between two basic “stages”: imputation and deconvolution. Because the imputed data is treated as fully-observed during deconvolution, while the deconvolution (via the deconvolution map) is treated as fully-observed during imputation, these two stages are fully modular and mathematically independent from each other. To illustrate this point, I implement two alternative deconvolution approaches: an LDA-based topic model (alg. 1), and non-negative least-squares (alg. 2). In addition to demonstrating the adaptability of our model, these two deconvolution approaches offer complementary benefits for single-cell reference data: the regression-based approach is based on the MuSiC deconvolution [118] algorithm which automatically performs batch correction based on inter-sample variation in the reference data. However, this approach requires multiple biological donor subjects, which are not always available. In contrast, our implementation of LDA makes no assumptions about sample origin, and can be used in cases where only a single biological replicate is available, or when sample origin is unknown.

1.3.4 Deconvolution via Topic-Model

To allow scalable deconvolution of large data sets, I introduce an exact, distributed implementation of stwdLDA that runs faster than the originally-described model (introduced for processing of natural-language text) without sacrificing accuracy. Deconvolution of the bulk data in DURIAN fits a mixture model using a matched single-cell dataset. I define a semi-supervised generative model based on standard latent dirichlet allocation [12].

In the deconvolution of bulk transcriptomic data, the composition of each sample corresponds to biological celltypes. In this context, the use of a known expression profile serves three purposes: first, it allows DURIAN to be utilized in modular fashion alongside state of the art unsupervised clustering methods, to identify celltypes of interest prior to imputation. Second the use of these celltype profiles imposes a constraint that makes deconvolution and imputation by DURIAN repeatable, in contrast to unsupervised topic models such as NMF (without a sparsity constraint) or

standard LDA which suffer from non-identifiability [99, 45, 109]. The third benefit of using known celltype expression profiles is that deconvolution is more scaleable on multiple bulk samples, which can be deconvoluted in parallel due to D-separation between the individual bulk samples [39], allowing the celltype assignment of reads from distinct bulk samples to be sampled independently. I therefore refer to our implementation of Gibbs sampling for the stwdLDA model as distributed, static topic word distribution LDA (dsLDA).

1.3.5 dsLDA Likelihood and Posterior Distribution

The model for dsLDA is based on the bulk read likelihood, where each read is assigned a ‘celltype’. The entire bulk sample of reads is then a mixture of various celltypes and their ratios are the latent composition of the tissue, aka the “deconvolution”. I can now give the likelihood for a sample of bulk reads $\mathbf{w}_m$, when the single-cell signature for celltype k is taken as the concentration parameter β_k on the bulk celltype gene distribution ϕ_k , and Φ is the $G \times K$ matrix representing the single cell references of K celltypes, over G genes:

$$p(\mathbf{w}_m, \mathbf{z}_m, \boldsymbol{\theta}_m | \Phi) = \prod_{i=1}^{N_m} p(w_{m,i} | \phi_{z_{m,i}}) p(z_{m,i} | \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m | \boldsymbol{\alpha}) \quad (1.4)$$

In order to perform the deconvolution of W , I sample the latent celltype associated with each read until the Brooks-Gelman criteria are satisfied [14] and take the MAP estimate of the $\tilde{\Theta}$ as the median of the highest density interval of MCMC samples.

The celltype assignment, which is drawn from the latent celltype proportion matrix $\boldsymbol{\theta}_m$, is solely dependent on the local statistics underlying the mixture of celltypes in the current bulk sample: the count $n_m^{(k)}$ of reads in the current sample assigned to celltype k . The second part of the conditional expresses how inference of the celltype giving rise to the current read relies on shared information

between samples. The likelihood $p(w_{m,i} = g)$ of a read given celltype k is:

$$p(w_{m,i} = g) = \frac{n_{k,-i}^{(g)} + \beta_{k,g}}{\sum_{g=1}^G n_{k,-i}^{(g)} + \beta_{k,g}} \quad (1.5)$$

Marginalization of θ_m allows sampling of the posterior for $\mathbf{z}_m$ via histogram representation of $p(\mathbf{w}_m, \mathbf{z}_m)$ [36, 100], giving rise to the “collapsed” explicit representation of stwdLDA [39]:

$$p(z_{m,i} = k | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{m,-i}^{(k)} + \alpha_k}{\left[\sum_{k=1}^K n_m^{(k)} + \alpha_k \right] - 1} \cdot \phi_{z_{m,i}} \quad (1.6)$$

Observation of ϕ_m (the single-cell gene expression profiles) makes the assignment vectors for any two bulk samples $\mathbf{z}_m, \mathbf{z}_{m'} : m \neq m'$ independent from each other. I exploit this so-called d-separation between bulk samples (alg. 1) by sampling all bulk samples simultaneously, greatly accelerating deconvolution compared to previous approaches [39].

1.3.6 Deconvolution via W-NNLS

I used the MuSiC implementation of weighted non-negative least-squares [118] as a high-efficiency alternative to dsLDA. As in eq. 1.4 above, each column ϕ_k of the reference matrix Φ represents the relative expression levels of all genes in celltype k . The latent celltype proportions $\tilde{\theta}_m$ in eq. 1.7 are solved via the Lawson-Hanson NNLS algorithm during iterative update of the latent scaling factors $\tilde{\alpha}_m$ [118]. Due to its dependence on the variance across bulk samples, $\tilde{\alpha}$ inherently deals with batch effects [118].

$$\tilde{\theta}_m = \underset{\alpha, \theta \geq 0}{\operatorname{argmin}} \left(\| \operatorname{diag}(\tilde{\alpha}_m)(\tilde{\mathbf{w}}_m - \Phi \tilde{\theta}_m) \|_2^2 \right) \quad (1.7)$$

1.3.7 DURIAN Iterative Algorithm

Our approach alternates between fitting the dsLDA mixture model to compute the mtSCRABBLE objective and optimizing the mtSCRABBLE objective (D-step and the I-step, respectively). During the initial D-step, the unimputed scRNA-seq data used as the signature for the deconvolution of the bulk. The subsequent I-step utilizes A based on this deconvolution, and estimates the imputed scRNA-seq data $\tilde{X}$. Starting with the next D-step, all deconvolutions Θ_{i+1} utilize the imputed scRNA-seq $\tilde{X}_i$. During the I-step, the penalty parameter β controls how closely the current imputation is required to match the current deconvolution estimate.

Algorithm 1: DURIAN [dsLDA]

```

Initialize:  $\tilde{X} \leftarrow X$ ;
while  $\tilde{X}$  not converged do
    Update:  $\tilde{\Phi} \leftarrow \tilde{X}, W$ ;
    for each bulk sample  $j$  in parallel do
        Sample  $z_{m_j}$  with global static  $\tilde{\Phi}$  ;
        Update:  $N_{w_j}, N_j$ 
    end
    Update:  $N_{w_j} \leftarrow N_w + \sum_p (N_{w_j} - N_w)$ ;
    Update:  $\tilde{\Theta} \leftarrow N_j, N_w$ ;
    Update:  $\tilde{S} \leftarrow \tilde{\Theta}, W, \tilde{X}$ ;
    Optimize:  $\tilde{X}$ ;
end

```

Algorithm 2: DURIAN [MuSiC]

```
Initialize:  $\tilde{X} \leftarrow X$ ;  
while  $\tilde{X}$  not converged do  
  Update:  $\tilde{\Phi} \leftarrow \tilde{X}, W$ ;  
  while  $\tilde{\alpha}$  not converged do  
    Update:  $\tilde{\Theta} \leftarrow \tilde{X}, W, \tilde{\alpha}$ ;  
    Optimize:  $\tilde{\alpha}$ ;  
  end  
  Update:  $\tilde{S} \leftarrow \tilde{\Theta}, W, \tilde{X}$ ;  
  Optimize:  $\tilde{X}$ ;  
end
```

1.3.8 Empirical Convergence Properties of DURIAN

The framework of DURIAN relies on an iterative scheme: the imputation stage treats the deconvolution map (reflecting the estimated celltype percentages of the bulk) as known, while deconvolution stage treats the current imputation estimate of the single-cell data as known. It thus remains to be shown that during successive DURIAN iterations, the deconvolution and imputation estimates are jointly improving. In Fig. 1.2, I show a 2-dimensional histogram of the scaled iteration-by-iteration mean error for down-sampled and splatter simulated runs in Fig. 1.3a, for all replicates corresponding to the two highest dropout rates in each strategy. Because each simulated replicate contains completely different cells, and each down-sampled replicate contains a completely different subset of the original pancreatic data, I adopted the following scaling approach for both mean and L2 error: $|(y_t - y_T)/(y_1 - y_T)|$, where y_1, y_t, y_T represent the value of the statistic at initiation (unimputed data), at iteration t and at convergence, respectively. For each respective strategy (e.g. down-sampling, dsLDA deconvolution) the histogram in the top facet of Fig. 1.2a contains binned imputation error for all replicates and the histogram in the bottom facet directly below it contains the binned deconvolution error (RMSE) for the predicted celltype x bulk sample matrix vs known, for the same replicates. Regardless of synthetic data strategy or deconvolution method, these his-

tograms suggest that as DURIAN approaches convergence, each successive round of imputation benefits from the previous round of deconvolution, and vice versa.

1.3.9 Imputation benchmarks for down-sampled and synthetic data

I compared the performance of DURIAN to existing imputation methods using two standard benchmarking approaches: down-sampled scRNA-seq data to introduce zeros at known locations [78, 92], and simulating single-cell data via the Splatter method [125]. For each benchmarking strategy I compare the performance of six algorithms: DrImpute [34], SCRABBLE [92], URSM [127] (scaled posterior celltype distributions as described in Methods), mtSCRABBLE (a single iteration of DURIAN with a deconvolution map based on the known celltype percentages in the bulk), DURIAN [MuSiC], and DURIAN [dsLDA]. The performance of both DURIAN and SCRABBLE is dependent upon the optimization parameters α, β , and γ (see Methods). In both methods, the α parameter governs the relative importance of low-rank, which favors smaller numbers of implicit cell clusters, and γ determines the relative rate of convergence (ADMM step-size). In DURIAN, the β parameter uniquely controls the relative importance of the deconvolution in the imputation estimate. For each benchmarking task below, I provide results for DURIAN, mtSCRABBLE, and SCRABBLE which utilize a moderate $\beta = 1e - 6$ value, and a high $\beta = 1e - 5$ value. The high value forces the imputation to lean more on celltype information present in the bulk data.

1.3.10 Benchmarking with Down-Sampling

Down-sampling is a common strategy for imputation benchmarking which takes real single-cell expression data as input, and subsequently sets individual entries to zero [78, 92]. Previous down-sampling benchmarks utilized a constant gene-wise dropout probability across single-cell donors, which does not reproduce the batch-related dropout effects introduced via the separate sequenc-

ing of biological replicates, which has previously been reported as a common source of technical bias [44]. I resolve this issue by utilizing donor-specific dropout rates rather than imposing a single, gene-wise dropout probability. Our approach sets the observed value to zero with probability $p = \exp(-\lambda \bar{x}_{g,j}^2)$ where $\bar{x}_{g,j}$ is the mean expression of gene g in single-cell donor j . Here λ represents the dropout rate, given the single-cell sample, where lower values of λ correspond to higher dropout rates. For each of three dropout rates ($\lambda = 1e - 5, 5e - 6, 1e - 6$), I ran imputation on 50 data replicates each composed of 1000 randomly selected genes and 800 randomly selected islet (alpha/beta/gamma/delta/epsilon) cells taken from published single-cell data from healthy adult human pancreatic tissue [10]. In this tissue, the differentiated celltypes provide a strong foothold for deconvolution [118], and this represents the best-case test for DURIAN’s performance compared to other methods. In the top (“Down-Sampling”) horizontal facet of Fig. 1.3a I present the cell-wise mean imputation error (log RMSE) and the corresponding spectral (L2) norm of imputation error in the top facet of Fig. 1.4. In each case, the average sparsity of the three tested λ values is shown in the second-level horizontal facet of Figs. 1.3,1.4.

For all three levels of down-sampling, DURIAN achieves a lower mean error than other methods, regardless of the sparsity of the unimputed data, or the method of deconvolution employed (Fig. 1.3a and tables 1.1,1.2). Each data point in Fig. 1.3a represents the log mean error across the entire data set. To better see how the error may relate to the underlying celltype, the mean across all genes in each individual cell for a single, representative down-sampled replicate is depicted graphically in Fig. 1.3b. Here, a single, unimputed replicate from the maximum dropout group ($\lambda = 1e - 6$, mean sparsity 0.908) was embedded via UMAP [87] and each cell is colored according to the error resulting from each of the imputation methods. In Fig. 1.3b I can see that DURIAN is more resistant than other methods to celltype related bias in dropout, as depicted by the relatively uniform coloring of the UMAP embedding. In contrast, both DrImpute and SCRABBLE show much higher celltype dependent error in the left-most lobe of the UMAP embedding, while URSM shows higher celltype dependent error in the right-most lobe of the embedding.

In contrast to the average error, which provides a domain-agnostic picture of how well imputation performs regardless of underlying celltypes or the individual genes, the L2 error (the largest singular value of the error matrix) shows how consistent the imputation is when considering system-level relationships between genes across all populations, represented by the inherent low-dimensional structure of the data [5]. I found that DURIAN’s measured improvement with respect to the other approaches, across multiple benchmarking strategies, was statistically significant (tables 1.7,1.8). For down-sampled data in particular, I found that for individual dropout levels DURIAN achieves significantly lower L2 error when the β parameter is increased from $1e - 6$ to $1e - 5$ (Fig. 1.4a and tables 1.3,1.4). For very sparse data (corresponding to $\lambda = 1e - 6$) DURIAN’s L2 error is significantly lower than all other methods when β is increased forcing the imputation estimate to rely more heavily on celltype information present in the bulk, which DURIAN obtains via deconvolution (tables 1.3,1.4). For the two lower dropout levels, the L2 error of DURIAN is second only to DrImpute. These results also show that our novel topic-model based deconvolution approach (dsLDA) provides a slight advantage in L2 error rate at lower dropout levels. A complementary perspective to the cell-specific L2 error across genes, is the gene-specific average error across cells via an adaptation of the MA plot for bulk sequencing [97]. Each of the facets of Fig. 1.4b show the gene-wise error of the true vs imputed version of a single down-sampled replicate for the respective method. For each method, log ratio of true vs imputed gene-wise average over the log average of true and imputed gene-wise counts is plotted, with over-estimates highlighted in red and under-estimates highlighted in blue, with error threshold of ± 2 . These plots clearly show that DURIAN’s use of celltype expression profiles in the bulk prevent over-estimation observed in SCRABBLE, a trend which was also reported previously [92].

1.3.11 Benchmarking on Synthetic Data

In contrast to the down-sampled pancreatic data above, a continuum of very closely related celltypes, in addition to batch effects provides a worst-case example of DURIAN’s performance relative to other methods. In this benchmark, batch effects separate simulated single-cell samples as well

as the simulated bulk samples. These challenges are compounded by the decreased distinction between individual celltypes based on a poorly-differentiated lineage. The splatter R package [125] implements a zero-inflated generative model of single-cell library simulation. This model is a standard platform for assessing the effects of biological and technical noise which has been repeatedly validated [70, 101, 84]. I used splatter to simulate 50 single-cell data sets, each composed of 1600 cells and 1000 genes, from four celltypes in a simulated lineage, subject to dropout rates defined by the logistic midpoint parameter ($x_0 = 4.5, 5.5, 6.5$), where higher values of x_0 correspond to higher dropout rates. I used splatter’s batch effect model to generate four batches for each simulation and reserved two batches each for the single-cell and pseudobulk data. For each simulated replicate, the two pseudobulk samples were constructed by taking the mean across all cells bearing their respective batch ids. Like the down-sampling approach above, I present the cell-wise mean imputation error (log RMSE) in the bottom facet of Fig.1.3a and the corresponding L2 error in the bottom facet of Fig. 1.4a. In each case, the average sparsity of the three tested λ values is shown in the second-level horizontal facet of figs. 1.3,1.4.

As with the down-sampled approach, the mean imputation error of DURIAN, regardless of deconvolution approach, is lower than competing methods, for dropout corresponding to $x_0 = 4.5, 5.5$ (Fig. 1.3a and tables 1.5,1.6). For more severe dropout ($x_0 = 6.5$, mean sparsity 0.927), the accuracy of DURIAN is still higher than other methods when β is decreased to $1e-6$, enforcing a weaker reliance on the bulk data.

Furthermore, even when a closer match to the deconvolution is enforced at $x_0 = 6.5$, only SCRABBLE achieves a lower mean error than DURIAN (Fig. 1.3a and tables 1.5,1.6). The graphical representation of mean error for simulated data in Fig. 1.3c clearly shows that splatter simulation is less challenging to other methods, relative to DURIAN than the down-sampling approach based on real data above. More importantly, the UMAP embedding plotted in Fig. 1.3c shows that globally, the four simulated celltypes are more closely related and less distinct than the adult islet cell populations in the down-sampled real data. Therefore it is expected that DURIAN’s utilization of population structure detected in the bulk data will be less important when imputing data where

the global distinction between celltypes is lost.

Even under this more challenging scenario, DURIAN achieves a lower norm error than other methods at all dropout levels, except for SCRABBLE (Fig. 1.4a and tables 1.11,1.12). This is consistent with the fact that SCRABBLE, while unable to use celltype-specific gene expression to its advantage when such global information is rich, is also immune to very challenging deconvolution tasks. The overall trends shown in the MA plots of splatter simulated data in Fig. 1.4c mirror those for down-sampled data in Fig. 1.4b above, except that SCRABBLE does not over-estimate the counts as badly in this plot. I attribute this difference to the fact that SCRABBLE’s regularizer based on the average bulk counts is more effective for very closely-related celltypes in this simulated lineage.

Overall, benchmarking of DURIAN on both the best and worst case scenarios suggests that the method can accurately address the simultaneous tasks of single-cell imputation and bulk-deconvolution, and is robust to concrete selection of deconvolution schemes. Furthermore, I found that when results were pooled across simulation strategies and dropout rates, DURIAN mean and L2 error were significantly lower than the the other methods (tables 1.7,1.8). This is due to DURIAN’s more consistent performance across down-sampled and simulated benchmarks, especially in contrast to SCRABBLE (Fig. 1.4a).

1.3.12 DURIAN improves clustering quality for real data

The down-sampling approach described above directly measures imputation performance on data that is as close as possible to real-world cases. Meanwhile, imputation is also expected to improve the score of raw single-cell data for some standard cluster-quality metrics. In Fig. 1.5a, I present five standard quality metrics (described below) [41] for DURIAN, SCRABBLE, and DrImpute applied to adult pancreas (healthy and diabetic), adult skin (healthy and eczema), and embryonic

(E13.5) mouse skin. The intra-cluster spanning-tree widest gap (`cwidegap`) takes the spanning tree computed for each cluster and calculates the longest distance between two cells in the tree. The cluster diameter is the longest distance between any two cells in a single cluster. For both of these metrics, DURIAN and DrImpute are equivalent on both of the pancreatic data sets, while DURIAN has the largest clustering quality on two of the three remaining data sets. The Dunn index (`dunn`) is the minimum separation over the maximum diameter for the data set. For this metric, DURIAN is equivalent to DrImpute on both pancreatic data sets, and on diseased skin, in the meantime yields better clustering quality for both healthy skin and embryonic skin. The silhouette index (`sindex`) is the mean of the bottom 10% of the distances from every cell in a cluster to the closest cell not in the same cluster. DURIAN is consistent with DrImpute for both pancreatic data sets, as well as healthy skin, and achieves the improved clustering quality for both diseased skin and embryonic skin. The sum-of-squares distance (`within.cluster.ss`) is half the sum of the squared distances for each pair of cells in the cluster divided by the cluster size. For this metric DURIAN is equivalent to DrImpute for both pancreatic tissue and embryonic skin, and achieved satisfactory results for both adult skin datasets.

In addition to the quantitative cluster-quality metrics above, I also compared the robustness of DURIAN’s output with respect to the global structure of the data by inspecting the low-dimensional embeddings of imputed datasets. Figs. 1.5b,c show scatter plots for the UMAP embeddings of each imputation method on the human skin and pancreatic data analyzed in Fig. 1.5a. From these figures, it is obvious that a fair amount of global heterogeneity already exists between mature islet cells in the raw data (shown as “dropout” method) for pancreatic tissue (Fig. 1.5b). In this case, I qualitatively measure the success of imputation based on how well it preserves this separation in the low-dimensional embedding, which could easily be obscured if error was introduced through over-estimation. Conversely, the skin data set (Fig. 1.5c) is composed of many celltypes, some of which are not well-separated, and were identified via clustering on a subset of the data in the original report [38]. Here, I can see that both DURIAN and DrImpute improve the global separation between celltypes, including closely-related fibroblasts which were found to express APOE

and FBN1 in a mutually-exclusive fashion [49, 40], in the cluster approximately centered at (0,-5) in the common UMAP embedding of Fig. 1.5c. Overall, DURIAN improves the quality of unsupervised clustering on single-cell datasets, but can also reveal separation between distinct celltypes not present in the original data.

1.3.13 DURIAN enhances intracellular ligand recovery in embryonic skin

Intercellular signaling relies on the presentation of one or more ‘sender’ molecules (ligands) by one cell, the presentation of one or more ‘receiver’ molecules (receptors and co-receptors) by another cell. The expression of these molecules in general, and ligands in particular, is vital to the cell communication analysis using single-cell data. For instance, CellChat determines significant intercellular interactions via a resampling approach across potential signaling partners such that all possible signaling events between two cells which are recapitulated by less than a threshold percentage of randomized samples (default is 5%) are declared significant [49]. However, the expression of signaling genes is notoriously difficult to detect in single-cell data from tissues where bulk sequencing and other molecular assays have shown activation of their respective pathways [49, 68]. I reason that unique feature of DURIAN to achieve single-cell data imputation using bulk data deconvolution can resolve the challenge and inconsistency in cell communication analysis.

I first compared the cell signaling analysis performance using the imputed single-cell data by DURIAN and DrImpute (as baseline control without bulk data) from mouse E13.5 embryonic skin [38]. In Fig. 1.6a, each row of the bar plot shows a pathway for which the relative abundance of predicted interactions for both DURIAN and DrImpute is depicted as a percentage of all recovered interactions for that pathway for the respective method. DURIAN dominates the recovery of signaling interactions in all 26 pathways except three (including MK and PTN, for which the two methods are tied). The remaining 23 pathways that highlighted DURIAN’s recovery include

non-canonical WNT (ncWNT) signaling, which is broadly implicated in skin pattern formation at E13.5 [6, 96].

Next, I explored the imputation effects of key markers in important pathways. The CellChat database contains a partial list of ligands, receptors, co-receptors and other molecular species for each available pathway. Fig. 1.6b shows the relative imputed expression of all ncWNT pathway markers involved in predicted interactions for the data set. Here I can see that DURIAN recovers ligand expression more broadly and at higher magnitude than DrImpute, and provides the enhancement of all ligands expression in several celltypes (immune, myeloid, melanocyte) for which DrImpute was unable to recover as robustly as DURIAN.

Interestingly, formation of the dermal placode at E13.5-14.5 relies on communication between the fibroblasts of the nascent dermal condensate and the cells of the basal epithelium [38]. I therefore examined the relative intensity of predicted signaling interactions between these two celltypes, based on imputation with DURIAN vs DrImpute in Fig. 1.6c. Here, I can see that DURIAN detects previously hidden interactions between members of the multicellular lineage of differentiated (FIB-A, FIB-B) and proliferating (FIB-P) fibroblasts. In summary, DURIAN can be utilized to enhance signaling analysis in single-cell data with the recovery of intracellular ligand expression.

1.4 Downstream Tasks and Prior-Knowledge

This chapter highlights the problem of scRNA-seq dropout, and introduces the DURIAN algorithm as an scRNA-seq pre-processing step which demonstrated enhanced performance and consistency across imputation benchmarks compared to several existing methods. Additionally, use of DURIAN to pre-process scRNA-seq was shown to enhance the prediction of CCC signaling networks. This chapter highlights the use of data-augmentation within an iterative optimization scheme, allowing

complementary sets of observations: bulk and single-cell RNA-seq to be used in conjunction to remove platform effects and enhance the performance of a modular bioinformatics workflow, whose aim (CCC inference) relies not only on the handling of platform effects arising from data collection, but also on the use of prior-knowledge available from signaling network interaction databases [49].

In the next chapter, I will introduce a data-mining tool (PCquery) to combine and utilize the linked information present in a subset of online databases that organize and expose relational knowledge on biological pathways through BioPAX [24]. In the context of the CCC problem, I will explain how the utilization of prior-knowledge is mainly a data-augmentation problem, whose solution requires the integration and annotation of existing knowledge curation units.

1.5 Figures and Tables

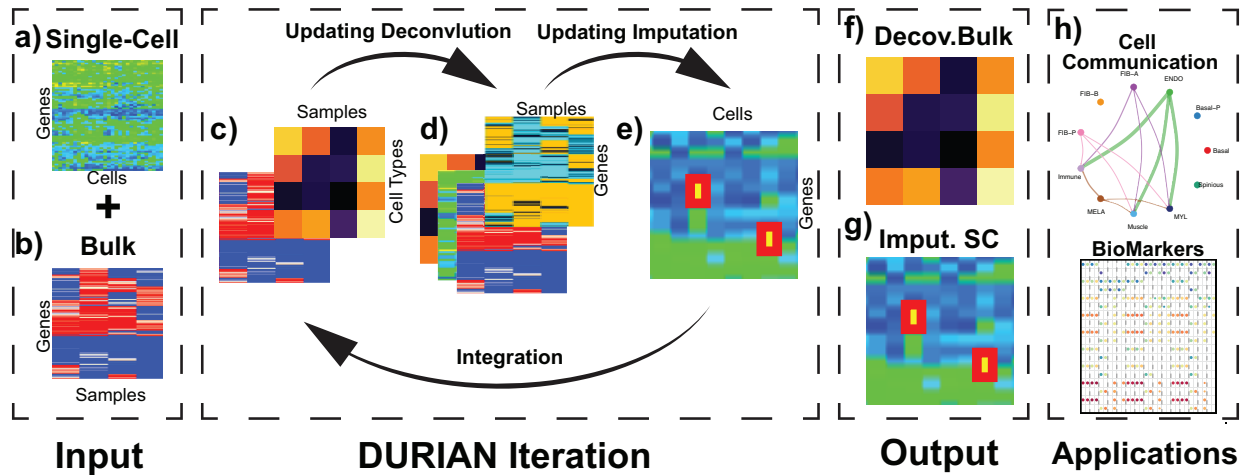


Figure 1.1: **Schematic of DURIAN method.** a) Input: Single-cell expression data ($m \times n$). b) Input: Bulk expression data ($m \times w$). c) Deconvolution of bulk data, yielding a matrix of celltype proportions ($w \times k$). d) Construct deconvolution map ($m \times w$) where each column corresponds to the pseudobulk average of the current imputation result according to the celltype ratios calculated in (c). e) Update the imputation objective with the current deconvolution map. Perform imputation to replace zeros in the single-cell expression data ($m \times n$) with estimated counts. Finally, update the deconvolution model with the imputation estimate. f) Output: Final deconvolution matrix of celltype proportions ($w \times k$). g) Output: Final estimate of single-cell data with dropout replaced by imputed reads. h) Application: Use imputed single-cell data to directly compute cell-signaling patterns.

1.5.1 Empirical Convergence Analysis Histograms

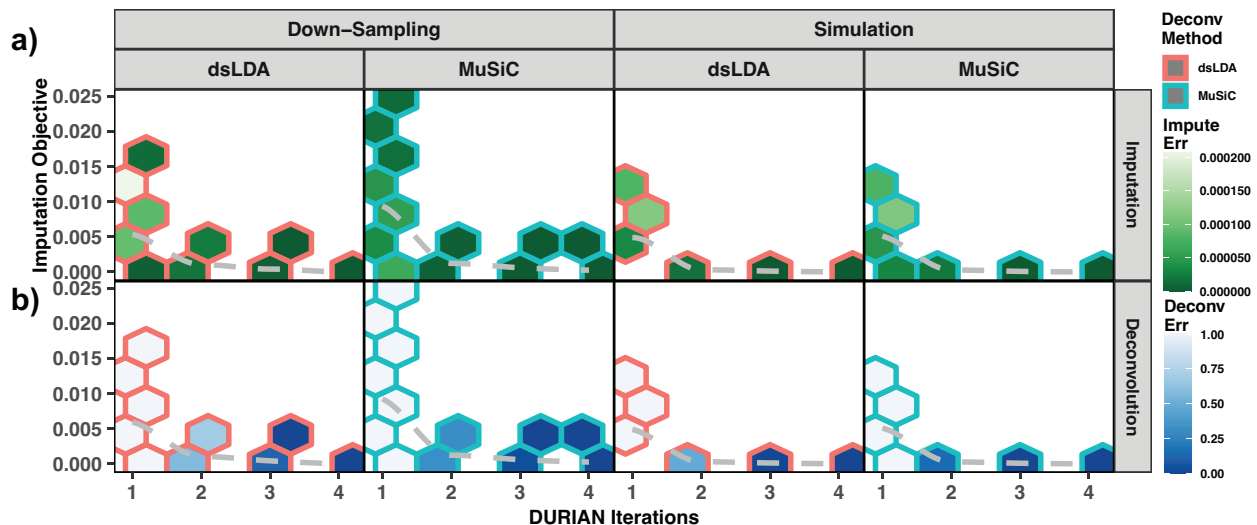


Figure 1.2: **Imputation and Deconvolution Error Mutually Improve as DURIAN Converges.** a-b) 2-Dimensional histograms of the scaled error (for imputation and deconvolution) for all replicates of the two highest dropout rates reported in figs 2-3. a) Scaled imputation error for DURIAN run with both dsLDA and NNLS (MuSiC) deconvolution strategies, with samples subdivided by synthetic data strategy: down-sampling and simulation. b) Scaled deconvolution error collected during the imputation runs plotted in (a).

1.5.2 Error Box-Plots

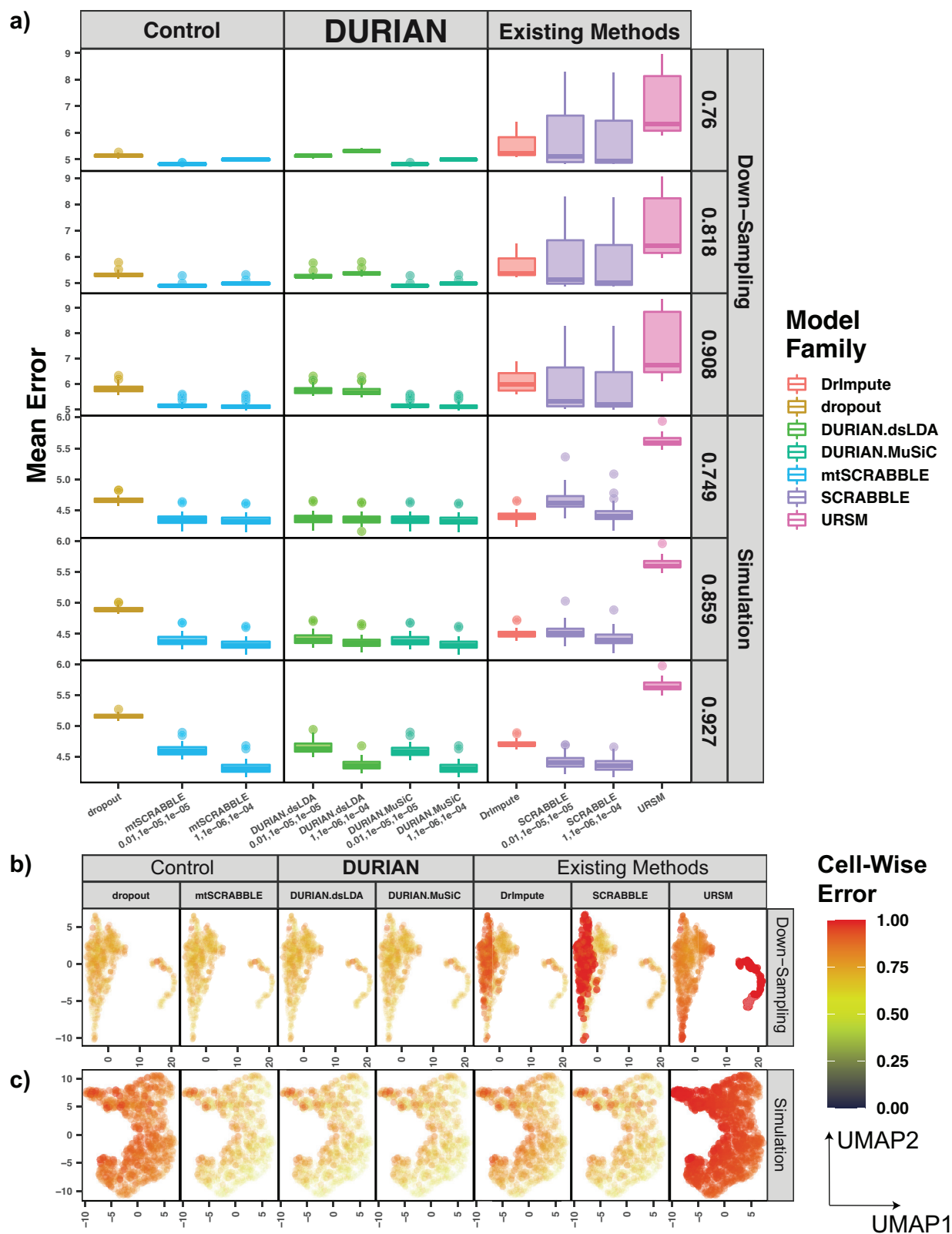


Figure 1.3: **Mean Error of Imputation Benchmarks.** a) Mean error (log RMSE) of imputation for both synthetic data strategies: Down-Sampling and Simulation. For each strategy, the mean sparsity (percentage of zeros in the unimputed data) of all replicates corresponding to each of three values for the strategy-specific dropout parameter is shown in the Supplemental Table 1: 0.76, 0.818, 0.908 (Down-Sampling); 0.749, 0.859, 0.927 (Simulation).

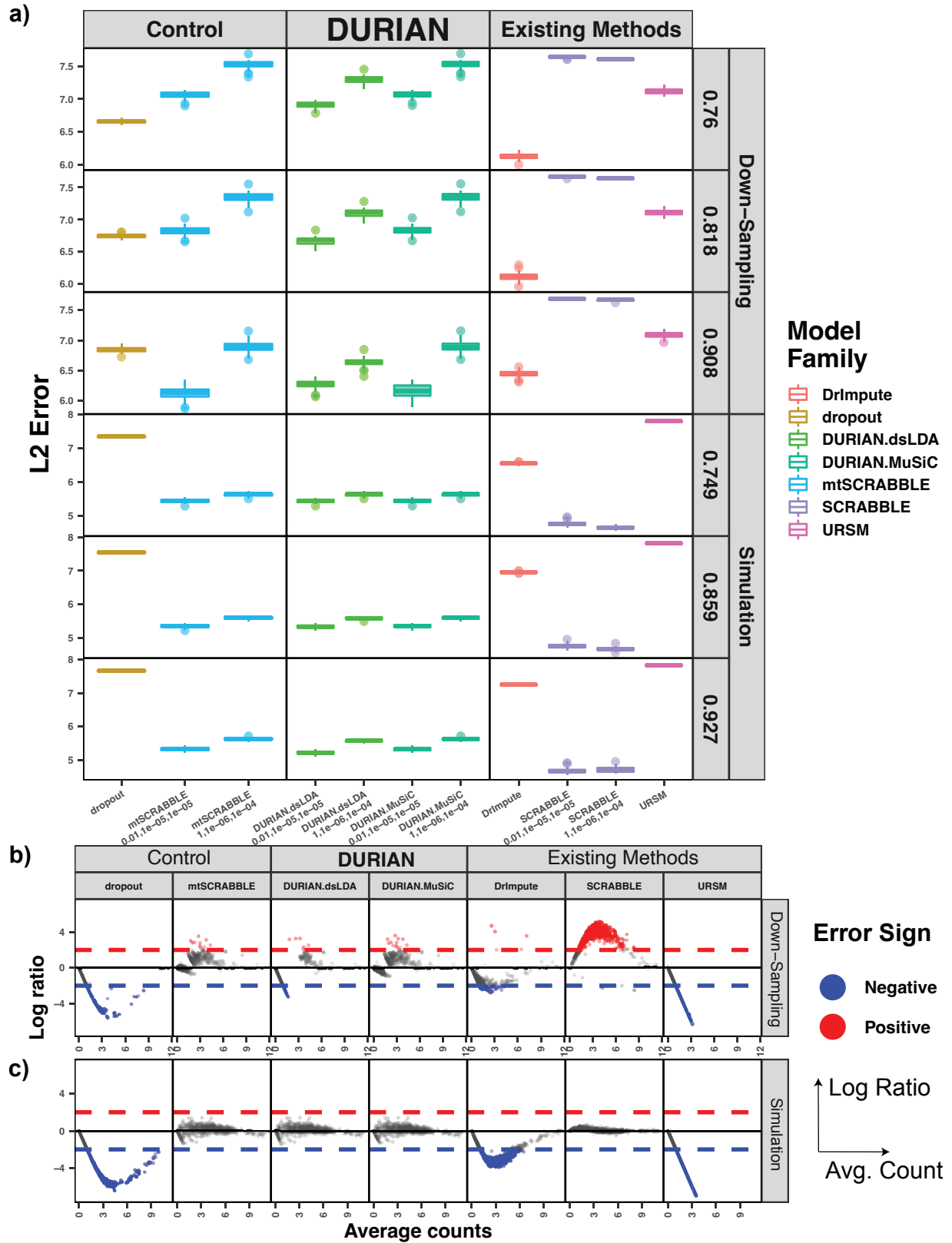


Figure 1.4: **L2 Error of Imputation Benchmarks.** a) The Y-axis is the log L2 norm of difference between the imputed data and the true data, for both synthetic data strategies: Down-Sampling and Simulation. For each strategy, the mean sparsity (percentage of zeros in the unimputed data) of all replicates corresponding to each of three values for the strategy-specific dropout parameter is shown in the 2nd level vertical label: 0.76, 0.818, 0.908 (Down-Sampling), 0.748, 0.859, 0.927 (Simulation). The x-axis is arranged into three method categories (Control, DURIAN, Existing Methods). Control methods include dropout, mtSCRABBLE, and URSM. DURIAN methods include DURIAN.dsLDA and DURIAN.MuSiC. Existing methods include DrImpute, SCRABBLE, and URSM. b) Error Sign. c) Error Sign.

1.5.3 Cluster Quality Box-Plots

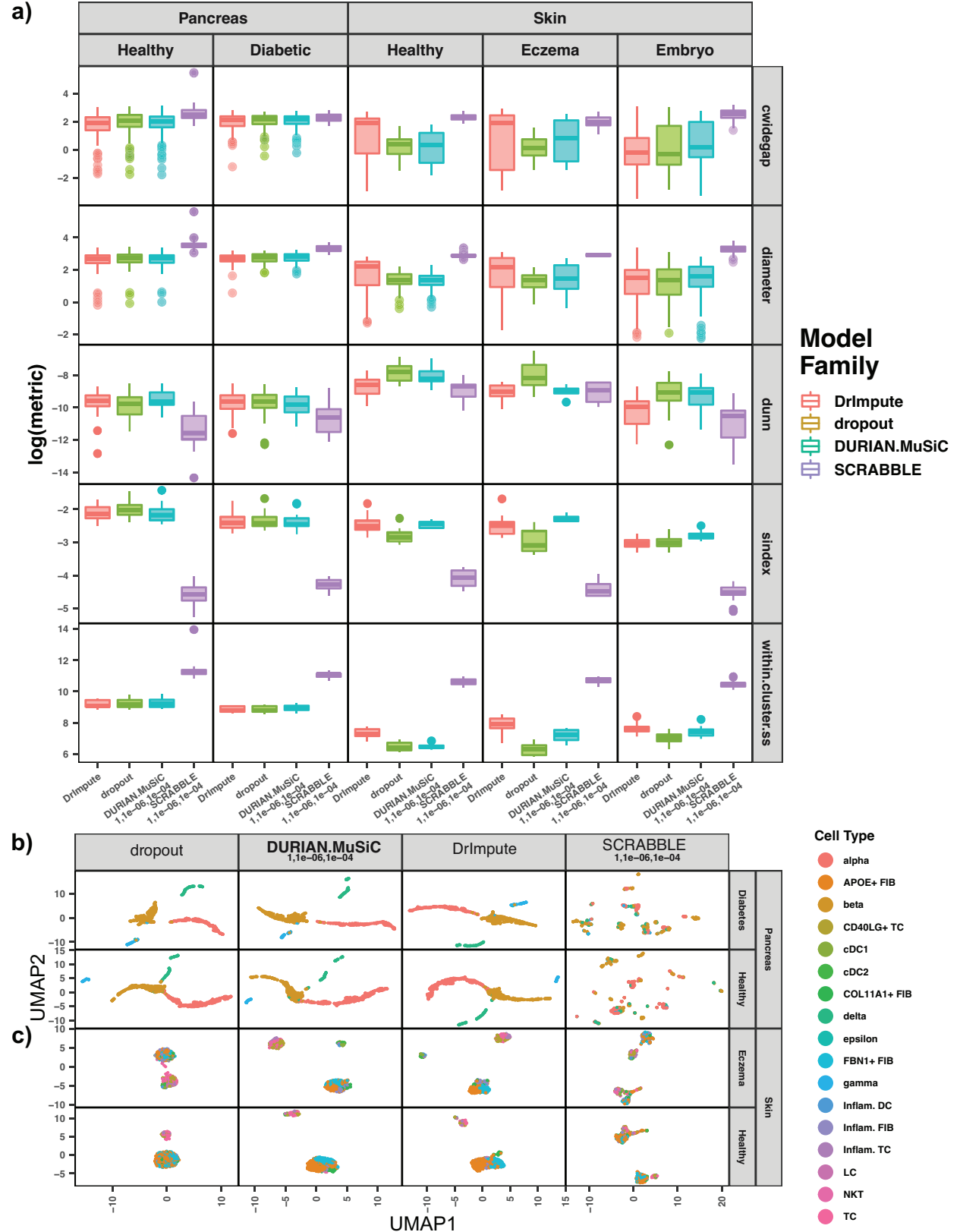


Figure 1.5: **Quantitation of Cluster-Quality for Imputed Single-Cell Data** a) Single-cell data from adult human pancreas (healthy and diabetic) and skin (healthy and eczema), and mouse embryonic skin was imputed via DrImpute, SCRABBLE, and DURIAN. The y-axis corresponds to the log of the following cluster statistics: “cwidgap” (lower is

1.5.4 Signaling Analysis

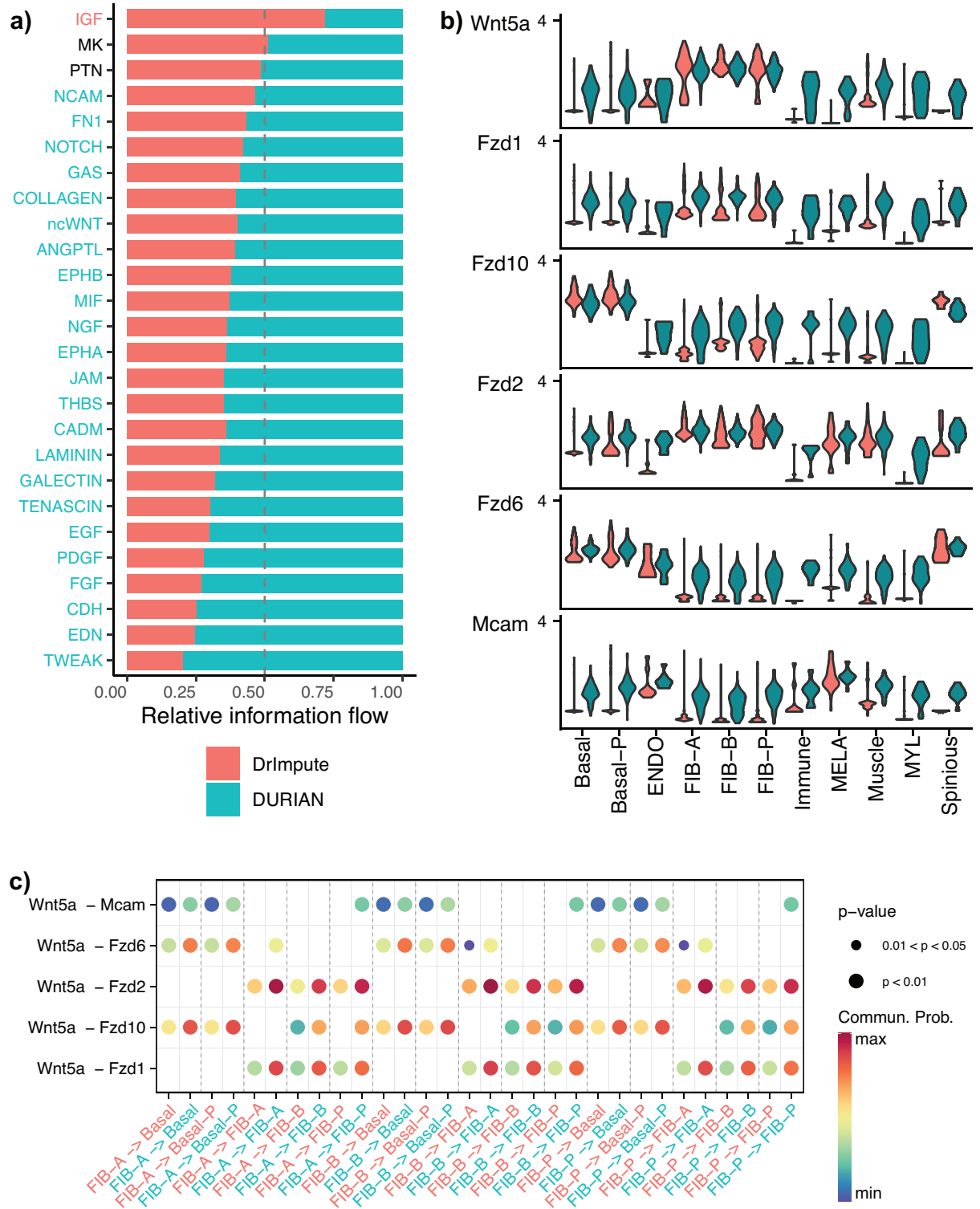


Figure 1.6: **Quantitative Comparison of Imputation via DURIAN and DrImpute on Mouse Embryonic Skin** a-c) CellChat comparative analysis of signaling in mouse E13.5 embryonic skin. a) Relative Information flow (higher is better) on each pathway (for which significant interactions were detected), defined as the sum of communication probability

1.5.5 Mean Error Tables: Aggregate by Method

	modelname	count	mean	sd
1	DURIAN.MuSiC 1,1e-06,1e-04	3.000E+02	4.687E+00	3.677E-01
2	mtSCRABBLE 1,1e-06,1e-04	3.000E+02	4.687E+00	3.679E-01
3	DURIAN.MuSiC 0.01,1e-05,1e-05	2.840E+02	4.698E+00	3.042E-01
4	mtSCRABBLE 0.01,1e-05,1e-05	2.860E+02	4.703E+00	3.044E-01
5	DURIAN.dsLDA 1,1e-06,1e-04	2.930E+02	4.933E+00	5.781E-01
6	DURIAN.dsLDA 0.01,1e-05,1e-05	2.820E+02	4.959E+00	5.145E-01
7	SCRABBLE 1,1e-06,1e-04	3.000E+02	5.048E+00	9.914E-01
8	DrImpute	2.830E+02	5.063E+00	6.472E-01
9	SCRABBLE 0.01,1e-05,1e-05	3.000E+02	5.159E+00	9.813E-01
10	dropout	3.000E+02	5.169E+00	3.696E-01
11	URSM	2.830E+02	6.330E+00	1.069E+00

Table 1.1: Stats for benchmark mean error (RMSE)

	DrImpute	dropout	DURIAN.dsLDA 0.01,1e-05,1e-05	DURIAN.dsLDA 1,1e-06,1e-04	DURIAN.MuSiC 0.01,1e-05,1e-05	DURIAN.MuSiC 1,1e-06,1e-04	mtSCRABBLE 0.01,1e-05,1e-05	mtSCRABBLE 1,1e-06,1e-04	SCRABBLE 0.01,1e-05,1e-05	SCRABBLE 1,1e-06,1e-04
dropout	6.974E-05									
DURIAN.dsLDA 0.01,1e-05,1e-05	5.894E-02	1.802E-07								
DURIAN.dsLDA 1,1e-06,1e-04	8.781E-04	4.061E-04	5.687E-01							
DURIAN.MuSiC 0.01,1e-05,1e-05	7.758E-10	1.590E-45	4.786E-09	4.061E-04						
DURIAN.MuSiC 1,1e-06,1e-04	9.734E-17	8.238E-40	4.685E-16	1.064E-10	6.996E-01					
mtSCRABBLE 0.01,1e-05,1e-05	1.726E-09	6.014E-45	1.001E-08	4.904E-04	8.492E-01	6.018E-01				
mtSCRABBLE 1,1e-06,1e-04	9.734E-17	8.568E-40	4.722E-16	1.064E-10	6.996E-01	9.400E-01	6.018E-01			
SCRABBLE 0.01,1e-05,1e-05	6.157E-01	6.695E-12	9.495E-01	4.330E-02	6.435E-08	6.616E-10	1.790E-07	6.717E-10		
SCRABBLE 1,1e-06,1e-04	1.226E-04	3.194E-19	2.353E-03	4.618E-01	2.435E-02	7.321E-04	3.736E-02	7.423E-04	1.794E-03	
URSM	1.269E-51	9.757E-68	9.757E-68	2.291E-73	7.575E-93	3.206E-95	4.775E-93	3.206E-95	8.203E-52	2.765E-56

Table 1.2: Adjusted p-values for paired Wilcox tests on benchmark mean error (RMSE)

1.5.6 Mean Error Tables: Down-Sampled

	strategy	modelname	count	mean	sd
1	Down-Sampling	DURIAN.MuSiC 0.01,1e-05,1e-05	1.340E+02	4.974E+00	1.711E-01
2	Down-Sampling	mtSCRABBLE 0.01,1e-05,1e-05	1.360E+02	4.977E+00	1.712E-01
3	Down-Sampling	DURIAN.MuSiC 1,1e-06,1e-04	1.500E+02	5.040E+00	1.087E-01
4	Down-Sampling	mtSCRABBLE 1,1e-06,1e-04	1.500E+02	5.041E+00	1.088E-01
5	Down-Sampling	DURIAN.dsLDA 0.01,1e-05,1e-05	1.470E+02	5.396E+00	2.936E-01
6	Down-Sampling	dropout	1.500E+02	5.430E+00	3.087E-01
7	Down-Sampling	DURIAN.dsLDA 1,1e-06,1e-04	1.500E+02	5.473E+00	2.115E-01
8	Down-Sampling	DrImpute	1.330E+02	5.651E+00	4.643E-01
9	Down-Sampling	SCRABBLE 1,1e-06,1e-04	1.500E+02	5.681E+00	1.072E+00
10	Down-Sampling	SCRABBLE 0.01,1e-05,1e-05	1.500E+02	5.781E+00	1.061E+00
11	Down-Sampling	URSM	1.330E+02	7.115E+00	1.124E+00

Table 1.3: Stats for benchmark mean error (RMSE) - Strategy:Down-Sampling

	DrImpute	dropout	DURIAN.dsLDA 0.01,1e-05,1e-05	DURIAN.dsLDA 1,1e-06,1e-04	DURIAN.MuSiC 0.01,1e-05,1e-05	DURIAN.MuSiC 1,1e-06,1e-04	mtSCRABBLE 0.01,1e-05,1e-05	mtSCRABBLE 1,1e-06,1e-04	SCRABBLE 0.01,1e-05,1e-05	SCRABBLE 1,1e-06,1e-04
dropout	4.731E-05									
DURIAN.dsLDA 0.01,1e-05,1e-05	8.217E-07	2.248E-01								
DURIAN.dsLDA 1,1e-06,1e-04	2.400E-01	1.076E-03	4.983E-06							
DURIAN.MuSiC 0.01,1e-05,1e-05	6.739E-38	3.178E-35	8.818E-34	4.008E-42						
DURIAN.MuSiC 1,1e-06,1e-04	1.195E-41	3.538E-40	3.896E-39	5.103E-45	2.155E-07					
mtSCRABBLE 0.01,1e-05,1e-05	5.278E-38	3.521E-35	9.770E-34	1.736E-42	8.715E-01	7.587E-07				
mtSCRABBLE 1,1e-06,1e-04	1.195E-41	3.949E-40	4.545E-39	5.103E-45	2.018E-07	8.715E-01	7.056E-07			
SCRABBLE 0.01,1e-05,1e-05	2.864E-03	1.041E-01	2.197E-01	2.723E-03	6.468E-18	1.529E-09	1.324E-17	1.707E-09		
SCRABBLE 1,1e-06,1e-04	7.587E-07	4.543E-05	1.134E-04	1.863E-07	9.345E-13	1.959E-03	2.102E-12	2.070E-03	9.075E-02	
URSM	1.432E-30	5.960E-45	5.960E-45	2.557E-45	1.659E-44	2.633E-46	8.883E-45	2.633E-46	9.805E-20	4.086E-21

Table 1.4: Adjusted p-values for paired Wilcox tests on benchmark mean error (RMSE) - Strategy:Down-Sampling

1.5.7 Mean-Error Tables: Simulated

	strategy	modelname	count	mean	sd
1	Simulation	DURIAN.MuSiC 1,1e-06,1e-04	1.500E+02	4.334E+00	9.188E-02
2	Simulation	mtSCRABBLE 1,1e-06,1e-04	1.500E+02	4.334E+00	9.186E-02
3	Simulation	DURIAN.dsLDA 1,1e-06,1e-04	1.430E+02	4.366E+00	8.835E-02
4	Simulation	SCRABBLE 1,1e-06,1e-04	1.500E+02	4.416E+00	1.302E-01
5	Simulation	DURIAN.MuSiC 0.01,1e-05,1e-05	1.500E+02	4.451E+00	1.410E-01
6	Simulation	mtSCRABBLE 0.01,1e-05,1e-05	1.500E+02	4.455E+00	1.430E-01
7	Simulation	DURIAN.dsLDA 0.01,1e-05,1e-05	1.350E+02	4.484E+00	1.549E-01
8	Simulation	SCRABBLE 0.01,1e-05,1e-05	1.500E+02	4.538E+00	1.688E-01
9	Simulation	DrImpute	1.500E+02	4.542E+00	1.412E-01
10	Simulation	dropout	1.500E+02	4.908E+00	2.042E-01
11	Simulation	URSM	1.500E+02	5.635E+00	8.879E-02

Table 1.5: Stats for benchmark mean error (RMSE) - Strategy:Simulation

	DrImpute	dropout	DURIAN.dsLDA 0.01,1e-05,1e-05	DURIAN.dsLDA 1,1e-06,1e-04	DURIAN.MuSiC 0.01,1e-05,1e-05	DURIAN.MuSiC 1,1e-06,1e-04	mtSCRABBLE 0.01,1e-05,1e-05	mtSCRABBLE 1,1e-06,1e-04	SCRABBLE 0.01,1e-05,1e-05	SCRABBLE 1,1e-06,1e-04
dropout	2.146E-34									
DURIAN.dsLDA 0.01,1e-05,1e-05	4.032E-04	2.807E-38								
DURIAN.dsLDA 1,1e-06,1e-04	2.599E-26	4.008E-48	6.118E-11							
DURIAN.MuSiC 0.01,1e-05,1e-05	1.224E-07	4.313E-44	9.112E-02	1.981E-07						
DURIAN.MuSiC 1,1e-06,1e-04	3.079E-33	1.853E-49	4.886E-18	5.556E-04	3.477E-14					
mtSCRABBLE 0.01,1e-05,1e-05	3.418E-07	6.706E-44	1.334E-01	9.992E-08	8.136E-01	1.379E-14				
mtSCRABBLE 1,1e-06,1e-04	3.232E-33	1.853E-49	4.889E-18	5.726E-04	3.510E-14	9.115E-01	1.405E-14			
SCRABBLE 0.01,1e-05,1e-05	6.012E-01	2.554E-36	6.620E-03	1.228E-22	4.898E-06	4.419E-29	1.526E-05	4.419E-29		
SCRABBLE 1,1e-06,1e-04	1.272E-14	6.959E-47	4.076E-04	2.456E-04	4.549E-02	1.044E-10	2.790E-02	1.061E-10	3.435E-12	
URSM	7.421E-50	7.421E-50	1.748E-47	8.029E-49	7.421E-50	7.421E-50	7.421E-50	7.421E-50	7.421E-50	7.421E-50

Table 1.6: Adjusted p-values for paired Wilcox tests on benchmark mean error (RMSE) - Strategy:Simulation

1.5.8 Error L2 Norm Tables: Aggregate by method

	modelname	count	mean	sd
1	mtSCRABBLE 0.01,1e-05,1e-05	2.860E+02	5.975E+00	6.956E-01
2	DURIAN.MuSiC 0.01,1e-05,1e-05	2.840E+02	5.981E+00	7.018E-01
3	DURIAN.dsLDA 0.01,1e-05,1e-05	2.820E+02	6.000E+00	6.744E-01
4	SCRABBLE 1,1e-06,1e-04	3.000E+02	6.159E+00	1.482E+00
5	SCRABBLE 0.01,1e-05,1e-05	3.000E+02	6.203E+00	1.463E+00
6	DURIAN.dsLDA 1,1e-06,1e-04	2.930E+02	6.321E+00	7.387E-01
7	mtSCRABBLE 1,1e-06,1e-04	3.000E+02	6.436E+00	8.414E-01
8	DURIAN.MuSiC 1,1e-06,1e-04	3.000E+02	6.438E+00	8.430E-01
9	DrImpute	2.830E+02	6.582E+00	4.292E-01
10	dropout	3.000E+02	7.133E+00	4.004E-01
11	URSM	2.830E+02	7.479E+00	3.537E-01

Table 1.7: Stats for L2 norm error

	DrImpute	dropout	DURIAN.dsLDA 0.01,1e-05,1e-05	DURIAN.dsLDA 1,1e-06,1e-04	DURIAN.MuSiC 0.01,1e-05,1e-05	DURIAN.MuSiC 1,1e-06,1e-04	mtSCRABBLE 0.01,1e-05,1e-05	mtSCRABBLE 1,1e-06,1e-04	SCRABBLE 0.01,1e-05,1e-05	SCRABBLE 1,1e-06,1e-04
dropout	7.013E-41									
DURIAN.dsLDA 0.01,1e-05,1e-05	1.753E-18	3.437E-61								
DURIAN.dsLDA 1,1e-06,1e-04	3.353E-04	2.500E-38	9.042E-16							
DURIAN.MuSiC 0.01,1e-05,1e-05	1.173E-22	2.876E-52	6.327E-01	5.062E-18						
DURIAN.MuSiC 1,1e-06,1e-04	2.835E-02	2.191E-17	4.485E-20	1.392E-04	3.542E-22					
mtSCRABBLE 0.01,1e-05,1e-05	8.459E-24	4.041E-53	7.687E-01	5.289E-18	1.000E+00	4.027E-22				
mtSCRABBLE 1,1e-06,1e-04	2.751E-02	1.243E-17	5.506E-20	1.514E-04	3.882E-22	1.000E+00	4.474E-22			
SCRABBLE 0.01,1e-05,1e-05	1.000E+00	8.991E-02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00		
SCRABBLE 1,1e-06,1e-04	1.000E+00	1.090E-02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	7.505E-07	
URSM	7.632E-65	6.513E-28	3.164E-92	1.011E-54	1.788E-86	1.405E-45	2.388E-87	1.405E-45	6.513E-28	6.513E-28

Table 1.8: Adjusted p-values for paired Wilcox tests on imputation error L2 norm

1.5.9 Error L2 Norm Tables: Down-Sampled

	strategy	modelname	count	mean	sd
1	Down-Sampling	DrImpute	1.330E+02	6.200E+00	1.519E-01
2	Down-Sampling	DURIAN.dsLDA 0.01,1e-05,1e-05	1.470E+02	6.613E+00	2.713E-01
3	Down-Sampling	mtSCRABBLE 0.01,1e-05,1e-05	1.360E+02	6.641E+00	4.043E-01
4	Down-Sampling	DURIAN.MuSiC 0.01,1e-05,1e-05	1.340E+02	6.664E+00	3.907E-01
5	Down-Sampling	dropout	1.500E+02	6.749E+00	8.337E-02
6	Down-Sampling	DURIAN.dsLDA 1,1e-06,1e-04	1.500E+02	7.012E+00	2.877E-01
7	Down-Sampling	URSM	1.330E+02	7.106E+00	4.616E-02
8	Down-Sampling	mtSCRABBLE 1,1e-06,1e-04	1.500E+02	7.252E+00	2.788E-01
9	Down-Sampling	DURIAN.MuSiC 1,1e-06,1e-04	1.500E+02	7.256E+00	2.788E-01
10	Down-Sampling	SCRABBLE 1,1e-06,1e-04	1.500E+02	7.637E+00	2.854E-02
11	Down-Sampling	SCRABBLE 0.01,1e-05,1e-05	1.500E+02	7.663E+00	2.277E-02

Table 1.9: Stats for imputation error L2 norm - Strategy:Down-Sampling

	DrImpute	dropout	DURIAN.dsLDA 0.01,1e-05,1e-05	DURIAN.dsLDA 1,1e-06,1e-04	DURIAN.MuSiC 0.01,1e-05,1e-05	DURIAN.MuSiC 1,1e-06,1e-04	mtSCRABBLE 0.01,1e-05,1e-05	mtSCRABBLE 1,1e-06,1e-04	SCRABBLE 0.01,1e-05,1e-05	SCRABBLE 1,1e-06,1e-04
dropout	7.013E-41									
DURIAN.dsLDA 0.01,1e-05,1e-05	1.753E-18	3.437E-61								
DURIAN.dsLDA 1,1e-06,1e-04	3.353E-04	2.500E-38	9.042E-16							
DURIAN.MuSiC 0.01,1e-05,1e-05	1.173E-22	2.876E-52	6.327E-01	5.062E-18						
DURIAN.MuSiC 1,1e-06,1e-04	2.835E-02	2.191E-17	4.485E-20	1.392E-04	3.542E-22					
mtSCRABBLE 0.01,1e-05,1e-05	8.459E-24	4.041E-53	7.687E-01	5.289E-18	1.000E+00	4.027E-22				
mtSCRABBLE 1,1e-06,1e-04	2.751E-02	1.243E-17	5.506E-20	1.514E-04	3.882E-22	1.000E+00	4.474E-22			
SCRABBLE 0.01,1e-05,1e-05	1.000E+00	8.991E-02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00		
SCRABBLE 1,1e-06,1e-04	1.000E+00	1.090E-02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	7.505E-07	
URSM	7.632E-65	6.513E-28	3.164E-92	1.011E-54	1.788E-86	1.405E-45	2.388E-87	1.405E-45	6.513E-28	6.513E-28

Table 1.10: Adjusted p-values for paired Wilcox tests on benchmark L2 norm - Strategy:Down-Sampling

1.5.10 Error L2 Norm Tables: Simulated

	strategy	modelname	count	mean	sd
1	Simulation	SCRABBLE 1,1e-06,1e-04	1.500E+02	4.681E+00	6.180E-02
2	Simulation	SCRABBLE 0.01,1e-05,1e-05	1.500E+02	4.744E+00	8.528E-02
3	Simulation	DURIAN.dsLDA 0.01,1e-05,1e-05	1.350E+02	5.332E+00	1.065E-01
4	Simulation	mtSCRABBLE 0.01,1e-05,1e-05	1.500E+02	5.371E+00	7.508E-02
5	Simulation	DURIAN.MuSiC 0.01,1e-05,1e-05	1.500E+02	5.372E+00	7.683E-02
6	Simulation	DURIAN.dsLDA 1,1e-06,1e-04	1.430E+02	5.597E+00	4.851E-02
7	Simulation	mtSCRABBLE 1,1e-06,1e-04	1.500E+02	5.620E+00	4.163E-02
8	Simulation	DURIAN.MuSiC 1,1e-06,1e-04	1.500E+02	5.620E+00	4.164E-02
9	Simulation	DrImpute	1.500E+02	6.920E+00	2.885E-01
10	Simulation	dropout	1.500E+02	7.516E+00	1.373E-01
11	Simulation	URSM	1.500E+02	7.809E+00	2.623E-02

Table 1.11: Stats for benchmark L2 norm - Strategy:Simulation

	DrImpute	dropout	DURIAN.dsLDA 0.01,1e-05,1e-05	DURIAN.dsLDA 1,1e-06,1e-04	DURIAN.MuSiC 0.01,1e-05,1e-05	DURIAN.MuSiC 1,1e-06,1e-04	mtSCRABBLE 0.01,1e-05,1e-05	mtSCRABBLE 1,1e-06,1e-04	SCRABBLE 0.01,1e-05,1e-05	SCRABBLE 1,1e-06,1e-04
dropout	7.013E-41									
DURIAN.dsLDA 0.01,1e-05,1e-05	1.753E-18	3.437E-61								
DURIAN.dsLDA 1,1e-06,1e-04	3.353E-04	2.500E-38	9.042E-16							
DURIAN.MuSiC 0.01,1e-05,1e-05	1.173E-22	2.876E-52	6.327E-01	5.062E-18						
DURIAN.MuSiC 1,1e-06,1e-04	2.835E-02	2.191E-17	4.485E-20	1.392E-04	3.542E-22					
mtSCRABBLE 0.01,1e-05,1e-05	8.459E-24	4.041E-53	7.687E-01	5.289E-18	1.000E+00	4.027E-22				
mtSCRABBLE 1,1e-06,1e-04	2.751E-02	1.243E-17	5.506E-20	1.514E-04	3.882E-22	1.000E+00	4.474E-22			
SCRABBLE 0.01,1e-05,1e-05	1.000E+00	8.991E-02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00		
SCRABBLE 1,1e-06,1e-04	1.000E+00	1.090E-02	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	1.000E+00	7.505E-07	
URSM	7.632E-65	6.513E-28	3.164E-92	1.011E-54	1.788E-86	1.405E-45	2.388E-87	1.405E-45	6.513E-28	6.513E-28

Table 1.12: Adjusted p-values for paired Wilcox tests on benchmark L2 norm - Strategy:Simulation

Chapter 2

Automated Mining, Integration, and Annotation of Biological Signaling Pathways

2.1 Introduction

Quantitative tasks in biology involve two general kinds of aims: 1) the "forward problem" of prediction, and 2) the "inverse problem" of parameterizing predictive models [20]. These tasks all have the general aim of using data to fit the parameters of a model. Exploratory tasks in computational biology, such as the identification of cell-cell signaling networks, are inverse problems. While the scope of biological inverse problems is broad, prior-knowledge is critical to the effective use of data in the estimation of model parameters. Implementing the inverse problem involves constraining the representation of the system [2], based on the modality and structure of the data being used to fit the model. Inverse problems in bioinformatics often involve additional, higher-level constraints to the representation of the model in the form of genomic information, or information derived from genome-level analysis [121, 53]. In the case of cell-cell signaling networks, a network model parame-

terized by data on gene expression requires prior-knowledge of the network markers and their roles [25].

Experimental knowledge of cell-signaling informs the teleological design of computational models. However, the curation of annotated pathways is both incomplete and highly-centralized, leading to inefficient use of existing knowledge and limited scalability in the face of increasingly rapid accumulation of new data.

I present PCQuery: an automated pipeline that delivers decentralized, on-demand integration of biological signaling pathways. PCQuery facilitates regularized and high-dimensional cell signaling models by exposing long-range signal-transduction cascades whose provenance is fully-annotated and verifiable. The modular package includes a collection of SPARQL clients to pull current annotations from UniProt, OntoDB, and NextProt. This structural, evolutionary, and clinical knowledge can be graphically displayed by a comprehensive set of pathway visualization tools. Finally, the integrated pathways output by PCQuery support direct combinatorial operations (like signaling-trajectory inference) and are annotated in a subset of BioPAX v3 to support downstream application-development 3.2. Currently, the PCQuery package is available on JuliaHub [1] and at the project github repository [56].

2.2 Pathway Knowledge Curation and Integration

2.2.1 Biological Knowledge Databases and Ontologies

The systematic organization and digital storage of biological knowledge is a key component of the genomic revolution which began almost a century ago, and is closely related to the development and increasing availability of the internet and world wide web [53]. Simultaneously, ontologies have emerged in order to more effectively annotate this information and support linkage across databases [65]. Biological databases can be roughly divided among those that curate 1) pathways,

2) single molecules including sequence information, and 3) whole genomes. Category (1) includes resources such as PathwayCommons, WikiPathways, KEGG, and Reactome [19]. These resources describe the hierarchy of biological interactions and their context, while linking to other databases in Category (2). Databases in Category (2) includes resources such as UniProt [21], NextProt [30], and ChEMBL [88]. Finally, Category (3) includes resources that host reference sequences for species and the comparative analysis of these sequences: [22] and OrthoDB [66]. The most two critical features shared by all of these databases are 1) that they allow programmatic access through public internet addresses (sometimes referred to as "endpoints"), and 2) that they encompass this knowledge from a unified ontological perspective: all biological entities, from individual proteins to entire species, have the same unique digital identifier across all databases [105]. Within this ecosystem, the automation of knowledge retrieval and synthesis is possible using programs that can execute queries on the public endpoints and manipulate this information based on the published ontologies.

2.2.2 Use of Curated Pathways in Cell-Signaling Models

Computational models of cell-cell-communication (CCC) [4] provide state-of-the-art predictions of developmental processes, tissue-plasticity, and clinical treatment outcomes [31, 108, 37]. These predictions are the output of pipelines such as SoptSC, CellChat, CellPhoneDB, SpaOTsc, and COMMOT, which convert machine-readable pathway knowledge stored in online databases (such as KEGG) into a teleological model of cell-signaling constrained by factors such as binding stoichiometry, competition between reactants, and physical proximity [117, 50, 28, 16, 17, 54].

The scope of CCC models can be evaluated based on two criteria: 1) *dimensionality*: the level of specificity at which the set of candidate pathways is represented, and 2) *regularization*: how well the model rejects false-positives. To illustrate the scope of dimensionality in existing methods, all of the pipelines above take into account the co-expression of experimentally-verified ligands and receptors for candidate pairs of cells (or spatial units) but only SoptSC takes into account the

expression of transcriptional target genes [117]. Regarding regularization, two types of constraints can be devised: 1) *teleological constraints* or constraints which rely on knowledge of the signaling pathway itself (e.g. transcriptional targets downstream of a ligand/receptor binding event), and 2) *integration-based constraints* which rely on a separate dataset (e.g. spatial transcriptomics) to eliminate physically-impossible predictions such as juxtacrine signaling between non-neighboring cells [117, 16, 17].

Although *low-dimensional representations* of biological data are critical to achieving systems-level understanding [80], when considering the *observation model* for a CCC pipeline, limited dimensionality means that our model is less sensitive and therefore has a higher false-negative rate, while sub-optimal regularization means that our model has a higher rate of false positives. These limitations on dimensionality and regularization are usually a result of insufficient experimental knowledge, insufficient data, or both [4]. On one hand, the integration of single-cell multi-modal data is an ongoing challenge resulting from the necessity of working with tiny amounts of material. However, more efficient use of new and existing experimental knowledge can open the door to much higher-dimensional (and therefore more sensitive) CCC pipelines, and greater potential for the application of multi-modal data in regularization.

2.2.3 Challenges to Pathway Integration

Two main issues confront the design of high-dimensional teleological models: 1) That only a subset of a single CCC model such as the SoptSC Ligand/Receptor/Target tuple, may be covered by any single "pathway" available from databases like PathwayCommons, and 2) that a human cannot reliably discern the (sometimes implicit) overlap between any two such "pathways" without being able to parse (and occasionally extend) the barely-human-readable XML code in which these online databases store this knowledge. The reason for issue (1) above is a practical problem of converting published results into machine-readable form: the ongoing generation of this knowledge,

and the aggregation of existing knowledge face the practical limitations of humans reading research publications, and connecting the dots between experiments. The reason for issue (2) is related: even when experienced researchers undertake focused literature review, individual experimental studies in turn present challenges to knowledge integration. Thus, the fragmentation of signaling pathway knowledge characterizing the primary literature persists in the curation of online pathway databases.

2.2.4 Current Approaches to Pathway Integration

Existing strategies for backend pathway knowledge management in bioinformatics take the form of pre-curated signaling-pathway graphs such as OmniPath. Although this static framework is more straightforward for package developers to support, and does not require the user to run a query – allowing immediate access to pathway knowledge present in each release, it is limited in several key ways: 1) their annotation cannot be updated by the user, 2) they do not offer the user the ability to build ad-hoc pathways to include recently-published pathway information that has been curated but not yet included in the next package release, 3) they do not allow automated integration of existing pathways based on shared identifiers or identifiers imputed as reactants involved in the formation of complexes whose identifiers are already present in the pathway, 4) they do not leverage the extensibility of BioPAX which allows custom annotation of encoded pathways in a form that is easily transmitted to document downstream modeling or curation efforts. PCQuery addresses all of these limitations in a package that is written in native Julia, and callable in Python via a lightweight wrapper.

2.2.5 Graph Databases and RDF

Graph databases support the storage and retrieval of information in a similar way to traditional table-based databases. In contrast tabular databases which store single pieces of information in rows of a table, graph databases store pieces of information as edges in a graph (called "pairs"). Nevertheless, any tabular database can be converted to a graph and vice-versa. However, graph databases offer three unique capabilities that make them ideal for bioinformatics: 1) the knowledge structure (or ontology) of the database is encoded in the data graph itself, 2) a query on one database can be composed as a query on all internet-reachable databases whose graphs share common identifiers ("federated" queries), 3) the expansion of the underlying ontology can take place on-line, without refactoring the existing database.

A graph database is a type of database that uses graph structures to store, represent, and query data. In a graph database, data is organized as nodes (vertices) and relationships (edges) between those nodes. Nodes represent entities, while relationships capture the connections between them. This structure allows for rich modeling of complex relationships, making graph databases particularly well-suited for biological data.

In biology, graph databases are used to model and analyze various biological networks, such as protein-protein interaction networks, gene regulatory networks, metabolic networks, and more. By representing biological entities as nodes and their relationships as edges, graph databases enable efficient querying and traversal of these networks, facilitating the discovery of patterns, pathways, and interactions [110]. These advantages can be summarized as follows:

1. Flexibility: Graph databases allow for the dynamic addition and modification of nodes and relationships, making them adaptable to evolving biological data.
2. Efficient querying: Graph databases excel at traversing complex relationships, enabling fast and efficient querying of biological networks.
3. Relationship analysis: Graph databases enable the analysis of relationships between biological

entities, helping uncover hidden connections and patterns.

4. Scalability: Graph databases can handle large-scale biological data, making them suitable for storing and analyzing extensive datasets.

RDF (Resource Description Framework): RDF is a standard model for representing knowledge on the web. It provides a framework for describing resources and their relationships in a structured manner. RDF represents information as triples, consisting of subject-predicate-object statements. In the context of biology, RDF can be used to model and represent biological data, allowing for interoperability, integration, and querying across diverse data sources.

RDF is widely used in bioinformatics and biological research to represent and connect data from various domains such as genomics, proteomics, metabolomics, and ontologies. It enables the integration of data from multiple sources by providing a common framework for expressing relationships between biological entities. Efficient access to these resources is facilitated by:

1. Semantic interoperability: RDF provides a semantic framework for integrating data from different sources, enabling researchers to connect and query diverse biological datasets.
2. Linked data: RDF facilitates the creation of linked data by using URIs (Uniform Resource Identifiers) to uniquely identify resources, allowing for the establishment of connections and relationships between data.
3. Ontologies: RDF can represent ontologies, which are formalized vocabularies that define concepts and relationships in a specific domain. Ontologies play a crucial role in organizing and integrating biological data, enabling semantic querying and reasoning.

RDF data can be queried using the SPARQL query language, which allows researchers to retrieve and analyze specific information from RDF datasets.

2.2.6 BioPAX

BioPAX is a computer-based annotation framework that represents experimental knowledge of signaling-pathways in a format that is extendable, machine-readable and intuitive according to the common thematic distinction between high-level pathway architectures, reactions, reactants, and macromolecular classes ???. Within the BioPAX representation of a given pathway, structurally-detailed descriptions of complexes and complex-forming reactions are possible ???. Although the volume of experimental knowledge on biological signaling has outgrown existing curation efforts, online databases such as PathwayCommons provide unprecedented access to this knowledge in machine-readable form. Furthermore, BioPAX identifiers support the annotation of these pathways with structural, clinical, and evolutionary knowledge hosted by well-established structural databases such as UniProt, OMIM, and OntoDB.

Pathways in BioPAX [23] are heterogeneous graphs whose edges denote both conceptual (relational) knowledge and dynamics.

2.3 PCquery

2.3.1 Containerized RDF Endpoints for PathwayCommons and NextProt

The current implementation of PCquery is paired with a containerized instance of OpenLink Virtuoso Open-Source server (VOS), allowing fast, local deployment of NextProt and PathwayCommons RDF endpoints [55]. This not only allows greater control over support for federated queries, but also increases scalability in cases where high-bandwidth institutional access is required.

2.3.2 Pathway Mining and Reactant Imputation

Curated pathway databases, such as Reactome [29] and KEGG [52], are aggregated by the Pathway-Commons [98] initiative and ported to BioPAX, allowing PCquery to mine sets of related pathways using structured queries:

```
psParams = Dict(  
    "q"=>"notch1", :filter=>[["notch1", "nucleus"], ["notch1", "transcription"]],  
    "type"=>"pathway", "datasource"=>"reactome")
```

The above query will filter results through the PathwayCommons VOS relational endpoint, although a SPARQL query interface is also available to obtain the initial pathway set from the VOS RDF endpoint. Often, however, the biological pathway matching a given set of modeling criteria will have a fragmented representation within the database. An example of such a situation is the human Notch membrane pathway 2.2, and the human Notch nuclear pathway ?? . Despite the fact that these pathways both include the Notch intracellular complex, they are separate knowledge graphs from the standpoint of BioPAX.

As a backend to computational modeling, one key objective of PCquery is to generate a single knowledge graph encompassing the knowledge domain of the model. Many pathways are curated at the complex level, often reflecting the minimum unit of biological activity that could be used to validate experimental findings. However, in order to fit the model to the data, the knowledge graph must include assayable labels in its vertex annotations. In order to facilitate this, PCquery automatically expands complexes, creating nodes for complex forming reactions whose parents (individual proteins). In figure 2.4 below, this process is highlighted by the red edges for the human Notch membrane pathway.

2.3.3 Pathway Integration and Annotation

Individual pathways are integrated via common molecular identifiers. Once this integrated graph is constructed, PCquery traverses the graph and further labels all vertices with annotations from other online databases, including gene, transcript, and GO terms from NextProt [71] (figure 3.2b), and inter-species gene and transcript orthologs from OrthoDB [67]. Model-selection tasks requiring marker IDs from multiple contexts (e.g. cytosolic, nuclear) can be executed within the context of the integrated graph, as shown below for ligand-receptor interactions in the Notch pathway, including the transcriptional targets of Jag1 2.5.

Finally, a pathway from one species can be automatically mapped to another species by specifying a phylogenetic distance threshold and the resulting markers can be retrieved as annotations to the curated pathway. Figure 2.6 shows the average expression of all zebrafish homologs to the markers of the integrated human Notch pathway described above based on the scRNA-seq data of a single zebrafish neural crest cell.

2.4 A Unified Pipeline for Distributed Pathway Integration and Annotation

Two main approaches exist for the compilation and usage of biological signaling data: monolithic approaches [111, 49], and distributed approaches [63, 117]. Monolithic approaches include pipelines or tools that redistribute explicit biological pathway knowledge so that knowledge integration and annotation takes place during method development. Benefits of this approach include high reproducibility because the curation status of prior-knowledge sources is fixed to a specific release of the package. Furthermore, for pathway knowledge pipelines, packages like Omnipath [111] are straightforward to use can save significant computational and network resources related to integration. In the case of signaling pathway inference, monolithic tools like CellChat [49] go one step further so that inverse problems [20] can be solved based on the input of experimental data alone, based on a CCC model that is pre-configured based on protein-protein interactions mined from KEGG [54]. In contrast to monolithic approaches, distributed approaches for integration and annotation of prior-knowledge are pipelines that are run by the end user to acquire and manipulate pathway knowledge on-demand [63]. Additionally, there exist 'distributed analogs' to model-specific inference packages like CellChat, including the SoptSC package, which simultaneously offers expanded pipeline features, such as embedding and lineage analysis, while requiring the user to supply the gene labels of signaling pathway markers [117].

Although methods like CellChat [50] represent a turn-key approach to the use of curated pathway knowledge in CCC inference, there are several reasons why manipulation of the knowledge required to build such tools is also warranted: First, modeling extended pathway connectivity often requires integration of curated pathways, and annotation of their components. Such efforts support regularization of predicted interaction networks supported by models that take into account the dogma of biological signal transduction and gene expression, as in the SoptSC method, which takes into account the expression of down-stream transcriptional targets when considering putative ligand/receptor interaction [117]. Second, existing integration tools are

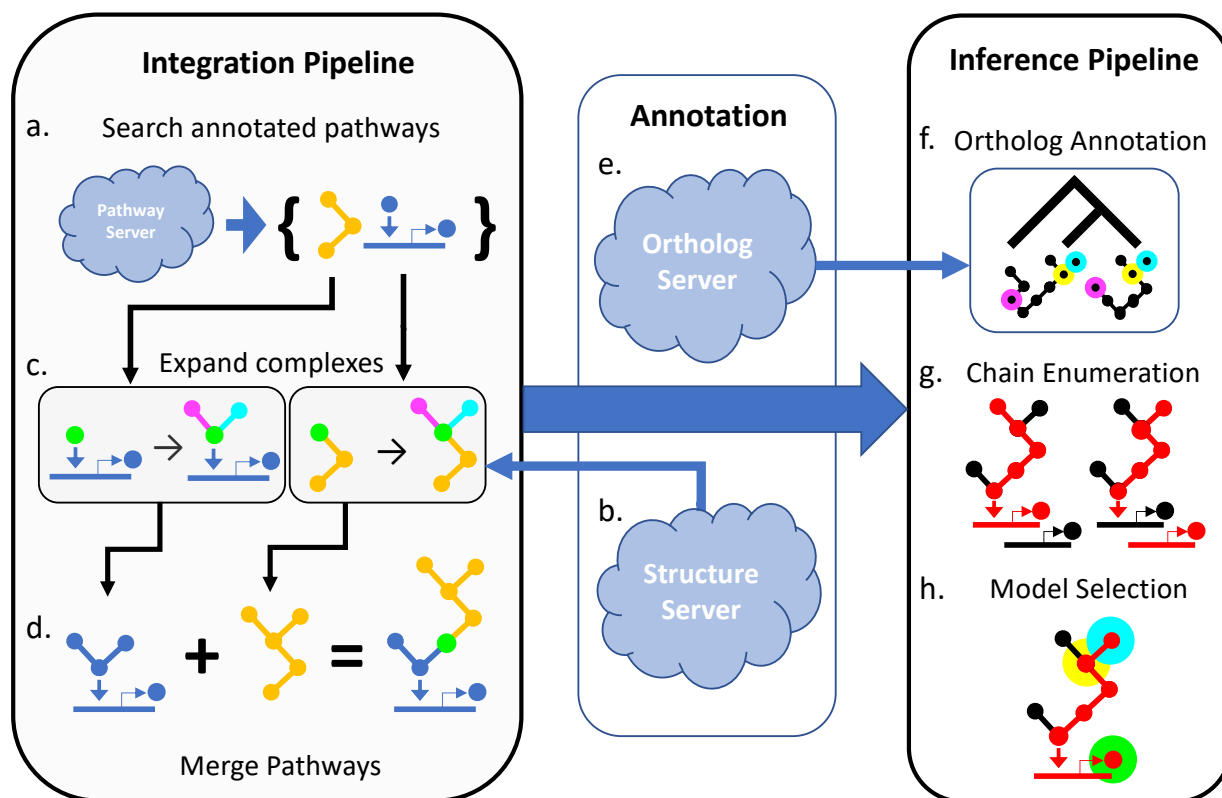


Figure 2.1: A) query BioPAX pathway repository based on organism and keyword search. B) for each annotated pathway (blue,yellow) recursively impute complex-forming reactions (RHS blue arrow) based on existing product annotations (green) and their BioPAX-annotated constituents (purple,blue). C) Integrate augmented pathways (blue, yellow) based on common molecular identifiers (green). D) Mine inter-species pathway markers based on sequence-homology between reference integrated pathway protein gene loci and the gene loci of the target species. E) Enumerate signaling trajectories for integrated pathways. Query gene regulatory interactions downstream of cell-surface interactions based on GO annotation of integrated pathways.

restricted to performing integration tasks, but don't directly support the annotation of knowledge graphs [64]. For tasks such as CCC inference, this annotation is a key step in utilizing pathway knowledge to decide model structure, since the mathematical representation of the objective for models like SoptSC requires explicit labeling of pathway roles such as "receptor", "ligand", "up target", and "down target" [117].

2.5 Figures and Tables

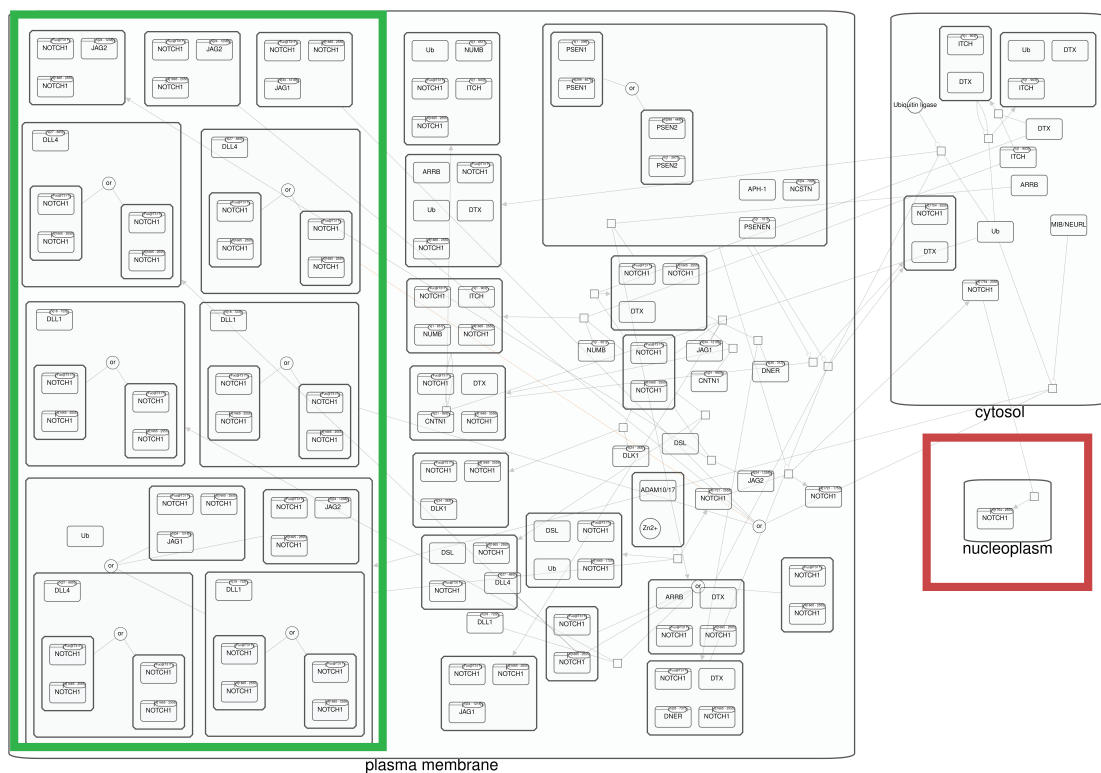


Figure 2.2: Notch1 Cytosolic Pathway
(Reactome HSA-2122948)

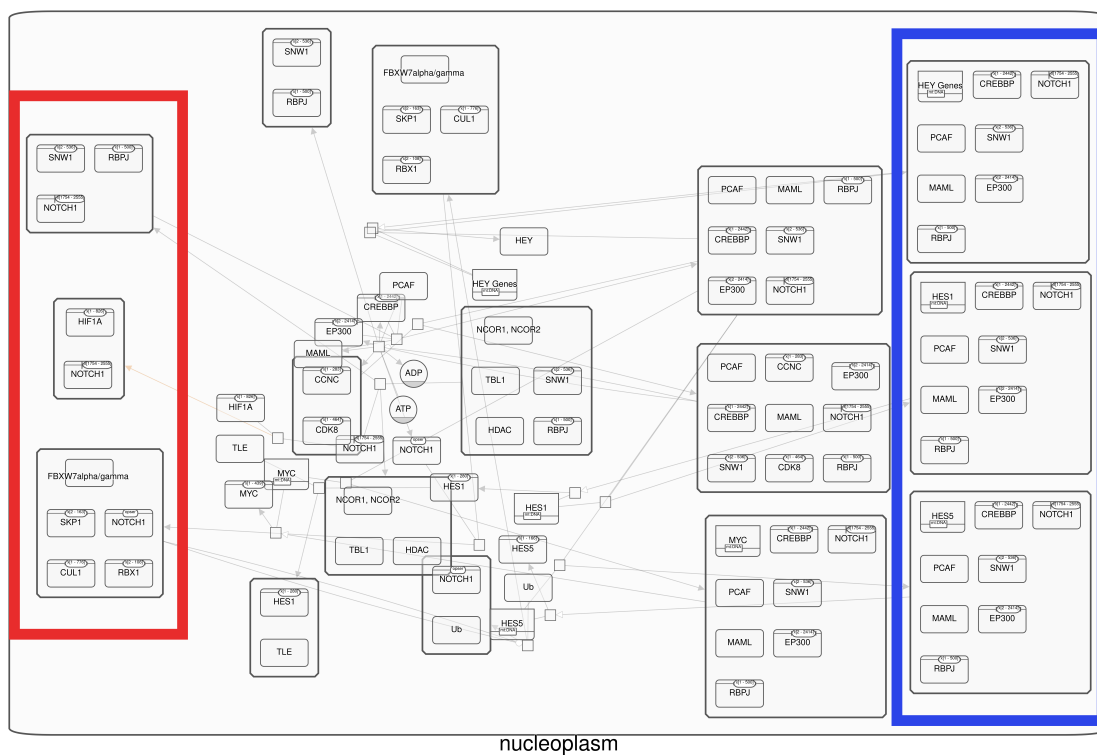


Figure 2.3: Notch1 Nuclear Pathway
(Reactome HSA-2122947)

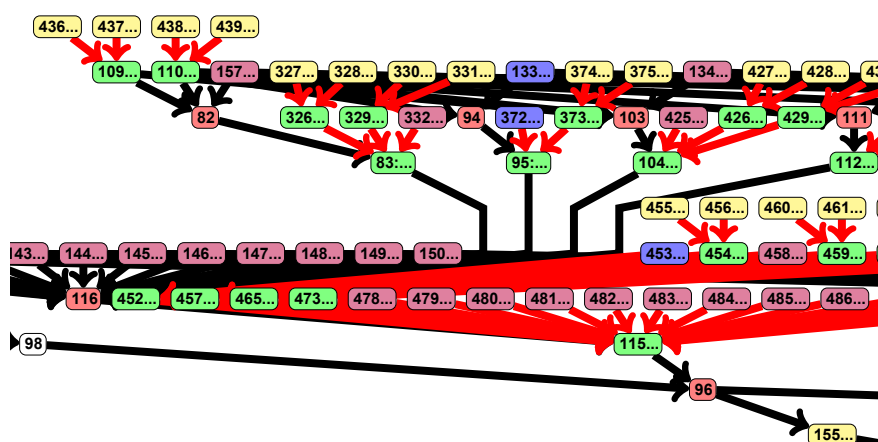


Figure 2.4: Expansion of complexes imputes complex-forming interactions and allows GO annotation of reactants to be mined from NextProt [71]

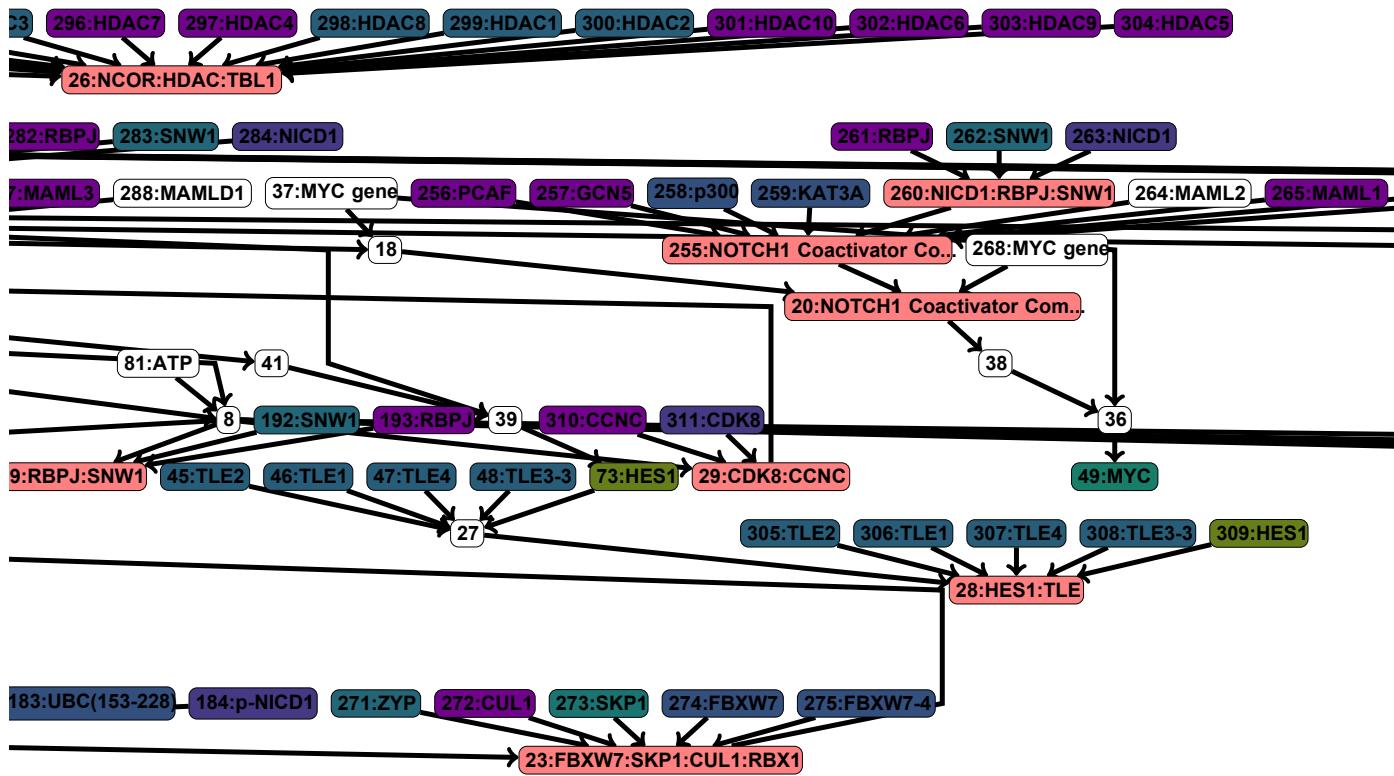


Figure 2.6: Mapping the gene expression of a single zebrafish neural crest cell to human orthologs in the integrated Notch1 pathway from human

Chapter 3

Single-Cell Multiomic scRNA/scATAC Integration

3.1 Introduction

Inverse problems refer to the task of using data to parameterize a model whose structure describes how a system behaves [20]. In biological inverse problems, the utilization of high-dimensional data (especially for high-throughput, single-cell modalities) is necessary to parameterize models that predict complex interactions [2]. These tasks have motivated the development and adoption of so-called "black box" machine-learning methods that operate efficiently on high-dimensional data [115, 9, 120]. The term "black box" refers to the fact that the flexibility of these methods comes at the cost of more interpretable, linear models [107, 9].

Single-cell multiomics allow heterogeneity in the co-regulation between distinct processes to be detected in a population of cells. In the case of gene transcription and chromatin accessibility, measured through scRNA/scATAC-seq, the relationship between the accessibility of a gene and its transcription can be quantified. In particular, the heterogeneity of coupling between these two processes can be assessed both individually and collectively. Recently, decoupling between accessibility and transcription has been investigated as a mechanism for epigenetic memory within populations of epithelial stem cells, with important consequences for wound healing and inflammatory disorders. In this chapter, I investigate a highly-interpretable

method for the investigation of such decoupling phenomena. I introduce a deep, variational factor model that learns a multimodal ATAC/RNA embedding where global accessibility changes are linearly related to the transcription of individual genes.

3.2 Single-Cell Multiomic Profiling of Chromatin Accessibility and Transcription

3.2.1 Single-Cell Multiomics

In contrast to gene expression, which is measured by transcription, chromatin accessibility refers to the ease with which the transcriptional machinery, including transcription factors and RNA polymerase, can access and interact with specific regions of DNA. In general, open or accessible chromatin is associated with active gene expression, while closed or inaccessible chromatin is associated with gene silencing or repression [3].

Single-cell multiomics allows the joint observation of modalities like RNA-seq and ATAC-seq in a single cell [114]. The integration of such measurements at the single-cell level, combined with analysis of unspliced transcripts, has led to a better understanding of both global and local regulatory changes in gene expression [8, 114].

3.2.2 Complementary Regulation of Accessibility and Transcription

Due to the commercial availability of single-cell ATAC/RNA sequencing [114], the integration of these measurements has shed light on mechanisms of lineage priming that involve decoupled accessibility and transcription [85]. This "chromatin potential" is characterized by accessibility which precedes gene expression at regulatory domains. This decoupling between accessibility and transcription reflects the fact that the regulation of chromatin accessibility and transcriptional regulation are complementary [59]. Mechanisms contributing to this separation include:

1. Regulatory elements: The presence of regulatory elements, such as enhancers or silencers, can influence gene expression independently of chromatin accessibility. These elements can modulate gene transcription by interacting with specific transcription factors or by forming long-range interactions with gene promoters [27].
2. Transcription factor binding: The binding of transcription factors to DNA can occur in regions of closed chromatin, allowing them to regulate gene expression without requiring changes in chromatin accessibility. Transcription factors can recruit co-activators or co-repressors to promote or inhibit gene transcription, respectively [60].
3. Epigenetic modifications: Chromatin accessibility is influenced by various epigenetic modifications, including DNA methylation and histone modifications. However, the effects of these modifications on gene transcription can be context-dependent [46].

3.3 Lineage Plasticity and Tissue Memory

3.3.1 Existence and Characterization of Memory in Epithelial Lineages

Differentiated epithelial cells determine the structure and function of organs and tissues [3]. Within these tissues, multiple lineages retain plasticity, able to alter their fate in response to injury or exposure to environmental factors, prompting a response that is characterized by homeostasis and regeneration [32]. However, these adaptive mechanisms also encompass forms of tissue memory characterized by transcriptionally silent alterations to chromatin in stem cells that is distinct from the chromatin state of naive stem cells [72, 91, 35].

3.4 Single-Cell Multiomics Integration and the Identification of Modality-Specific Features

Joint computational analysis of the modalities presented by single-cell multiomics provide the opportunity to explore both global and local regulation of cell lineages revealed by this data. Current approaches therefore

seek to learn a representation of cell state that unifies these modalities [114].

3.4.1 Paired vs Non-Paired Integration Techniques

Due to the continuing generation, and historic availability of single-modality data such as scRNA-seq, scATAC-seq before the wide availability of single-cell multiomics, computational methods for the analysis of multiomics data have focused on the integration of both multi-modal and single-modality data [106, 33, 119]. Accordingly, several types of integration have been proposed [7]: 1) horizontal integration involves the embedding of cells onto the same feature space, 2) vertical integration uses the cell as a common reference point when linking features from multiple modalities, while 3) diagonal integration allows the integration of multiple modalities where cells cannot directly serve as a reference point. In practice, the use of both multiome and single-modality data allows diagonal integration to proceed by horizontal integration of single-modality data with the corresponding features of cells for which multiomic measurements were collected, combined with the vertical integration of the two modalities [7].

3.4.2 Interpretable Integration and Modality-Separation

The training of machine learning models on single-cell transcriptomics and multiomics data has allowed complex patterns to be identified in this high-dimensional data, such that important down-stream tasks such as genotype/phenotype dependencies and drug interactions [42, 83, 124, 75]. The complexity of these predictive tasks has driven the adoption of machine learning models with state-of-the-art performance on predictive task benchmarks [120, 9]. This tradeoff between flexibility and interpretability is a classical problem a machine learning [26] whose existence motivates the separate evaluation of these criteria [107].

3.5 A Variational Factor Model for Detection and Interpretation of Global Decoupling

3.5.1 Single-Cell Embeddings and Lineage-Uncertainty

The discrete nature of cell lineage dynamics and the complexity of their regulation means that rare or poorly-differentiated states will always be problematic to observe, and therefore problematic to learn from data [86, 102]. In recognition of this, mathematical representations of single-cell data explicitly address the uncertainty regarding fate and cell identity [126, 103, 8, 82, 33].

3.5.2 Unsupervised Learning With Deep, Generative Models

Low-dimensional embeddings of high-dimensional data seek to identify a feature space where complex behavior in the high-dimensional space can be represented by more straightforward interaction between perceptual or functional features [74, 117, 51]. In the case of generative models, these low-dimensional features are the parameters of an unknown probability distribution whose estimation relies on posterior inference [89]. Such models aid the interpretation of single-cell data by addressing two previously-stated challenges: 1) sparse data with missing information, and 2) inability to directly observe transition state cells. In the case of (1), higher read-depth can be simulated by averaging the reconstructed features of individual embeddings [81]. In the case of (2), generative models equipped with flexible non-linear decoder networks have been shown simulate high-dimensional transitions through latent-space interpolation [83, 123, 11].

3.5.3 Variational Autoencoders in Single-Cell Data Analysis

Any probabilistic embedding of high-dimensional data must minimally express the high-dimensional observation x in terms of an unobserved embedding z whose prior distribution is $p(z)$, leading to a model of the form $p(z)p(x|z)$, whose posterior distribution (the embedding) $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$ involves the intractable marginal likelihood $\int p(z)p(x|z)dz$, and for which the dimensionality of the feature space precludes both mean-field approximation and MCMC [62]. This problem led to the adaptation of the denoising autoencoder whose middle layer is replaced by separate networks for the mean and variance of a standard Gaussian embedding

distribution z , and whose decoder network represents the conditional distribution of the data $p(x'|z)$. The key innovation of this model is to sample $z \sim p(\mu, \sigma)$ as a transformation of a second variable $\epsilon \sim N(0, 1)$, allowing both back-propegation and sampling-based reconstruction [61].

3.5.4 VAE-based Multiomic Integration Models

Currently, two forms of VAE-based model have been explored for the integration of multiomic single-cell data: the MVAE [122] whose product-of-experts integration strategy was implemented by Cobolt [33], and a conjoined VAE model whose embedding is formed by annealing separately parameterized single-modality latent spaces [8]. Both of these models perform vertical integration in the latent embedding z and rely on deep MLP-based decoder networks to reconstruct individual modalities from the multiomic embedding of each cell.

3.5.5 Representation Learning With Variational Factor Models

Existing adaptations of the VAE for single-cell data include both deep [81] and shallow [107] decoder models of scRNA-seq data. While sacrificing the non-linear flexibility of deep decoder models, the shallow decoder model learns an embedding whose dimensions are linear factors of the reconstruction, allowing direct interpretation of a cell’s position in the embedding with respect to the genes it expresses. Additionally, like the loadings of PCA components, the high-variance genes of each of the latent dimensions of the VAE-based scRNA-seq factor model can be easily identified. But unlike PCA, the VAE-based model directly represents discrete count-based sequencing data through a ZINB observation likelihood [107].

3.5.6 Augmenting the Multiome Sequencing with Unsupervised ATAC Clustering

In order to provide global accessibility to the model, the single-cell multiome data is augmented with unsupervised clustering of the ATAC modality. Since this model aims to explore unknown population-specific changes in the relationship between accessibility and transcription, the scale at which we must condition the decoder is also a parameter of the model. Rather than making assumptions about this population parameter,

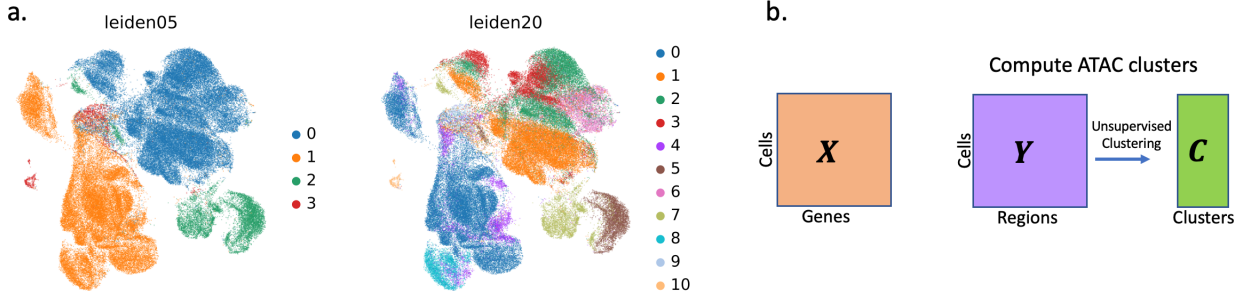


Figure 3.1: a) Unsupervised clustering of the ATAC modality. b) Augmentation of the paired single-cell multiome ATAC/RNA sequencing with ATAC clustering.

a general feature expansion approach is adopted so that the ATAC data is clustered at several scales (3.1a) and the labels concatenated into a label feature vector (3.1b).

3.5.7 A Multiomic VAE-Based Factor Model

I adapt the integration scheme of [8] to generate a multiomic embedding of paired single-cell data whose RNA and ATAC decoders are shallow single-layer networks. These decoders are conditioned on the a set of fully observed labels representing the global structure of the ATAC modality 3.2.

3.5.8 Synthetic Downsampling of PBMC scATC/RNA Multiome Data

Because the regulation of global decoupling is likely complex, a synthetic benchmark to detect this phenomenon was carried out by selectively down-sampling the RNA reads in cells from selected clusters (3.3a). The cluster-specific weight of $W_C^{(RNA)}$ was averaged across downsampled genes for downsampled clusters in each benchmarking batch and compared to the average non-downsampled gene weight across non-downsampled clusters (3.3b) and a two-tailed Student's T-test was performed on the means of the downsampled vs non-downsampled giving a test statistic of -4.078 ($p = 7.11e-5$), suggesting that a lower ATAC cluster weight for a given gene was strongly suggestive of suppressed RNA expression at that gene for cells in the corresponding ATAC cluster.

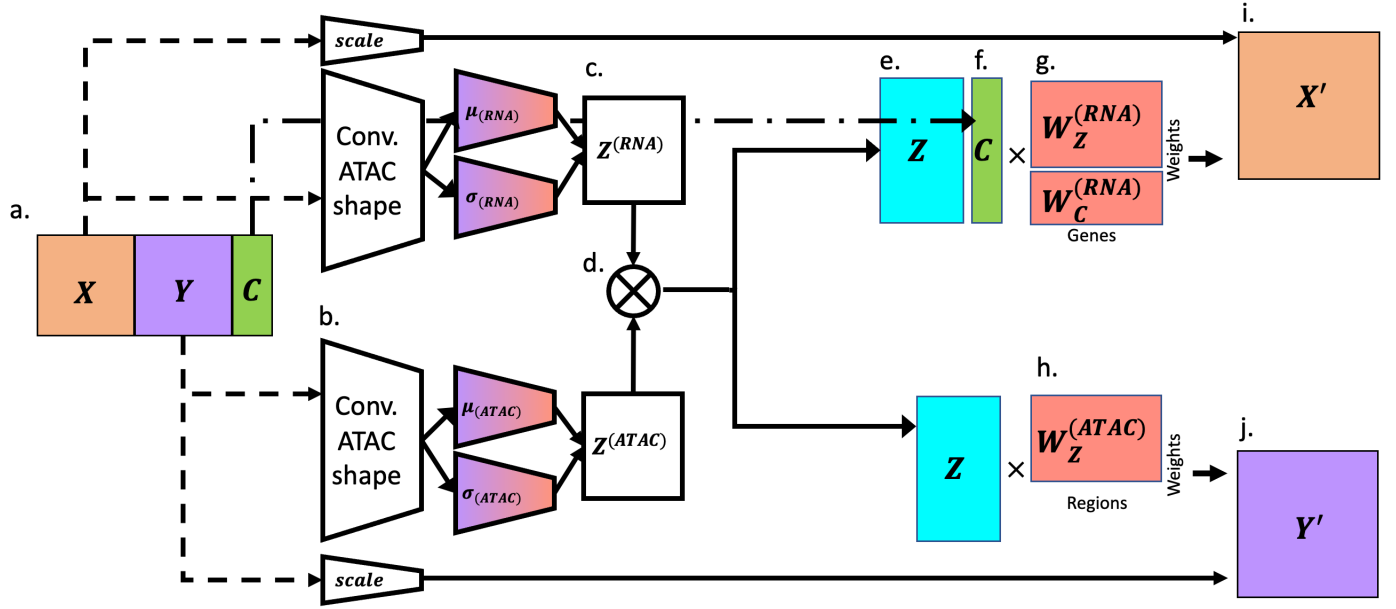


Figure 3.2: a) Input data augmented with ATAC clustering. b) Mean, variance, and scale encoders for the RNA and ATAC data. c) Single-modality embedding of the RNA. d) Annealing operator that averages the two single-modality embeddings. e) Multiomic embedding to be processed by the RNA decoder. f) ATAC clusters conditioning the RNA decoder weights. g) RNA decoder weights. h) ATAC decoder weights. i) RNA reconstruction. j) ATAC reconstruction

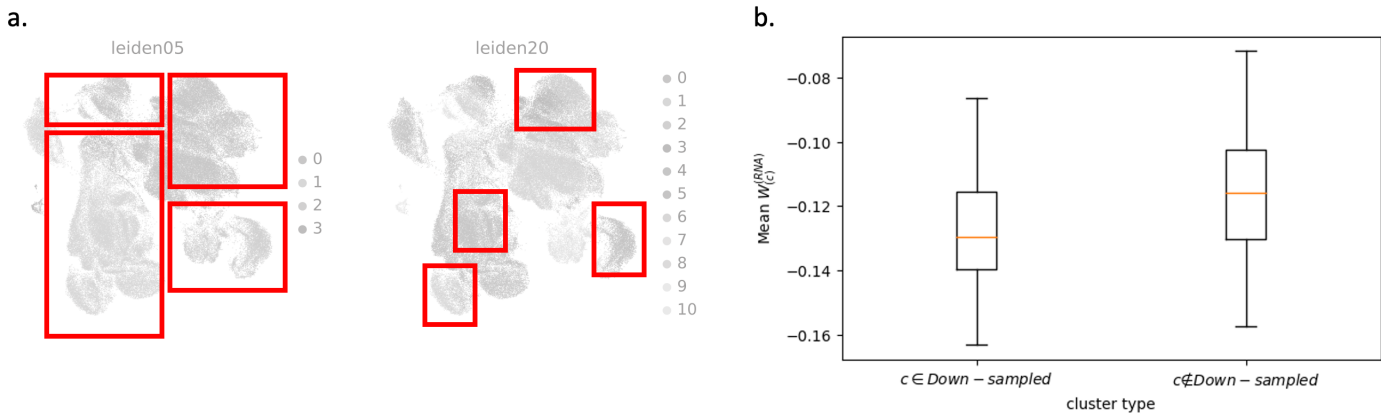


Figure 3.3: a) Cells from ATAC clusters labeled via unsupervised Leiden clustering were selected for RNA downsampling. b) Box plot prediction of synthetic decoupling in down-sampled cells based on $W_C^{(RNA)}$.

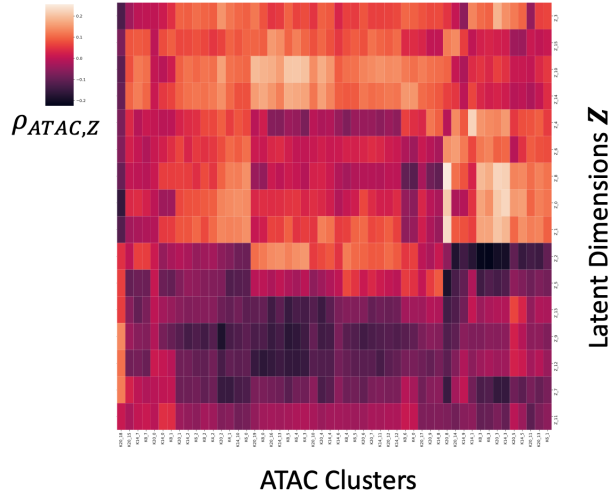


Figure 3.4: Correlation between ATAC Leiden cluster labels and the dimensions of the multiomic embedding for decoder weight sub-matrix $W_C^{(RNA)}$

3.5.9 Exploratory Analysis of the Decoder Weights

In order to investigate whether the model can identify meaningful relationships between global patterns in accessibility and the dimensions of the multiomic latent embedding Z , I analyzed PBMC data from cancer immunotherapy patients [13]. Epigenetic memory is a well-studied feature of the adaptive immune system [73]. I analyzed two complementary trends in the weight matrix of the RNA decoder: 1) the correlation between the weights corresponding to specific ATAC cluster labels and specific embedding dimensions and 2) the correlation between the weights of specific genes. The analysis in (1) revealed that embedding dimensions Z_2 and Z_5 represent significant heterogeneity in the scale of decoupling with all specific accessibility clusters from all three Leiden resolutions showing strong decrease in transcriptomic output for accessible genes 3.4. Likewise, there is significant heterogeneity in the decoupling rate among genes 3.5.

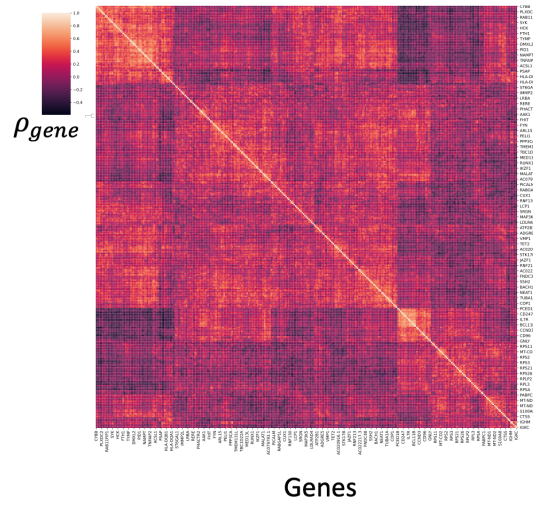


Figure 3.5: Correlation between genes for decoder weight sub-matrix $W_C^{(RNA)}$

Chapter 4

Discussion

In the preceding chapters, I have addressed several ways in which transformations and annotations can be applied to single-cell omics data to facilitate its analysis through machine learning approaches. A common theme in these tasks is the use of multiple data sources, representing the opportunity in many workflows to increase exploratory scope and robustness, while also introducing challenges associated with the annotation of data and various platform affects.

4.1 scRNA-seq Imputation

This chapter reviewed current approaches to scRNA-seq imputation and the use of deconvolution to provide enhanced perspective on bulk RNA-seq data. This chapter introduces DURIAN [57], an imputation method that utilizes bulk data to improve the specificity of minimum-rank matrix completion by constraining the objective with a pseudobulk loss. This scheme of utilizing an 'outer' imputation objective, constrained by an 'inner' deconvolution objective resembles classical data-augmentation in the context of EM [113]. In order to use bulk data to supervise the imputation, I augmented the objective with the pseudobulk representation of the single-cell data.

Addressing dropout in scRNA-seq is a longstanding problem that has been addressed both indirectly and

directly. Direct approaches can take an optimization-based objective, such as in DURIAN [57], SCRABBLE [93], or DrImpute [34], but may also take the form of a generative model, such as URSM [128] or scVI [81]. conversely by implicitly handling dropout within downstream analytical pipelines, such as cell-type identification [95]. There are benefits and drawbacks to both strategies. Direct imputation leads to bioinformatics pipelines that are more modular. This modularity is embraced by DURIAN in its deconvolution approach, where two very different inference schemes are used to estimate the cell-type mixture of the deconvolution.

4.2 Pathway Mining and Integration

In this chapter, I introduce an integrated pipeline accessing and integrating multi-scale prior-knowledge. The interpretability of PCquery makes it suitable for many downstream tasks, all falling under the umbrella of model parameterization, which represents a shift in the historic focus of quantitative biology. Due to the availability of high-dimensional data, the need to integrate domain-specific prior-knowledge is increasingly important. The integration of BioPAX pathways forms a scaffold upon which multiple levels of biological information can be systematically added. In order to facilitate integration of multiple pathways, the individual publicly-stored pathway graphs must be augmented by expansion of complex-forming reactions, allowing unique molecular identifiers to serve as the intersection for pathway integration.

PCquery relies on semantic web resources to provide continued access to public online databases. Semantic web databases explicitly represent knowledge as a tree, allowing database queries to implement logical operations on the pathway graphs, such as transitivity [?]. This framework allows the SPARQL client modules currently-included in PCquery to augment public curated pathways as described above by performing inference on pathway graphs. These modules also allow PCquery to more-effectively leverage structural databases to fully annotate pathways. However, there are also limitations to this approach, including the fact that endpoints supporting SPARQL are still less common among public bioinformatic resources than more traditional endpoints supporting only SQL.

4.3 Conditional Factor Models for Single-Cell Multi-omics

This chapter introduced a method for deep, generative integration of single-cell multiomics data in the case of paired RNA and ATAC sequencing. This model augments the multiome sequencing with a set of ATAC-specific labels. The use of single-cell multiome data to study the relationship between accessibility and transcription is motivated by the observation that the timescale of these processes is regulated dynamically, and that this regulation appears to have both global [76] and local [77] structure, and that decoupling of these processes plays a critical adaptive role in epithelial tissues [90].

In contrast to existing multiomics integration approaches, I prioritize the interpretability of the model regarding global inter-modality relationships in the latent space. This leads to an exploration of how gene-specific transcriptional affects can be predicted based on the global heterogeneity of the underlying accessibility landscape.

Bibliography

- [1] PCquery · JuliaHub. <https://juliahub.com/ui/Packages/PCquery/csULp/0.3.6>.
- [2] M. Alber, A. Buganza Tepole, W. R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W. W. Lytton, P. Perdikaris, L. Petzold, and E. Kuhl. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine*, 2(1):1–11, Nov. 2019.
- [3] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walter, J. Wilson, and T. Hunt. *Molecular Biology of the Cell*. Garland Science, Taylor and Francis Group, New York, NY, sixth edition edition, 2015.
- [4] A. A. Almet, Z. Cang, S. Jin, and Q. Nie. The landscape of cell–cell communication through single-cell transcriptomics. *Current Opinion in Systems Biology*, 26:12–23, June 2021.
- [5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.
- [6] T. Andl, S. T. Reddy, T. Gaddapara, and S. E. Millar. Wnt signals are required for the initiation of hair follicle development. *Developmental cell*, 2(5):643–653, 2002.
- [7] R. Argelaguet, A. S. E. Cuomo, O. Stegle, and J. C. Marionni. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10):1202–1215, Oct. 2021.
- [8] T. Ashuach, M. I. Gabitto, M. I. Jordan, and N. Yosef. MultiVI: Deep generative model for the integration of multi-modal data, Sept. 2021.
- [9] C. B. Azodi, J. Tang, and S.-H. Shiu. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics*, 36(6):442–455, June 2020.
- [10] M. Baron, A. Veres, S. Wolock, A. Faust, R. Gaujoux, A. Vetere, J. Ryu, B. Wagner, S. Shen-Orr, A. Klein, D. Melton, and I. Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360.e4, Oct. 2016.
- [11] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow. Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer. *arXiv:1807.07543 [cs, stat]*, July 2018.
- [12] D. M. Blei, M. I. Jordan, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
- [13] G. M. Boukhaled, R. Gadalla, H. J. Elsaesser, D. Abd-Rabbo, R. Quevedo, S. Y. C. Yang, M. Guo, B. X. Wang, B. Noamani, D. Gray, S. C. M. Lau, K. Taylor, K. Aung, A. Spreafico, A. R. Hansen, S. D. Saibil, N. Hirano, C. Guidos, T. J. Pugh, T. L. McGaha, P. S. Ohashi, A. G. Sacher, M. O. Butler, and D. G. Brooks. Pre-encoded responsiveness to type I interferon in the peripheral immune system defines outcome of PD1 blockade therapy. *Nature Immunology*, 23(8):1273–1283, Aug. 2022.

- [14] S. P. Brooks and A. Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, Dec. 1998. Publisher: Taylor & Francis.
- [15] J.-F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization*, 20(4):1956–1982, Jan. 2010. Publisher: Society for Industrial and Applied Mathematics.
- [16] Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1):2084, Apr. 2020.
- [17] 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*, 20(2):218–228, Feb. 2023.
- [18] M. Chen and X. Zhou. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biology*, 19(1):196, Nov. 2018.
- [19] S. Chowdhury and R. R. Sarkar. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database*, 2015:bau126, Jan. 2015.
- [20] G. Clermont and S. Zenker. The inverse problem in mathematical biology. *Mathematical Biosciences*, 260:11–15, Feb. 2015.
- [21] T. U. Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, Jan. 2019.
- [22] F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C. G. Giron, T. Genez, J. G. Martinez, C. Guizarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D. N. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, J. G. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.-M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T. A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. E. Hunt, G. R. Ilesley, J. E. Loveland, F. J. Martin, B. Moore, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, S. Dyer, P. W. Harrison, K. L. Howe, A. D. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, Jan. 2022.
- [23] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D’eustachio, C. Schaefer, J. Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.
- [24] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D’eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, Ö. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Ruebenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney,

- J. Collado-Vides, S. Goto, M. Hucka, N. Le Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, and G. D. Bader. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, Sept. 2010.
- [25] D. Dimitrov, D. Türei, M. Garrido-Rodriguez, P. L. Burmedi, J. S. Nagai, C. Boys, R. O. Ramirez Flores, H. Kim, B. Szalai, I. G. Costa, A. Valdeolivas, A. Dugourd, and J. Saez-Rodriguez. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications*, 13(1):3224, June 2022.
- [26] I. D. Dinov. Black Box Machine-Learning Methods: Neural Networks and Support Vector Machines. In I. D. Dinov, editor, *Data Science and Predictive Analytics: Biomedical and Health Applications Using R*, pages 383–422. Springer International Publishing, Cham, 2018.
- [27] N. Dogan, W. Wu, C. S. Morrissey, K.-B. Chen, A. Stonestrom, M. Long, C. A. Keller, Y. Cheng, D. Jain, A. Visel, L. A. Pennacchio, M. J. Weiss, G. A. Blobel, and R. C. Hardison. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics & Chromatin*, 8(1):16, Apr. 2015.
- [28] 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*, 15(4):1484–1506, Apr. 2020.
- [29] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.
- [30] P. Gaudet, P.-A. Michel, M. Zahn-Zabal, A. Britan, I. Cusin, M. Domagalski, P. D. Duek, A. Gateau, A. Gleizes, V. Hinard, V. Rech de Laval, J. Lin, F. Nikitin, M. Schaeffer, D. Teixeira, L. Lane, and A. Bairoch. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Research*, 45(D1):D177–D182, Jan. 2017.
- [31] D. L. Gibbs, B. Aguilar, V. Thorsson, A. V. Ratushny, and I. Shmulevich. Patient-Specific Cell Communication Networks Associate With Disease Progression in Cancer. *Frontiers in Genetics*, 12, 2021.
- [32] A. Gola and E. Fuchs. Environmental control of lineage plasticity and stem cell memory. *Current Opinion in Cell Biology*, 69:88–95, Apr. 2021.
- [33] B. Gong, Y. Zhou, and E. Purdom. Cobolt: Integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, 22(1):351, Dec. 2021.
- [34] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, 19(1):220, June 2018.
- [35] K. A. U. Gonzales, L. Polak, I. Matos, M. T. Tierney, A. Gola, E. Wong, N. R. Infarinato, M. Nikolova, S. Luo, S. Liu, J. S. S. Novak, K. Lay, H. A. Pasolli, and E. Fuchs. Stem cells expand potency and alter tissue fitness by accumulating diverse epigenetic memories. *Science*, 374(6571):eabh2444, Nov. 2021.
- [36] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, Apr. 2004.
- [37] C. F. Guerrero-Juarez, G. H. Lee, Y. Liu, S. Wang, M. Karikomi, Y. Sha, R. Y. Chow, T. T. L. Nguyen, V. S. Iglesias, S. Aasi, M. L. Drummond, Q. Nie, K. Sarin, and S. X. Atwood. Single-cell analysis of human basal cell carcinoma reveals novel regulators of tumor growth and the tumor microenvironment. *Science Advances*, 8(23):eabm7981, June 2022.

- [38] K. Gupta, J. Levinsohn, G. Linderman, D. Chen, T. Y. Sun, D. Dong, M. M. Taketo, M. Bosenberg, Y. Kluger, K. Choate, and P. Myung. Single-Cell Analysis Reveals a Hair Follicle Dermal Niche Molecular Differentiation Trajectory that Begins Prior to Morphogenesis. *Developmental Cell*, 48(1):17–31.e6, Jan. 2019.
- [39] J. A. Hansen, E. K. Ringger, and K. D. Seppi. Probabilistic Explicit Topic Modeling Using Wikipedia. In I. Gurevych, C. Biemann, and T. Zesch, editors, *Language Processing and Knowledge in the Web*, Lecture Notes in Computer Science, pages 69–82, Berlin, Heidelberg, 2013. Springer.
- [40] H. He, H. Suryawanshi, P. Morozov, J. Gay-Mimbrera, E. Del Duca, H. J. Kim, N. Kameyama, Y. Estrada, E. Der, J. G. Krueger, J. Ruano, T. Tuschl, and E. Guttman-Yassky. Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 145(6):1615–1628, June 2020.
- [41] C. Hennig. *fpc: Flexible Procedures for Clustering*, 2020. R package version 2.2-9.
- [42] L. Hetzel, S. Böhm, N. Kilbertus, S. Günnemann, M. Lotfollahi, and F. Theis. Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution.
- [43] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, Oct. 2018.
- [44] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.
- [45] P. O. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, 5(Nov):1457–1469, 2004.
- [46] Q. Hu, Z. Lian, X. Xia, H. Tian, and Z. Li. Integrated chromatin accessibility and DNA methylation analysis to reveal the critical epigenetic modification and regulatory mechanism in gonadal differentiation of the sequentially hermaphroditic fish, *Monopterus albus*. *Biology of Sex Differences*, 13(1):73, Dec. 2022.
- [47] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542, July 2018.
- [48] T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1):29, Feb. 2016.
- [49] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, and Q. Nie. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1):1088, Feb. 2021. Number: 1 Publisher: Nature Publishing Group.
- [50] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, and Q. Nie. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1):1088, Feb. 2021.
- [51] S. Jin, A. L. MacLean, T. Peng, and Q. Nie. scEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics*, 34(12):2077–2086, June 2018.
- [52] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl_1):D480–D484, 2007.

- [53] M. Kanehisa and P. Bork. Bioinformatics in the post-sequence era. *Nature Genetics*, 33(3):305–310, Mar. 2003.
- [54] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, Jan. 2017.
- [55] M. Karikomi. PathwayCommons v12 on Virtuoso 7, Feb. 2023.
- [56] M. Karikomi. PCquery, Jan. 2023.
- [57] M. Karikomi, P. Zhou, and Q. Nie. DURIAN: An integrative deconvolution and imputation method for robust signaling analysis of single-cell transcriptomics data. *Briefings in Bioinformatics*, 23(4):bbac223, July 2022.
- [58] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, July 2014.
- [59] K. Kiani, E. M. Sanford, Y. Goyal, and A. Raj. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. *Molecular Systems Biology*, 18(9):e10979, Sept. 2022.
- [60] K. Kiani, E. M. Sanford, Y. Goyal, and A. Raj. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. *Molecular Systems Biology*, 18(9):e10979, Sept. 2022.
- [61] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014.
- [62] D. P. Kingma and M. Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.
- [63] F. Kramer, M. Bayerlová, and T. Beißbarth. R-Based Software for the Integration of Pathway Data into Bioinformatic Algorithms. *Biology*, 3(1):85–100, Mar. 2014.
- [64] F. Kramer, M. Bayerlová, F. Klemm, A. Bleckmann, and T. Beißbarth. rBiopaxParser—an R package to parse, modify and visualize BioPAX data. *Bioinformatics*, 29(4):520–522, Feb. 2013.
- [65] F. Kramer and T. Beißbarth. Working with Ontologies. In J. M. Keith, editor, *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*, Methods in Molecular Biology, pages 123–135. Springer, New York, NY, 2017.
- [66] E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, and E. M. Zdobnov. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1):D807–D811, Jan. 2019.
- [67] E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, and E. M. Zdobnov. Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*, 47(D1):D807–D811, 2019.
- [68] M. P. Kumar, J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. C. Drummond, D. A. Lauffenburger, and A. Raue. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Reports*, 25(6):1458–1468.e4, Nov. 2018.
- [69] A. T. L. Lun, K. Bach, and J. C. Marionni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, Apr. 2016.

- [70] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.
- [71] L. Lane, G. Argoud-Puy, A. Britan, I. Cusin, P. D. Duek, O. Evalet, A. Gateau, P. Gaudet, A. Gleizes, A. Masselot, et al. nextprot: a knowledge platform for human proteins. *Nucleic acids research*, 40(D1):D76–D83, 2012.
- [72] S. B. Larsen, C. J. Cowley, S. M. Sajjath, D. Barrows, Y. Yang, T. S. Carroll, and E. Fuchs. Establishment, maintenance, and recall of inflammatory memory. *Cell Stem Cell*, 28(10):1758–1774.e8, Oct. 2021.
- [73] C. M. Lau, N. M. Adams, C. D. Geary, O.-E. Weizman, M. Rapp, Y. Pritykin, C. S. Leslie, and J. C. Sun. Epigenetic control of innate and adaptive immune memory. *Nature Immunology*, 19(9):963–972, Sept. 2018.
- [74] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct. 1999.
- [75] H. Lee and J. D. Welch. MorphNet Predicts Cell Morphology from Single-Cell Gene Expression, Oct. 2022.
- [76] C. Levra Levron, M. Watanabe, V. Proserpio, G. Piacenti, A. Lauria, S. Kaltenbach, A. Tamburrini, T. Nohara, F. Anselmi, C. Duval, L. Elettrico, D. Donna, L. Conti, D. Baev, K. Natsuga, T. Hagai, S. Oliviero, and G. Donati. Tissue memory relies on stem cell priming in distal undamaged areas. *Nature Cell Biology*, 25(5):740–753, May 2023.
- [77] C. Li, M. C. Virgilio, K. L. Collins, and J. D. Welch. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature Biotechnology*, 41(3):387–398, Mar. 2023.
- [78] W. V. Li and J. J. Li. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1):997, Mar. 2018.
- [79] P. Lin, M. Troup, and J. W. K. Ho. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, 18(1):59, Mar. 2017.
- [80] W.-C. Lo, C.-S. Chou, K. Gokoffski, F. Wan, A. Lander, A. Calof, and Q. Nie. Feedback regulation in multistage cell lineages. *Mathematical Biosciences and Engineering*, 6(1):59–82, Dec. 2008.
- [81] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, Dec. 2018. Number: 12 Publisher: Nature Publishing Group.
- [82] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, Dec. 2018.
- [83] M. Lotfollahi, F. A. Wolf, and F. J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, Aug. 2019.
- [84] M. D. Luecken and F. J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.
- [85] S. Ma, B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, T. Law, C. Lareau, Y.-C. Hsu, A. Regev, and J. D. Buenrostro. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.e20, Nov. 2020.

- [86] A. L. MacLean, T. Hong, and Q. Nie. Exploring intermediate cell states through the lens of single cells. *Current Opinion in Systems Biology*, 9:32–41, June 2018.
- [87] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [88] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, and A. R. Leach. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, Jan. 2019.
- [89] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [90] S. Naik and E. Fuchs. Inflammatory memory and tissue adaptation in sickness and in health. *Nature*, 607(7918):249–255, July 2022.
- [91] S. Naik, S. B. Larsen, N. C. Gomez, K. Alaverdyan, A. Sendoel, S. Yuan, L. Polak, A. Kulukian, S. Chai, and E. Fuchs. Inflammatory memory sensitizes skin epithelial stem cells to tissue damage. *Nature*, 550(7677):475–480, Oct. 2017.
- [92] T. Peng, Q. Zhu, P. Yin, and K. Tan. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biology*, 20(1):88, May 2019.
- [93] T. Peng, Q. Zhu, P. Yin, and K. Tan. SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biology*, 20(1):88, May 2019.
- [94] E. Pierson and C. Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, Nov. 2015.
- [95] P. Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 11(1):1169, Mar. 2020.
- [96] S. Reddy, T. Andl, A. Bagasra, M. M. Lu, D. J. Epstein, E. E. Morrissey, and S. E. Millar. Characterization of wnt gene expression in developing and postnatal hair follicles and identification of wnt5a as a target of sonic hedgehog in hair follicle morphogenesis. *Mechanisms of development*, 107(1-2):69–82, 2001.
- [97] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
- [98] I. Rodchenkov, O. Babur, A. Luna, B. A. Aksoy, J. V. Wong, D. Fong, M. Franz, M. C. Siper, M. Cheung, M. Wrana, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497, 2020.
- [99] C. E. Rodríguez and S. G. Walker. Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, Jan. 2014.
- [100] D. Rohde and O. Cappé. Online maximum-likelihood estimation for latent factor models. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 565–568, June 2011. ISSN: 2373-0803.
- [101] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.
- [102] Y. Sha, D. Haensel, G. Gutierrez, H. Du, X. Dai, and Q. Nie. Intermediate cell states in epithelial-to-mesenchymal transition. *Physical Biology*, 16(2):021001, Jan. 2019.

- [103] Y. Sha, S. Wang, F. Bocci, P. Zhou, and Q. Nie. Inference of intercellular communications and multilayer gene-regulations of epithelial–mesenchymal transition from single-cell transcriptomic data. *Frontiers in genetics*, 11:1700, 2021.
- [104] R. Stark, M. Grzelak, and J. Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, Nov. 2019.
- [105] L. D. Stein. Integrating biological databases. *Nature Reviews Genetics*, 4(5):337–345, May 2003.
- [106] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019.
- [107] V. Svensson, A. Gayoso, N. Yosef, and L. Pachter. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, June 2020.
- [108] D. Tatarakis, Z. Cang, X. Wu, P. P. Sharma, M. Karikomi, A. L. MacLean, Q. Nie, and T. F. Schilling. Single-cell transcriptomic analysis of zebrafish cranial neural crest reveals spatiotemporal regulation of lineage decisions during development. *Cell Reports*, 37(12):110140, Dec. 2021.
- [109] F. J. Theis, K. Stadthanner, and T. Tanaka. First results on uniqueness of sparse non-negative matrix factorization. In *2005 13th European Signal Processing Conference*, pages 1–4, Sept. 2005.
- [110] S. Timón-Reina, M. Rincón, and R. Martínez-Tomás. An overview of graph databases and their applications in the biomedical domain. *Database*, 2021:baab026, Sept. 2021.
- [111] D. Türei, A. Valdeolivas, L. Gul, N. Palacio-Escat, M. Klein, O. Ivanova, M. Ölbei, A. Gábor, F. Theis, D. Módos, T. Korcsmáros, and J. Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, Mar. 2021.
- [112] K. Van den Berge, F. Perraudeau, C. Soneson, M. I. Love, D. Risso, J.-P. Vert, M. D. Robinson, S. Dudoit, and L. Clement. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biology*, 19(1):24, Feb. 2018.
- [113] D. A. van Dyk and X.-L. Meng. The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, Mar. 2001.
- [114] K. Vandereyken, A. Sifrim, B. Thienpont, and T. Voet. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, pages 1–22, Mar. 2023.
- [115] A. Wagner, A. Regev, and N. Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, Nov. 2016.
- [116] A. Wagner, A. Regev, and N. Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, Nov. 2016. Number: 11 Publisher: Nature Publishing Group.
- [117] 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*, 47(11):e66–e66, June 2019.
- [118] X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):1–9, Jan. 2019.
- [119] J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, June 2019.

- [120] H. Wen, J. Ding, W. Jin, Y. Wang, Y. Xie, and J. Tang. Graph Neural Networks for Multimodal Single-Cell Data Integration. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4153–4163, Aug. 2022.
- [121] A. K. Wong, R. S. G. Sealfon, C. L. Theesfeld, and O. G. Troyanskaya. Decoding disease: From genomes to networks to phenotypes. *Nature Reviews Genetics*, 22(12):774–790, Dec. 2021.
- [122] M. Wu and N. Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [123] H. Yu and J. D. Welch. MichiGAN: Sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome Biology*, 22(1):158, May 2021.
- [124] H. Yu and J. D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations, July 2022.
- [125] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, Sept. 2017.
- [126] P. Zhou, S. Wang, T. Li, and Q. Nie. Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nature Communications*, 12, 2021.
- [127] L. Zhu, J. Lei, B. Devlin, and K. Roeder. A unified statistical framework for single cell and bulk RNA sequencing data. *The Annals of Applied Statistics*, 12(1):609–632, Mar. 2018.
- [128] L. Zhu, J. Lei, B. Devlin, and K. Roeder. A unified statistical framework for single cell and bulk RNA sequencing data. *The Annals of Applied Statistics*, 12(1):609–632, Mar. 2018.