Integration and disentanglement of single-cell and spatial transcriptomics in health and disease

Novella Rausell, C.

6

DISCUSSION

THE integration of diverse datasets in single-cell and spatial transcriptomics is essential for advancing our understanding of complex biological systems and disease mechanisms. In each chapter, we have explored different facets of these challenges whilst leveraging machine learning for robust cross-study and cross-technology integration. Additionally, we developed a novel computational tool to tackle representation disentanglement in single-cell datasets, demonstrating the practical applications and impact of these advancements in real-world biological datasets. This discussion will address the challenges and opportunities in cross-study reference building, including standardizing quality control, leveraging biological information, and improving reference atlases through multi-modal data integration and foundational models.

6.1 CROSS-STUDY REFERENCE BUILDING

Upon generation of a new single-cell dataset, the identification of cell types is one of the first tasks. Unsupervised clustering assigns cells to groups of similar omic profiles, which are then annotated as cell types. This process has been widely reported to be subjective and time-intensive, ultimately leading to biased annotations. To address these challenges, the construction of harmonised references aims to establish cellular identities that can be generalised. In Chapter 2, we introduced a novel reference atlas, the Mouse Kidney Atlas (MKA), in which we used unsupervised learning to integrate single-cell datasets from several studies and supervised learning and manual curation to harmonise annotations across these. The main usage of our atlas is to perform annotation used as a reference[267] or leverage our trained model to do transfer learning[127] and contextualise disease datasets. In our own work (Chapter 3), we extended the MKA to create a comprehensive reference of ADPKD-associated cell states, enabling the deconvolution of spatial transcriptomics data from diseased tissue. Additionally, the MKA was used to automatically annotate the cells of a mouse model of acute kidney injury, in which they identified that impaired hematopoiesis potential (i.e., the ability of progenitor cells to develop into various types of immune cells) is associated with worse outcomes after kidney injury via an aberrant renal macrophage inflammatory response[269].

6.1.1 STANDARDIZING QUALITY CONTROL

Quality control in single-cell RNA sequencing is a critical step in the analysis, with approaches varying across studies, technologies, and tissues. Distinguishing between genuine cellular signals and technical artifacts is a fundamental challenge that can obscure true cell-type-specific expression. Current best practices recommend applying different thresholds for each sample due to inter-sample variation. This variability in quality control standards poses a significant challenge when integrating data from multiple sources, as was the case in our study.

In Chapter 2, we approached this challenge by applying the same quality filters as those

used in the original publications to remove low-quality cells and nuclei. This decision was made to maintain consistency with the source data, but it highlights the broader issue of varying quality control standards between studies, technologies, and tissues. For instance, the kidney's high metabolic activity makes mitochondrial read content biologically relevant, not necessarily indicative of poor-quality cells. Stringent thresholds might remove important populations, whereas lenient thresholds could include low-quality cells, leading to inconsistent filtering criteria that mask cellular identities and complicate accurate cell type identification[270].

A complementary approach during quality control is to use computational doublet detection algorithms[271–273]. However, these require careful parameter optimization on a dataset-by-dataset basis[274,275]. The variability in doublet rates across different experimental conditions and the lack of ground truth labels in most datasets make doublet detection a persistent challenge in scRNA-seq analysis.

Recently, intronic content (i.e., reads mapping to intronic regions, indicating transcriptional activity and intact cells or nuclei) has been proposed as a new quality metric to distinguish between single cells and poor-quality cells or ambient RNA[273,276]. Compared to other quality metrics such as mRNA content, which can be confounded with biological variability, intronic content provides a more consistent measure of cell integrity. However, this metric itself requires careful fine-tuning on a dataset-by-dataset basis when building large cross-dataset atlases. This process is time-consuming and shares the subjectivity issues inherent in cell type annotation.

Alternatively, a machine learning model could be trained to confidently detect low-quality cells across a variety of scenarios. Such a model would distinguish between high-quality and low-quality cells based on multiple features derived from the scRNA-seq data. These features could include traditional quality metrics like the number of unique genes detected, total UMI counts, and mitochondrial read content, as well as more advanced metrics like the proportion of intronic reads or the ratio of unspliced to spliced transcripts. Training such a classifier presents its own challenges, primarily in obtaining reliable labels for low-quality cells. One approach could be to use positive-unlabeled (PU) learning[277], where we have a set of cells confidently labelled as high-quality (positive) and a larger set of unlabeled cells that may contain both high-quality and low-quality cells. PU learning algorithms can learn to distinguish between classes even when only one class is reliably labelled. Alternatively, we could use simulation-based approaches to generate synthetic low-quality cells, or leverage expert-annotated datasets.

6.1.2 LEVERAGING BIOLOGICAL INFORMATION TO ENHANCE CROSS-STUDY INTEGRATION AND CELL TYPE IDENTIFICATION

In theory, a reference containing as many cells as possible from diverse origins can promote a highly generalisable cell type ontology. However, significant technical and biological

differences between organs (e.g., kidney and brain), organisms (e.g., mice and humans) and starting material (e.g., single-cell and single-nuclei) make this task computationally challenging. While technical differences can be partially solved by state-of-the-art methods[51,59,267,278], significant batch effects require regularisation techniques that enhance batch correction in baseline models. For example, in VAE-based models, tuning the weight of the Kullback-Leibler (KL) divergence in the training loss[279] or using adversarial training[280,281] can help the model learn a batch-corrected space even with significant technical effects, albeit at the cost of losing biological conservation (i.e., ground truth cell types mix in the latent space). This can be helpful in single-study scenarios, in which minimising technical variability can help accurately identify cell populations. When building a reference, however, obtaining a universal cellular abstraction necessitates balancing the reduction of technical noise with the preservation of true biological diversity, enabling robust cross-study comparisons and annotation transfer.

More recently, an expressive prior over the latent space has been suggested to improve biological conservation[282]. Combining regularisation techniques such as adversarial training with this prior has been shown to outperform state-of-the-art unsupervised integration methods when significant batch effects are present[283]. Furthermore, in cross-study reference building, labels are usually available and can be harmonised[34,35]. In Chapter 2, we demonstrated that using these harmonised labels as input for a semi-supervised method (scANVI[59]) significantly improves large cross-study integration performance. Further, we showed that leveraging this biological information to tune hyperparameters with an objective designed to maximise cell type information can significantly enhance performance over baseline models. Ideally, reference-building methods should use available labels in supervised and semi-supervised learning to enhance model accuracy and robustness. By incorporating harmonised cell type labels as priors, the integration process can better account for biological variability and produce more general representations.

6.1.3 IMPROVING REFERENCE ATLASES

While many atlasing efforts have primarily utilised scRNA-seq datasets due to their abundance and relative ease of measurement, it is important to recognize that comprehensive cell atlases increasingly incorporate multiple data types. These can include epigenomic, proteomic, and spatial transcriptomic data, among others. Nevertheless, transcriptomic data remains the most common, serving as the foundation for many large-scale atlasing projects. However, transcriptomes often offer a limited view of a cell's identity. For example, proximal tubules are the most abundant population in the kidney and can be split into three segments—S1, S2, and S3—that display segment-specific expression of transporters. These transporters are lowly expressed, and differences between the segments are hard to capture in scRNA-seq. In Chapter 2, we highlighted how annotating these cell types is a major bottleneck for newly generated datasets, with some authors annotating the three segments at full resolution (i.e., S1, S2, and S3), and others using a more conservative approach by annotating mixed populations such as S1/2 or Proximal Convoluted Tubule (PCT).

Although we performed extensive manual curation of such cell types at full resolution in the MKA, it has been recently shown that the chromatin accessibility patterns surrounding segment-specific transporters can help the identification of the different segments in a kidney multi-omic dataset (i.e., scRNA-seq and scATAC-seq)[284]. This evidence underscores the importance of building references with models capable of handling multiple modalities, as these can significantly improve the learning of cellular identities. However, since these models are often used to transfer annotations to new single-modality datasets, both transfer learning models and baseline models must handle missing modalities. An alternative approach is to disentangle the learned identities into modality-specific spaces, which can be leveraged in single-modality transfer learning scenarios.

To further improve reference atlases, foundational models pre-trained on large collections of data present a promising avenue. These models[285–287] can be fine-tuned for a variety of tasks, including cell type classification, making them versatile references for transferring annotations. However, a recent benchmark study found that simpler models, such as logistic regression, often perform comparably to these advanced models, particularly when high-quality labelled data is available[288]. The lack of standardised quality criteria across studies and technologies (as mentioned in Section 5.1.2) makes foundational models less reliable for consistent performance, as they are likely pre-trained on datasets that include poor quality data. Overall, these models will benefit from large-scale (automatic) data curation and harmonisation in two ways: (i) creating standardised cross-tissue datasets (e.g., ImageNet[289] in the computer vision field) to benchmark these models against and (ii) reducing the amount of poor-quality data introduced in the pre-training phase, improving downstream performance.

6.2 CELL-TO-CELL COMMUNICATION

In Chapter 3, we analysed cell-to-cell communication (CCC) patterns around cystic regions in a spatial transcriptomics dataset of a Polycystic Kidney Disease (PKD) mouse model. Using Visium technology, which captures transcriptional signals from spots containing approximately 20 cells, we studied how cysts influence their surrounding tissue environment. We leveraged factorization methods[68] to decompose the ligand-receptor signal into factors of variation across layers, cysts, and cell types. We then used the loadings of these factors to compare CCC events between different cystic layers.

6.2.1 DETECTING CHANGES IN COMMUNICATION

A significant challenge in the field of cell-cell communication (CCC) analysis is the need to compare communication patterns across multiple samples or conditions. This is crucial for understanding how cellular interactions change in different contexts, such as disease states or developmental stages. However, achieving this comparison while maintaining cellular resolution and avoiding bias presents several difficulties.

Current solutions often rely on factorization methods to address this challenge. These methods summarize (e.g., average) the ligand-receptor signal per cell type, allowing for cross-sample comparisons. However, this approach comes with notable limitations. By aggregating signals at the cell type level, factorization methods lose cellular resolution, potentially obscuring important heterogeneity within cell populations. Additionally, when applied to specific Regions of Interest (ROIs), these methods might lack generalizability and miss broader context or interactions. The selection of ROIs, often driven by specific hypotheses, can also introduce confirmation biases.

To overcome these limitations, we need approaches that can incorporate cellular resolution while enabling cross-sample comparisons in an unbiased, whole-tissue manner. Some methods have been developed that make use of cellular resolution and spatial contexts to infer communication events[65,67,290,291], but these are typically limited to single samples, making comparative analyses difficult. Developing methods that can detect and compare CCC patterns across multiple regions and conditions while maintaining cellular resolution would enhance our ability to generate more robust and generalizable findings. Such approaches would allow for a more comprehensive understanding of how cell-cell communication dynamics change across different biological contexts, ultimately leading to more accurate and insightful analyses of complex cellular systems.

6.2.2 mRNA-DERIVED CCC IS INSUFFICIENT

Cell-to-cell communication methods assume that mRNA levels of ligand and receptor genes are good proxies of their protein levels. The ratio between transcript and protein levels is organ-specific and depends on post-transcriptional and post-translational modifications as well as translation efficiency[292,293]. In the kidney, the Pearson correlation between a small subset of transcripts and their absolute protein levels is 0.56[294]. Proteome- and transcriptome-wide analyses lack absolute protein quantification, but several studies estimate the correlation to be in the 0.3–0.6 range for a variety of tissues[295]. This indicates that using transcript levels to estimate cell communication events is not entirely reliable. Consequently, interpreting results from these methods typically requires protein-level validation.

Mass spectrometry-based spatial proteomics[296] promises to bridge this gap by measuring protein abundances with cellular resolution in cell lines and tissues. However, this technology depends on laser capture microdissection of single cells, which destroys the tissue and prevents any further post-processing. Ideally, the paired measurement of transcripts and proteins in tissue would allow us to constrain the CCC events to interactions for which both the receptor and ligand are detected at the protein level. However, this poses several computational challenges, such as inconsistent cellular distributions and morphology between consecutive slides and alignment of unpaired measurements (i.e., proteomics and transcriptomics from different but related locations).

6.3 DISENTANGLEMENT

Deep representation learning aims to identify meaningful hidden features from observed data. Crucially, these features should be able to reconstruct the data through a generative process. To fully leverage such models in a biological context, it is desirable that the learned features correspond to non-arbitrary explanatory factors in the data[30]. We refer to these as *generative factors* from now on. For example, if we study a tissue in which the phenotype is only influenced by the size and shape of the cells, the disentangled deep representation of such tissue should capture variation only in these attributes. In practice, however, biological systems are several orders of magnitude more complex, so baseline deep representation learning models often capture confounding and correlated factors. This leads to features that do not align with true biological variation. In Chapter 4, we introduced *spVIPES*, a model for the supervised learning of disentangled representations from scRNA-seq datasets.

6.3.1 USING DISENTANGLED REPRESENTATIONS

Disentangled representations are more robust, generalizable, and explainable[297]. This is because their features are intended to be independent and sparse, potentially allowing for a better understanding of the underlying process. One potential application of such representations could be improved generalisation to out-of-distribution samples. For instance, in a scRNA-seq classifier, disentangled representations might more effectively identify and distinguish different cell types or states, including when presented with new or rare cell populations that were not part of the training data. However, more research is needed to conclusively demonstrate these benefits in practical biological applications.

Additionally, by isolating distinct factors of variation, disentangled representations facilitate counterfactual analysis. In scRNA-seq, this means generating hypothetical scenarios such as predicting how a cell's gene expression profile would change under different conditions. By separating dataset-specific and shared features, *spVIPES* could be used to predict gene expression changes of healthy cells in disease conditions. Once a model is trained, new cells can be fed into the model while keeping the network components corresponding to the shared representation frozen (i.e., not updating weights and biases) and only training the parameters of the network responsible for learning the disease-specific latent space.

6.3.2 UNSUPERVISED DISENTANGLEMENT OF DATASETS WITH UNPAIRED FEATURES

By design, *spVIPES* assumes no correspondence between the features from each dataset, a concept known as *unpaired feature integration*, which is a complex problem in machine learning. We achieved integration by using supervision; specifically, we computed a Product of Experts (PoE) between the latent representations obtained for each matching cell type. This allowed us, for example, to integrate datasets from different species without the need to

define orthologs.

Given that our model accepts unpaired features as input, extending it to integrate datasets from different modalities is a promising avenue. This could be achieved by including previously published modelling choices[250,268,298] into our method's generative module. However, obtaining matching cell types in this scenario is complex, as it requires accurate harmonisation across varied data sources. Pre-training the shared space of *spVIPES* in an unsupervised integration setting would allow us to use clusters as surrogates for labels in the disentangling phase. In order to achieve unsupervised integration of unpaired features, *cycle consistency*[299] could be leveraged. This approach constrains learning by ensuring that a feature translated from one modality to another and then back to the original modality remains unchanged. While this method has been previously applied to single-cell data[300] using Autoencoders (AEs), adapting it for *spVIPES* presents challenges. Specifically, the decoder in *spVIPES* is non-deterministic and outputs parameters of a Negative Binomial (NB) Mixture, from which samples are drawn for reconstruction. These samples would need to be translated through the encoder of the opposite modality. However, feeding these samples into an encoder network is problematic because the lack of reparameterization in the NB distribution prevents backpropagation. Another option would be to incorporate assumptions about the high-dimensional similarity between the datasets, such as both datasets sampling the same cell types. In this case, methods like *scTopoGAN*[301] could be integrated into our model.

6.3.3 WHAT AFFECTS DISENTANGLEMENT?

Disentanglement is traditionally achieved by limiting the amount of information encoded in the latent variables. For example, β-VAE[69] increases the weight of the KL divergence term, which results in latent variables closer to the prior (i.e., an isotropic Gaussian). If the latent space is all centred around zero with unit variance, there is less capacity to differentiate between subtle variations in the data. This limits the expressiveness of the latent variables, forcing the model to prioritise essential features for reconstruction, leading to more factorised and interpretable latent variables[302]. When applying this model to scRNA-seq data, increasing the hyperparameter β leads to better disentanglement (i.e., recovery of generative factors such as cell identity or pathways) at the cost of performance in downstream tasks such as clustering[303]. Given the high dimensionality and complexity of scRNA-seq data, the reduced expressiveness of the latent space limits the model's ability to capture more complex patterns, leading to oversimplified representations.

Recent findings indicate that achieving disentangled representations in a completely unsupervised scenario (as in β-VAE) is fundamentally impossible due to the non-uniqueness of z to reconstruct x (for a thorough mathematical derivation, see[304]). To address this, the learning of disentangled representations should leverage prior information about the generative process (e.g., cell type labels or dataset of origin) in supervised scenarios to constrain the recovery of the underlying generative factors. For example, *Biolord*[72] achieves disentanglement by training a model for each attribute (e.g., cell type, perturbation status, or

time point) and a single model capturing unknown variation in single-cell datasets. *Biolord* outputs representations for each known attribute and an extra representation for the unknown attributes.

In addition to supervision, other inductive biases (i.e., parameter and architectural choices) allow models to learn disentangled representations. One obvious parameter that might influence disentanglement performance is the (lower) dimensionality of the latent space, since the model has less capacity to capture complex factors and might focus on the most important generative factors for reconstruction. Crucially, baseline models such as *scVI*[51] are unable to achieve disentanglement with lower dimensionality alone. In Chapter 4, we showed how *spVIPES* significantly improves disentanglement when choosing different numbers of dimensions for private and shared spaces. In our model, we combine spaces of high and low dimensionality to reconstruct each cell's original feature space. It becomes apparent that this architectural choice ultimately affects the capacity (i.e., expressiveness) of each of the encoder networks, facilitating disentanglement between representations. Other methods[71] achieve disentanglement by limiting the information encoded using non-informative priors, which also limits the network capacity. As cohorts with more factors of variation (e.g., time, perturbation status, mutation burden, or treatment) become available, it would be beneficial to build supervised models that have the intra-dataset disentanglement capabilities of *Biolord* with cross-dataset disentanglement using architectures adapted from *spVIPES* or similar multi-sample disentanglement methods.

6.3.4 EVALUATING DISENTANGLEMENT

In Chapter 4, we evaluated disentangled representations by quantifying the degree of separation of known generative factors (e.g., cell type labels or genetic programs) in the latent space. However, assuming that these are the main generative factors is questionable, especially in complex biological contexts where more intricate factors are likely contributing to the observed phenotypes. As such, disentangled representations should be evaluated with more rigorous metrics that can quantify *informativeness* (i.e., how much information about x is captured in a disentangled factor z_i) and *independence* (i.e., the information contained in z_i should not be present in z_j). Several mutual information-based metrics have been proposed for both the supervised (i.e., the generative factors are known) and unsupervised (i.e., the underlying generative factors are unknown) settings in the machine learning literature[305], which could be adapted to the biological context.

6.3.5 IS DISENTANGLEMENT POSSIBLE?

Disentanglement typically assumes that the underlying generative factors of the observed data are independent. This assumption might hold for less complex datasets, such as synthetic or toy datasets, where the factors are explicitly designed to be independent (e.g., the position, shape, and color of objects in the dSprites dataset[306]). However, in biological contexts,

the generative factors are often dependent and influenced by intricate relationships and interactions. This raises questions about the validity of disentanglement methods, as the assumption of independence may not accurately reflect the true underlying structure of the data.

Models such as Biolord address this issue by learning a representation that captures unknown factors while constraining disentanglement based on attributes such as cell type. However, the hierarchical nature of cell types poses a significant challenge to the assumption of independence. For example, the generative factor corresponding to injured proximal tubule cells (Injured Proximal Tubule (PT-Inj)) needs some degree of shared representation with the factors underlying regenerating (Repairing Proximal Tubule (PT-R)) and failed repair (Failed Repair Proximal Tubule (PT-FR)) proximal tubule states. This hierarchical relationship between cell states directly contradicts the assumption of independent factors.

To overcome this limitation, it would be beneficial for disentanglement methods to leverage the cell type hierarchy. Such methods could learn a representation where groups of dimensions are active for parent nodes (e.g., proximal tubule cells) while individual dimensions correspond to leaf nodes in the hierarchy (e.g., PT-Inj and PT-R cells). This hierarchical approach could enable disentanglement methods to better capture the complex dependencies inherent in biological data.

6.4 FUTURE PERSPECTIVES AND BIOLOGICAL IMPLICATIONS

The field of computational single-cell analysis is rapidly evolving, with several key areas ripe for further exploration and innovation. This section outlines three important directions for future research: disentanglement in single-cell data, integration of longitudinal and spatial information, and refinement of cell-cell communication (CCC) analysis techniques.

Disentanglement has emerged as a crucial area of focus in single-cell analysis. The ability to separate complex biological processes into interpretable factors offers potential for deepening our understanding of cellular biology. By separating biological variability into distinct factors such as cell cycle stage, differentiation status, or environmental response, disentanglement could improve the accuracy of cell type classification, particularly for overlapping or ambiguous cellular phenotypes. For instance, it could enhance the differentiation between progenitor and early differentiated cell states, or subtle distinctions among cell type subpopulations. However, the application of disentangled representations to improve classification remains unexplored. Furthermore, achieving meaningful unsupervised disentanglement in biological data presents significant challenges.

Future research should concentrate on creating unsupervised disentanglement methods, establishing more rigorous evaluation metrics, and exploring how disentangled representations

can enhance the identification of rare cell populations or subtle cellular states.

The integration of longitudinal and spatial information represents another critical area for future development. Current methods often struggle to fully capture the temporal and spatial aspects of cellular processes. Developing computational approaches that can effectively leverage both temporal and spatial data will be crucial for understanding cellular origins and trajectories.

For example, elucidating the mechanisms behind failed-repair proximal tubules (FR-PT) could benefit from methods that integrate longitudinal data with spatial transcriptomics. Understanding how and why certain proximal tubules escape repair or undergo repair, both spatially and temporally, is key to understanding kidney injury and repair processes. This integration presents unique computational challenges, requiring novel algorithms that can handle both temporal and spatial contexts to obtain cellular trajectories.

Advancing cell-cell communication (CCC) analysis depends on overcoming several critical challenges. While current methods predominantly rely on transcriptomic data, future approaches must integrate diverse data types to reveal new dimensions of cellular interactions. A fundamental obstacle in CCC research is the scarcity of reliable ground truth data, stemming from the dynamic and context-dependent nature of intercellular signalling.

To address this, future research should focus on developing innovative strategies for ground truth data generation, such as genetic perturbations that systematically disrupt specific communication channels or *in situ* protein-protein interaction assays. The knowledge gained from these experimental approaches could be synthesized into comprehensive prior knowledge graphs, which could serve as a basis for generating benchmarking datasets. Such benchmarks would be valuable for evaluating and refining computational CCC methods, guiding the development of more algorithms capable of integrating multimodal data and leveraging biological priors.

In conclusion, the future of computational single-cell analysis lies in developing more sophisticated methods for disentanglement, integrating temporal and spatial information, and enhancing cell-cell communication analysis. These advancements aim to deepen our understanding of biological systems and disease mechanisms.