Genome Biology

METHOD

Open Access

scCross: a deep generative model for unifying single-cell multi-omics with seamless integration, cross-modal generation, and in silico exploration

Xiuhui Yang[1,2,3], Koren K. Mann[4], Hao Wu[1*] and Jun Ding[2,3,5*]

*Correspondence:
haowu@sdu.edu.cn; jun.ding@mcgill.ca

[1] School of Software, Shandong University, 1500 Shunhua, Jinan 250101, Shandong, China
[2] Meakins-Christie Laboratories, Department of Medicine, McGill University Health Centre, Montreal H4A 3J1, QC, Canada
[3] Quantitative Life Sciences, Faculty of Medicine & Health Sciences, McGill University, Montreal, QC H3G 1Y6, Canada
[4] Department of Pharmacology and Therapeutics, McGill University, Montreal, QC H3G 1Y6, Canada
[5] Mila-Quebec AI Institute, Montreal, QC H2S 3H1, Canada

Abstract

Single-cell multi-omics data reveal complex cellular states, providing significant insights into cellular dynamics and disease. Yet, integration of multi-omics data presents challenges. Some modalities have not reached the robustness or clarity of established transcriptomics. Coupled with data scarcity for less established modalities and integration intricacies, these challenges limit our ability to maximize single-cell omics benefits. We introduce scCross, a tool leveraging variational autoencoders, generative adversarial networks, and the mutual nearest neighbors (MNN) technique for modality alignment. By enabling single-cell cross-modal data generation, multi-omics data simulation, and in silico cellular perturbations, scCross enhances the utility of single-cell multi-omics studies.

Keywords: Single cell, Muti-omics, Cross-modal generation, In silico perturbations, Multimodal integration, Generative adversarial network, Autoencoder

Background

The advent of single-cell sequencing technologies has ushered in a new era of biological research, enabling scientists to deconvolve cellular heterogeneity in unprecedented detail [1–3]. This granularity has illuminated intricate cellular dynamics across myriad biological processes. To harness the full potential of this data deluge, a suite of computational tools has been developed, propelling advancements in fields as diverse as cancer biology, neurobiology, and drug discovery [4–11]. However, many of these tools are tailored to specific data modalities, such as scRNA-seq [12] or scATAC-seq [13, 14], often providing a piecemeal view of the cellular landscape.

As the single-cell sequencing paradigm matures, we are seeing the convergence of multiple data modalities, offering a holistic view of cellular states [15–20]. Effective single-cell multi-omics data integration remains challenging, particularly when integrating unmatched data. Many single-cell multi-omics measurements are unmatched, profiling varied cell populations across modalities and thus missing the matched multimodal profiling of the same cells.

Existing methods have emerged to tackle this data integration challenge, each with its own set of advantages and drawbacks. Current methods for single-cell multi-omic data integration exhibit several limitations. A significant number of these methods are designed to process data where different modalities originate from the same set of cells with matched information. Examples include scMoGNN [21], SAILERX [22], scMVP [23], MIRA [24], SCALEX [25], scMDC [26], and scAI [11]. However, these methods are often limited by their reliance on matched multi-omics data, which is not always readily available or feasible to obtain. Additionally, many existing methods are supervised or semi-supervised, requiring pre-annotated datasets, such as predefined cell types, for effective data integration analysis. This prerequisite poses a limitation when pre-annotated data is not available, thus restricting the applicability of these methods in various contexts. Examples of such methods include scJoint [10] and Portal [27]. Even the methods capable of handling both matched and unmatched single-cell multi-omics data suffer from various limitations. For instance, methods like Seurat [28] and Harmony [9] hinge on finding commonalities across modalities, while others, such as scglue [29], uniPort [30], sciCAN [31], and scDART [32], explore nonlinear transformations or lean into deep learning. Many of these techniques face issues ranging from information loss to noise susceptibility [33].

Furthermore, due to the cost and difficulty in experiments, there is often a significant scarcity of matched single-cell multi-omics data that comprehensively profiles cellular states. Instead, we typically have abundant data for the most dominant modality (like scRNA-seq) but limited or no profiling for other modalities (e.g., epigenetics). The presence of other matched single-cell multi-omics data offers the potential for cross-modal generation of missing modalities from the more abundant ones by transferring knowledge learned from the reference.

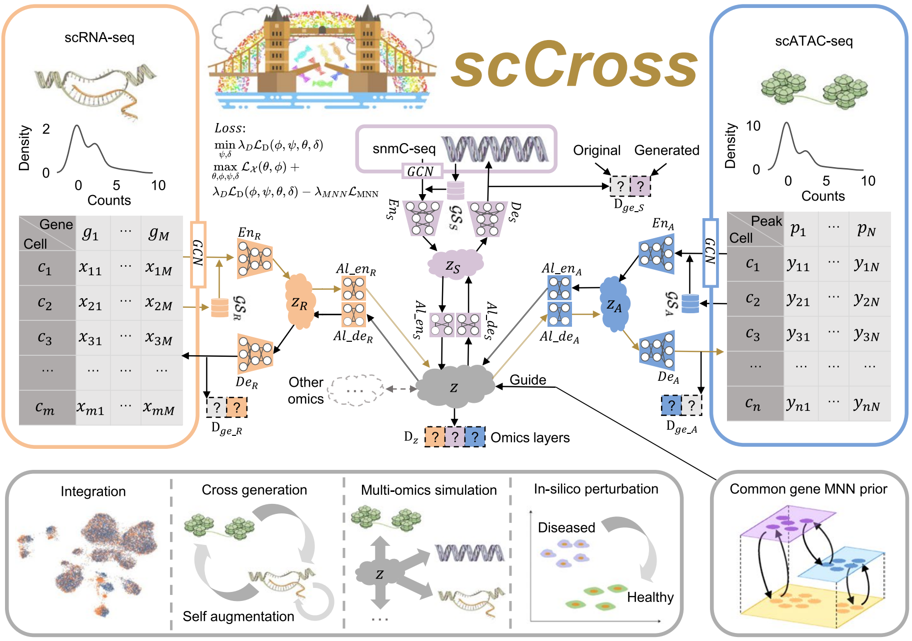
Here, we introduce scCross. At its foundation, scCross excels in its function of single-cell multi-omics data integration, bringing unparalleled precision to the assimilation of diverse data modalities. While preserving its primary proficiency in this arena, the most distinctive feature of scCross emerges: its prowess in cross-modal single-cell data generation. This capability, layered atop its integration strength, unlocks transformative potentials for researchers to bridge disparities between abundant and scant modalities. Building on these two principal functionalities, scCross further showcases its versatility by simulating single-cell multi-omics data with high fidelity. Moreover, the tool unveils opportunities for in silico perturbations within and across modalities, enabling researchers to postulate and evaluate potential cellular interventions. By capturing and elucidating intricate cellular states, scCross equips the scientific community with a thorough grasp of cellular dynamics across modalities. scCross stakes its claim as a vanguard in single-cell multi-omics integration, generation, and simulation.

Results

Overview of scCross

scCross employs the variational autoencoder (VAE)-generative adversarial network (GAN) deep generative framework to integrate single-cell multi-omics datasets, generate cross-modal data, simulate multi-omics data, and perform in silico perturbations within and across modalities (Fig. 1 and Additional file 1: Fig. S1). The methodology begins by training modality-specific variational autoencoders (VAEs) to capture low-dimensional cell embeddings for each data type, supplemented by the integration of gene set score vectors as representative features. With these embeddings learned, they are integrated into a common latent space. The harmonization of the modalities' data distributions then ensues. The generative adversarial network (GAN) refines this process, with a discriminator network juxtaposed against the VAE generator. This structure ensures a robust multi-modal data integration and confirms the alignment between the actual and generated single-cell inputs.

Venturing beyond mere integration and other functions, the model is specifically designed to facilitate cross-modal data generation. The bidirectional aligner is crucial for this cross-modal generation, as it decodes the shared latent embedding into a distinct modality, relying on the cell embedding sourced from the other modalities. This ambition of the model requires a dual alignment: a congruent data distribution and a meticulous coordination of individual cells across modalities. Mutual nearest neighbor (MNN) cell pairs are strategically employed as alignment anchors, guiding this intricate process. The optimization of the neural network aims to minimize discrepancies across



Cell \ Gene	g_1	⋯	g_M
c_1	x_{11}	⋯	x_{1M}
c_2	x_{21}	⋯	x_{2M}
c_3	x_{31}	⋯	x_{3M}
⋯	⋯	⋯	⋯
c_m	x_{m1}	⋯	x_{mM}

Cell \ Peak	p_1	⋯	p_N
c_1	y_{11}	⋯	y_{1N}
c_2	y_{21}	⋯	y_{2N}
c_3	y_{31}	⋯	y_{3N}
⋯	⋯	⋯	⋯
c_n	y_{n1}	⋯	y_{nN}

Fig. 1 Architecture of the scCross framework. scCross employs modality-specific variational autoencoders to capture cell latent embeddings $\mathbf{z}_R$, $\mathbf{z}_S$, $\mathbf{z}_A$, ..., for each omics type. During single-cell data integration, the method leverages biological priors by integrating gene set matrices $\mathcal{GS}_R$, $\mathcal{GS}_S$, $\mathcal{GS}_A$, ..., as additional features for each cell. The framework then harmonizes these enriched embeddings into shared embeddings z using further variational autoencoders and critically, bidirectional aligners. Bidirectional aligners are pivotal for the cross-modal generation, depicted by brown arrows signifying the transition from scRNA-seq to scATAC-seq. Mutual nearest neighbor (MNN) cell pairs ensure precision during alignment. Discriminator $\mathrm{D_z}$ maintains the integration of different omics and discriminators $\mathrm{D}_{\mathrm{ge}_R}$, $\mathrm{D}_{\mathrm{ge}_S}$, $\mathrm{D}_{\mathrm{ge}_A}$, ..., maintain the integrity and consistency of generated data. scCross offers a powerful toolbox for single-cell data integration, facilitating cross-modal data generation, single-cell data self-augmentation, single-cell multi-omics simulation, and in silico perturbations, making it versatile for a plethora of single-cell multi-omics challenges

modalities, especially those evident between MNN pairs. Once trained, the model enables the generation of single-cell data across modalities, encoding data from one modality into the latent space, and subsequently decoding it into another. It can also simulate multi-omics data and facilitates in silico perturbations within and across modalities, revealing potential modulations of cellular states. By merging single-cell multi-omic data into a unified latent space and enabling information flow across modalities, scCross sets the stage for various multi-omic tasks.

scCross improves single-cell multi-omics integration over established methods

Single-cell multi-omics data has emerged as a transformative tool that comprehensively captures cellular states, offering profound insights into cellular identity and function. Effective integration of the multi-omics data produces cell embeddings crucial for diverse analyses. These embeddings are pivotal for precise cell clustering and identification, and they support numerous downstream tasks. Metrics such as the Adjusted Rand Index (ARI) [34] and Normalized Mutual Information (NMI) [35] are invaluable in evaluating the quality of these embeddings, reflecting the proficiency of an integration method in retaining biological information.

We benchmarked scCross against several state-of-the-art single-cell data integration methods, including Seurat v4 [28], scglue [29], uniPort [30], sciCAN [31], scDART [32], and Harmony [9]. Our analysis utilized three gold-standard datasets derived from recent single-cell multi-omics sequencing, encompassing simultaneous scRNA-seq and scATAC-seq profiling. These datasets are matched mouse cortex [36], matched mouse lymph nodes [37], and matched mouse atherosclerotic plaque immune cells (GSE240753). Additionally, we benchmarked the methods on another unmatched dataset, which includes scRNA-seq [38], scATAC-seq sourced from the 10X Genomics website, and an snmC-seq dataset [39].

Against this landscape, scCross stands as a leading method in single-cell multi-omics data integration, showcasing comparable or better performance relative to other methods (Fig. 2 and Additional file 1: Figs. S2–S5). This distinction is evident from two primary angles. Firstly, scCross excels in downstream cell clustering, as demonstrated by its outstanding ARI, NMI, and cell type average silhouette width [29, 40] metrics across all benchmark datasets (Fig. 2a). Secondly, its prowess in cell mixing is apparent, seamlessly blending cells from different modalities. Effective cell mixing is indicative of high-quality integration and the method's ability to represent the intricate nuances of each modality. Please refer to the "Methods" section for a detailed description of the evaluation metrics.

To further underline our model's capacity for omics mixing, we assessed the FOSCTTM (Fraction of Samples Closer than the True Match) score [41], a metric quantifying the alignment error in single-cell multi-omics data integration. This score provides insights into how closely related cells from one omics modality are to those from another in the integrated latent space. A lower FOSCTTM score signals a more accurate integration, as the true match is typically closer than many of the other potential matches. On all three matched datasets, scCross consistently achieved the comparable or better FOSCTTM on all sizes of subsampled datasets, underscoring its alignment capabilities (Fig. 2b).

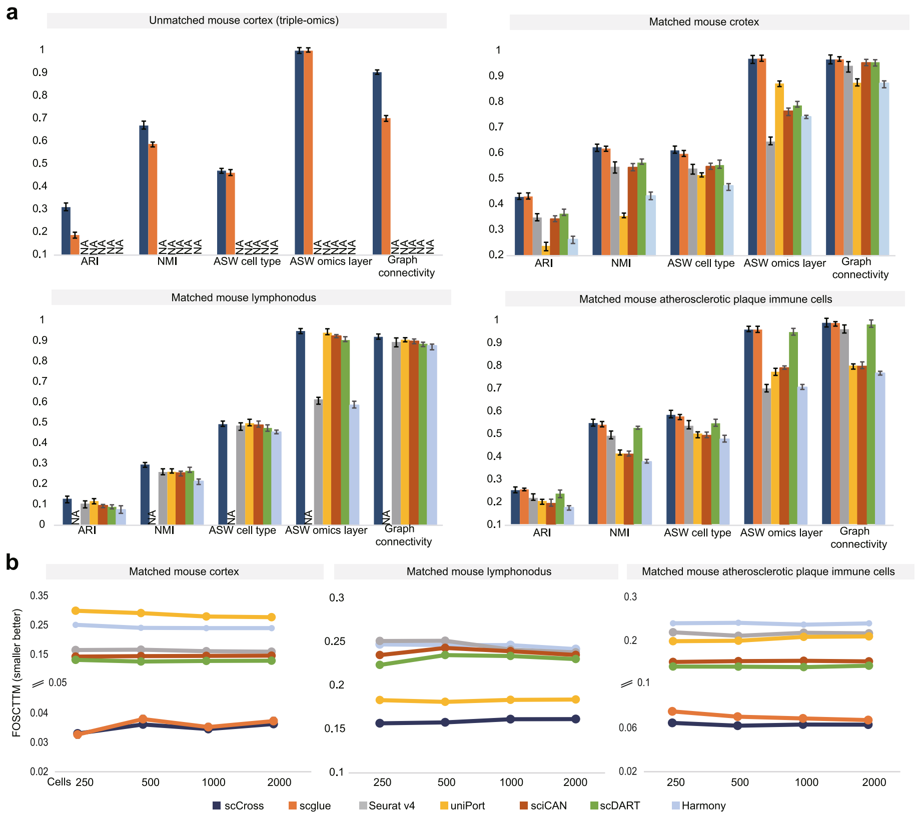


Fig. 2 Benchmarking superiority of scCross in single-cell multi-omics data integration. **a** A comprehensive comparative analysis showcases the comparable or better performance of scCross against other integration methods. Metrics such as average ARI, NMI scores, and ASW for cell types provide insights into clustering quality on the integrated single-cell multi-omics datasets, signifying the retention of biological characteristics after integration. Additionally, the ASW for omics layers and graph connectivity metrics shed light on the effective mixing of different omics post-integration. The datasets studied include the unmatched and matched mouse cortex, matched mouse lymph nodes, and matched mouse atherosclerotic plaque immune cells. "NA" indicates cases where results were not obtainable (e.g., methods do not support three omics integration or only obtain non-numeric outputs). **b** FOSCTTM scores further underscore the exceptional performance of scCross on subsampled datasets of varying sizes, demonstrating its consistent comparable or better performance over other integration techniques. This section delves into datasets like the matched mouse cortex, matched mouse lymph nodes, and matched mouse atherosclerotic plaque immune cells

Moreover, as detailed in Additional file 1: Fig. S6a–b, scCross demonstrates efficient consumption of computational resources in terms of both time and memory. This efficiency is particularly critical as the scale of single-cell data increases. For large datasets with more than 10,000 cells, our method surpasses the performance of most benchmarked tools in terms of running time and memory cost, demonstrating the computational effectiveness of scCross.

Effective integration is vital for successful downstream activities such as cross-modal generation, multi-omics simulation, and in silico perturbation. Stellar performance in these subsequent tasks speaks to the exemplary data integration and information retention capabilities of scCross. Notable differences between modalities in the combined latent space can obstruct generation, simulation, or in silico perturbation, underscoring the necessity of seamlessly merging all modalities into a unified space.

With the advent of high-throughput single-cell technologies, there's an increasing demand for methods that can effectively process large-scale single-cell multi-omics data, often comprising millions of cells [29, 42]. In our evaluation of scCross's integrative capabilities, we tapped into over four million human single-cell atlases, consisting of gene expression data for four million cells [43] and chromatin accessibility data for 0.7 million cells [44]. The challenge with such large-scale integration lies not just in the data volume but also in handling extensive heterogeneity, low per-cell coverage, and imbalances in cell type compositions [29]. Moreover, there is a common pitfall where existing methods tend to blend minority cell types with the majority, skewing accurate representation and estimation of these pivotal cell groups [45].

scCross addresses these challenges effectively and retains a significant amount of biological data. Unlike existing methods, this approach can differentiate and represent minority cell populations distinctly from the dominant ones. Thanks to the minibatch neural networks [46], scCross achieves sublinear time complexity and roughly linear memory consumption, as highlighted in Additional file 1: Fig. S6c. Through the strategic use of the generative adversarial network and MNN prior, scCross brings together gene expression and chromatin accessibility data into a comprehensive multi-omics human cell atlas [44], distinguishing the minority cell populations from the majority as demonstrated in Additional file 1: Fig. S7a–b.

In scglue's analysis by Cao et al. [29], notable discrepancies between cell type labels across omics modalities were uncovered. Cells labeled as Astrocytes in scATAC-seq data were found to be aligned with Excitatory neuron clusters in scRNA-seq, suggesting a potential mislabeling in the scATAC-seq dataset, which was inferred from biomarker analysis. Furthermore, a cluster identified as astrocytes/oligodendrocytes in scATAC-seq was divided and realigned to separate Astrocytes and Oligodendrocytes clusters in scRNA-seq, as indicated by blue cycles in Additional file 1: Fig. S8a–b. These discrepancies identified by scglue were also replicated by scCross, shown in the blue-highlighted clusters in Additional file 1: Fig. S7b and Additional file 1: Fig. S8c. Our analysis further confirms the potential inaccuracy of the Astrocytes annotation in scATAC-seq data, as its top marker genes are not consistent with the biomarkers of Astrocytes in scRNA-seq data (Additional file 1: Fig. S8d). Both scglue and scCross demonstrate the ability to detect and address potential discrepancies in cell annotations across different omics datasets.

Compared to methods such as scglue [29], scCross demonstrates enhanced capability in identifying minority cell type populations. This is particularly evident in the case of critical minority populations like Extravillous trophoblasts [47–49]-key cells in placental development and function-and Microglia [50–52], which are the primary immune cells in the brain involved in neuroinflammation and tissue repair. These cells are crucial in reproductive and neural systems, respectively. These critical populations are often prone to over-integration with majority cell types in other method like scglue [29] (Additional file 1: Fig. S8a). However, as highlighted by the red cycle in Fig. S8B, scCross distinguishes them, further validating the integration prowess of our approach. Comprehensive benchmark comparisons between scCross and scglue [29] for this dataset are available in Additional file 1: Fig. S8e. Alternative approaches such as Seurat v4 [28], uniPort [30], sciCAN [31], scDART [32] and Harmony [9] have not demonstrated their effectiveness in integrating large-scale single-cell multi-omics data with millions of cells.

Given its intrinsic modularity, scCross is not limited to merely a binary omic integration but boasts the capability to consolidate multiple omic layers (e.g., more than 3), attesting to its comprehensive applicability. To showcase this extensive integration potential, we employed scCross on single-cell multi-omics data of the adult mouse cortex, encompassing three distinct modalities: gene expression [38], chromatin accessibility, and DNA methylation [19].

In Fig. 3a–d, scCross exemplifies its prowess in aligning the cell types across the three modalities. The alignment is significantly enriched by the incorporation of common genes MNN prior and specific gene sets, thereby ensuring that corresponding cell types from various omic layers converge harmoniously within the latent space. Such alignment not only amplifies the clarity of cell type demarcation across the omic layers but also paves the way for enhanced cell typing efficacy. scCross demonstrates its proficiency in navigating triple-omics datasets, providing a shared cell embedding that effectively distinguishes between cell types for each modality, as depicted with RNA in Fig. 3a. This clear separation is mirrored in the ATAC-seq and snmC-seq modalities, where cell types also form distinct groupings in Fig. 3b and c, respectively. The embedding further shows its strength by closely aligning identical cell types across different modalities, which are visible as clusters in the same regions of the UMAP space. Adding to this, Fig. 3d illustrates the embedding's capacity for mixing cells from different modalities in the shared space, facilitating the study of cross-modality interactions without the loss of cell-type-specific clustering. This nuanced approach ensures that while cell types from different modalities interweave, they still maintain the fidelity of their distinct embeddings.

Upon establishing the shared cell embeddings, our evaluation centered on the coherence of cell type-specific markers across modalities to validate the alignment accuracy.

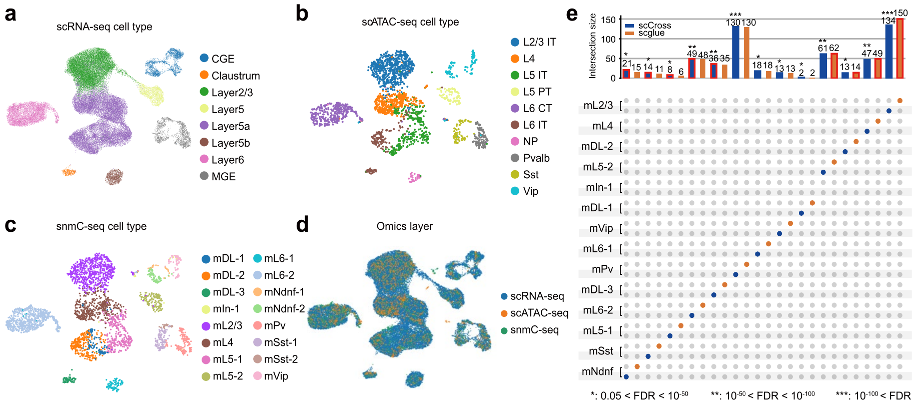


Fig. 3 Efficient integration of three omics layers in the mouse cortex by scCross. **a**–**c** UMAP visualizations display the aligned cell embeddings, clustering, and cell type annotations for scRNA-seq (**a**), snmC-seq (**b**), and scATAC-seq (**c**), on the shared joint latent space, exemplifying the coherence in cellular representations across modalities. **d** A comprehensive UMAP presentation underlines the flawless intermixing of cells from distinct modalities within the shared latent space, reflecting a true amalgamation of multi-omics information. **e** An UpSet plot emphasizes the alignment potency of our model, revealing through a three-way biomarker comparison that the distinct modalities are seamlessly integrated within the latent space. Leveraging a three-way Fisher's exact test [53], our model is showcased to either match or surpass scglue [29] in the alignment of over half of the cell types, with standout results accentuated by a red outline

As depicted in Fig. 3e, there is a pronounced intersection of markers for each cell type between modalities, substantiating the precision of our cell type alignments. This significant overlap was confirmed by a three-way Fisher's exact test, with three cell types showing a staggering FDR below 10^{-100}, and four additional cell types displaying FDRs between 10^{-100} and 10^{-50}. All cell types, with the sole exception of mIn-1, had FDR values indicating strong significance below the 0.05 benchmark. This degree of marker congruence emphasizes the successful integration of the omics layers, as our approach identifies consistent key markers across different modalities for equivalent cell types. In comparison to scglue, our analysis uncovered a greater number of overlapping gene markers in the majority of cell types (9 out of 13), further reinforcing our model's efficacy in multi-omic data integration.

scCross empowers cross-modal single-cell data generation

Single-cell multi-omics data offers deep insights into the complexities of cellular states. Yet, factors such as cost and experimental challenges often impede our capability to fully harness its advantages, leading to some modalities being less explored than others. In this context, scCross emerges as a crucial tool, focused on bolstering cross-modal single-cell data generation to tap into the vast potential of single-cell multi-omics. It achieves this by seamlessly integrating knowledge across modalities, drawing from established reference single-cell multi-omics datasets encompassing all desired modalities or from multi-omic datasets where the underrepresented modalities were insufficiently profiled.

While tools like scglue have their merits in single-cell multi-omics data integration, they inherently lack the capability for cross-modal generation. Specifically, scglue's entire framework emphasizes its unsuitability for cross-modal data generation, especially when noting that its decoder is designed to process embeddings from the input of the same modality. Contrastingly, scCross is architected to bridge such analytical gaps. To showcase its prowess and endorse its robustness and evaluate the method's performance in situations with limited or partial single-cell data of another modality of interest, we divided the matched mouse cortex dataset [36] into two parts—20% for training and the remainder for testing. Leveraging the model trained on the limited matched (20%) single-cell data, we applied the method to cross-generate the single-cell RNA-seq for the other missing 80% single-cell RNA-seq. The results, depicted in Fig. 4a-e, shed light on scCross's ability not only to generate but also to uphold the biological nuances of the original data. The UMAP visualization [54] in Fig. 4a showcases both the original and generated scRNA-seq data clustering closely, hinting at a high similarity in their distributions. A compelling confirmation of this is observed in Fig. 4b, where the cell type proportion between the original and generated datasets demonstrates a correlation of 0.89 with a p-value of 0.001311. In Fig. 4c, the gene expression of the top 5 genes for each cell type between both datasets remains largely consistent, with a correlation of 0.93 and a p-value of 5.356×10^{-177}. The remarkable similarity between the actual data and the crossed modality-generated data is apparent in those downstream analysis outcomes. Cell-cell interaction patterns, when analyzed through the CellChat tool [55] as seen in Fig. 4d, not only closely emulate those in the original dataset but also provide a window into potential interactions, leveraging insights from the scATAC-seq data of the other modality. This is quantitatively expressed by a correlation coefficient of 0.82 and a

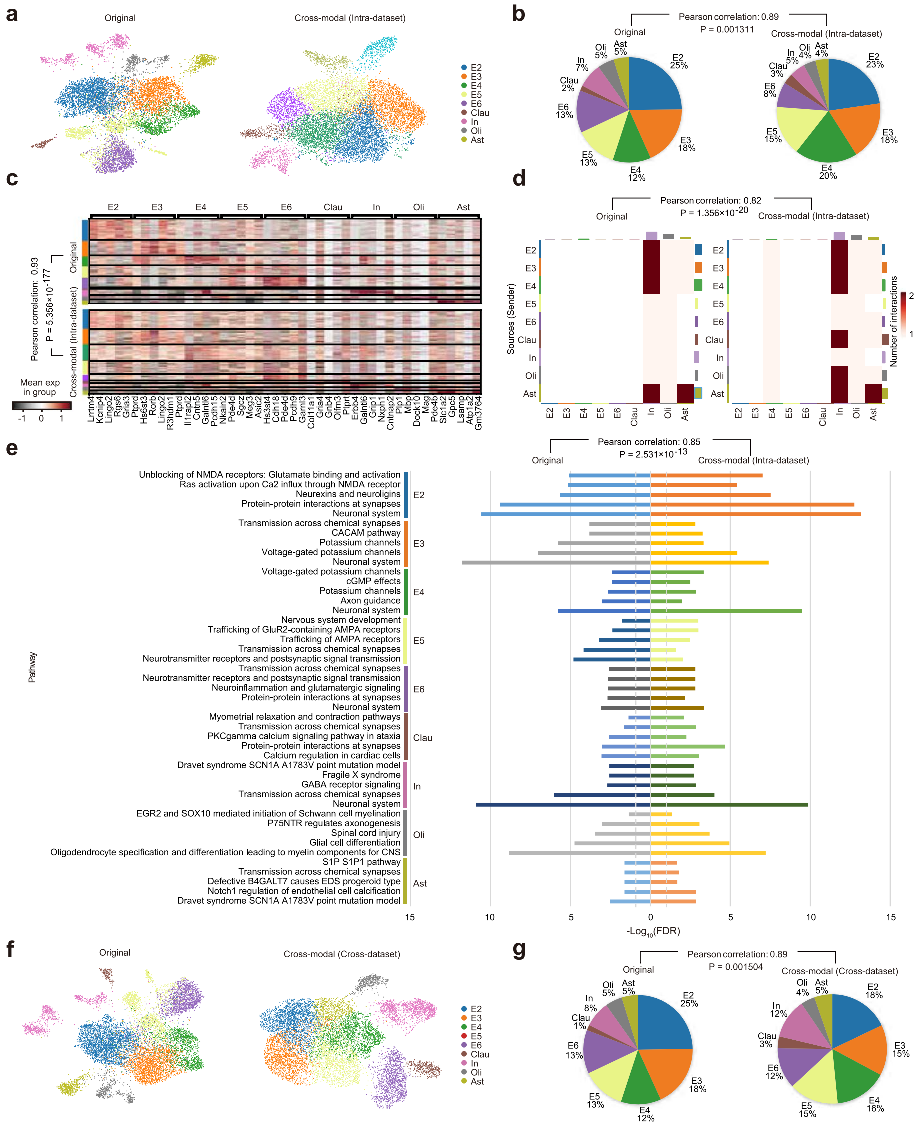


Fig. 4 Performance of scCross in cross modal single-cell data generation: emphasis on similarity between actual and cross-generated data. **a**–**e** Results when the model was trained on the partial dataset: (**a**) UMAP visualization emphasizes the overlapping distribution between the original and cross-generated scRNA-seq data. **b** Comparison of cell type proportions reveals a robust correlation of 0.89 ($P = 0.001311$). **c** Expression levels of the top 5 genes per cell type exhibit a correlation of 0.93 ($P = 5.356 \times 10^{-177}$). **d** Cell-cell interaction metrics show congruity across cell types with a correlation of 0.82 ($P = 1.356 \times 10^{-20}$). **e** Ratios of genes in the most significant pathways for each cell type validate a correlation of 0.85 ($P = 2.531 \times 10^{-13}$) between the original and cross-generated datasets. **f**, **g** Results when the model was trained on independent reference multi-omics datasets: (**f**) UMAP visualization illustrates the clustering resemblance between the original and cross-generated matched mouse scRNA-seq data. **g** Cell type proportions underscore a correlation of 0.89 ($P = 0.001504$) between the original and cross-generated scRNA-seq data

compellingly significant p-value of 1.356×10^{-20}. Furthermore, pathway analysis based on the top 100 marker genes for each cell type, conducted with the ToppGene platform [56] and depicted in Fig. 4e, reinforces the crossed data's relevance for robust biological interpretation. This analysis maintains a high correlation of 0.85 with a p-value of 2.531×10^{-13}. The integrity of the crossed data is further substantiated by UMAP analyses presented in Additional file 1: Figs. S9–S12, which cover a comprehensive set of benchmark datasets, both matched and unmatched. Collectively, these analyses affirm the crossed data's exceptional ability to mirror the biological intricacies and preserve the analytical properties of the actual data.

The task becomes more challenging when training and testing on different datasets. scCross's capability for cross-modality generation was further rigorously tested in a scenario devoid of any direct training set for the missing modality. Utilizing an unmatched reference single-cell multi-omics dataset containing both RNA-seq and ATAC-seq data from the same samples [38], scCross was trained to harness and transfer this integrated knowledge. It was then tasked with generating single-cell RNA-seq data from single-cell ATAC-seq data alone, sourced from another matched multi-omics dataset [36]. The authenticity of this cross-generated RNA-seq data was critically assessed against the actual RNA-seq dataset, with the UMAP overlap in Fig. 4f serving as a visual testament to scCross's crossmodal generation precision. The correlation in cell type composition between the cross-generated RNA-seq data and the actual RNA-seq data further substantiates scCross's performance, achieving a correlation of 0.89 and a *p*-value of 0.001504 as illustrated in Fig. 4g. The practicality and relevance of scCross's output are reinforced through various downstream analyses, detailed in Additional file 1: Fig. S13. This showcases scCross's ability to not just mimic, but also potentially extrapolate and fill in gaps in single-cell omics profiles, reinforcing its role as a potent tool for advancing biological insights in the absence of comprehensive multi-omic datasets.

Intra-modal simulation of matched single-cell multi-omics data

Another capability of scCross is its ability to generate matched single-cell multi-omics data. This computational prowess opens the door to numerous practical applications. One immediate use is in benchmarking existing single-cell multi-omics integration methods, especially in scenarios where true matched single-cell multi-omics data is lacking for accurate evaluations. Moreover, the simulated data can also serve as a basis for deciphering intricate inter-omics relationships, predicting states of unmeasured omics, or filling in the gaps in studies with incomplete modalities.

One of the standout features of scCross is its tailored generation capability, targeting specific cell types of interest across various omics layers. This becomes indispensable when focusing on rare cell populations, which often are underrepresented in sequenced datasets. The limited number of cells in these populations poses challenges for in-depth analysis and examination. Here, scCross provides a valuable solution. By simulating data up to 5X the original count, as shown in panels Fig. 5a and b, scCross can substantially upscale the representation of these rare populations, enabling more robust and comprehensive statistical analyses.

Panel Fig. 5a illustrates the simulated single-cell RNA-seq data of Ast cells in the matched mouse cortex dataset, both at its original count and a 5X upscale with the whole original matched mouse cortex dataset as cluster background. Panel b mirrors this information for the scATAC-seq data. The high-quality cell clustering shown in these panels testifies to the effectiveness of our simulation approach. Other cell types in the matched mouse cortex dataset also give out the same results (Additional file 1: Fig. S14).

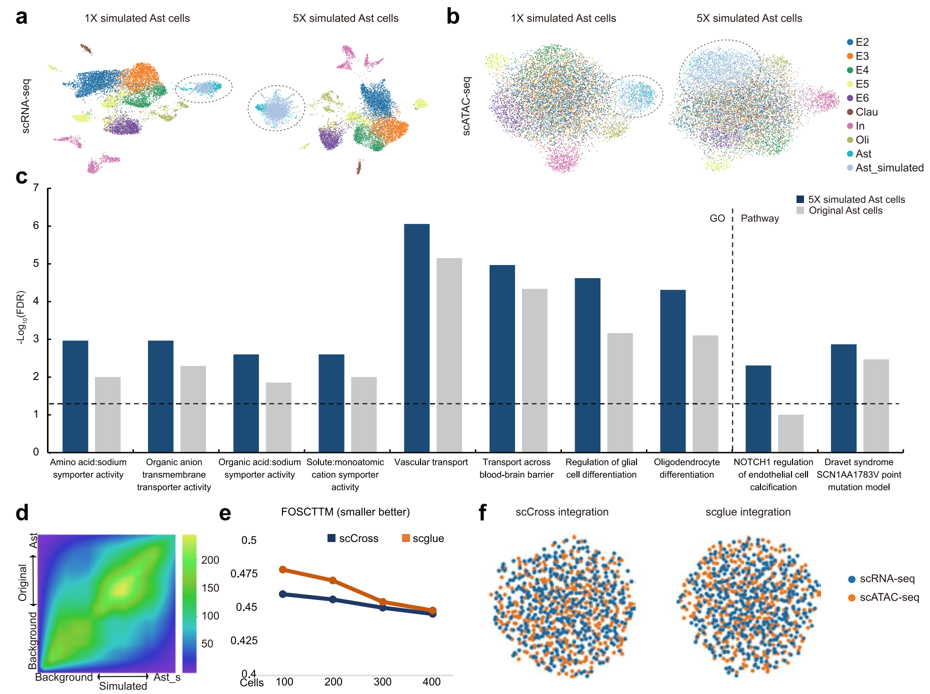


Fig. 5 Intra-modal simulation of matched single-cell multi-omics data. **a** UMAP visualization showcasing the prowess of scCross in simulating the scRNA-seq data of rare Ast cells in matched mouse cortex dataset (highlighted by gray cycles), presented at 1X and 5X their original count with the whole original dataset as cluster background. The 5X scale highlights the model's capability to upscale the representation of underrepresented populations like Ast cells. **b** Parallel to panel (**a**), this UMAP visualization delineates the simulated scATAC-seq data for Ast cells at the same 1X and 5X scales (highlighted by gray cycles), reinforcing the model's consistent performance across omic modalities. **c** The analysis contrasts the Gene Ontology (GO) terms and pathways enriched among the top biomarkers of the original Astrocyte (Ast) cells (scRNA-seq) with those of its 5X augmented simulated equivalent. The comparison underscores the model's proficiency in amplifying key biological signals within the framework of multi-omics data simulation. **d** RRHO plot emphasizing the significant correlation between the original Ast scRNA-seq data and its 1X simulated counterpart. This highlights the model's accurate capture of key biomarker genes during the simulation. **e** FOSCTTM integration metrics underscore the potential of the 1X simulated Ast cells' single-cell multi-omics data (encompassing both scRNA-seq and scATAC-seq) as an evaluative benchmark. Both scCross and scglue are compared; however, the inherent bias towards scCross due to the data's origin should be acknowledged. **f** UMAP visualization displaying the robust cell mixing achieved by both scglue and scCross when processing the 1X simulated Ast cells' single-cell multi-omics data, providing a vivid demonstration of their respective capabilities

To showcase the efficacy of scCross in simulating single-cell multi-omics data for rare cell populations, we targeted the Astrocyte (Ast) cells within the dataset, augmenting their representation by a fivefold increase. This upscaling was performed to better understand the potential implications of such rare cell types in subsequent analyses. We then conducted a comparative study between the original and the simulated dataset, centering on the enriched Gene Ontology (GO) terms and pathways linked to the Ast cell marker genes prior to and following their expansion. In Fig. 5c, the enriched terms are ranked and displayed, with particular attention to the leading terms. The two principal terms, "Vascular transport" [57] and "Transport across blood-brain barrier" [58], were found to be closely associated with the physiological roles of the Ast cells. The pathway "NOTCH1 regulation of endothelial cell calcification" shows significance only in the 5x upscaled Ast cell simulated data, but not in the original dataset. These findings were verified to be associated with the signaling processes in Ast cells [59], thereby further substantiating the potential effectiveness of scCross in enhancing rare cell populations. Further validating the simulation's accuracy, Fig. 5d compares the simulated data against the actual dataset for these rare Ast cells. The RRHO plot therein confirms that scCross preserves the key attributes of the Ast cells, thereby substantiating the simulated data's fidelity for this specific subset within the larger dataset.

The utility of this simulated data extends beyond mere generation. Panels Fig. 5e and f spotlight its potential applications. The panels underscore that the generated single-cell multi-omics data (1X simulated Ast) are "matched" since cells from distinct modalities derive from a consistent joint latent space. Panel **e** showcases the potential of using simulated data as an effective ground-truth to appraise single-cell data integration techniques. This perspective is reinforced in the panel **f**'s UMAP scatter plot, revealing exceptional cell mixing across modalities by both scglue and scCross. However, when examining the comparison in panel **e**, caution is advised. Since these simulated datasets originate from the scCross model, there might be a slight bias in favor of scCross. Nevertheless, this comparison underscores that these simulated datasets can serve as robust matched single-cell multi-omics benchmarks for examining a range of other integration methods especially when ground truth is experimentally unattainable.

Beyond the cross-generational capabilities from one modality to another and the intra-modal simulation, our exploration extended to intra-modal augment the inputs within the same modality, leveraging the inherent reconstruction and augmentation strength of the VAE-GAN framework [23, 29, 30]. This captivating line of inquiry uncovered that the data underwent an augmentation phase, which in turn enriched its quality and expanded its informational depth. Detailed results and visualizations can be found in Additional file 1: Fig. S15.

scCross enables multi-omics in silico perturbation for exploring potential cellular state intervention

In our endeavor to showcase the capabilities of scCross, we highlight its expertise in conducting in silico multi-omic perturbations. These capabilities not only enable the exploration of potential intervention strategies for cellular state manipulation but also provide insights into the intricacies of molecular responses, particularly those seen in COVID-19 infections using the single-cell multi-omic dataset derived from a comprehensive study [20].

We began by measuring the scaled cosine distance between the in silico perturbed latent matrix of COVID-19 cells and that of healthy cells (Fig. 6a) in the joint latent space produced by scCross. This allowed us to pinpoint signature marker genes that bridge these two conditions (i.e., diseased vs. healthy) via in silico perturbations (see "Methods" section for details). The identified critical disease-associated marker genes from this in silico perturbation, presented in Fig. 6b, are in line with recent findings. A notable observation was the down-perturbation of CCL3 in COVID-19 samples, suggesting its overexpression in COVID-19 patients. This concurs with prior studies that identified CCL3's role in inflammatory macrophages and its contribution to inflammatory tissue

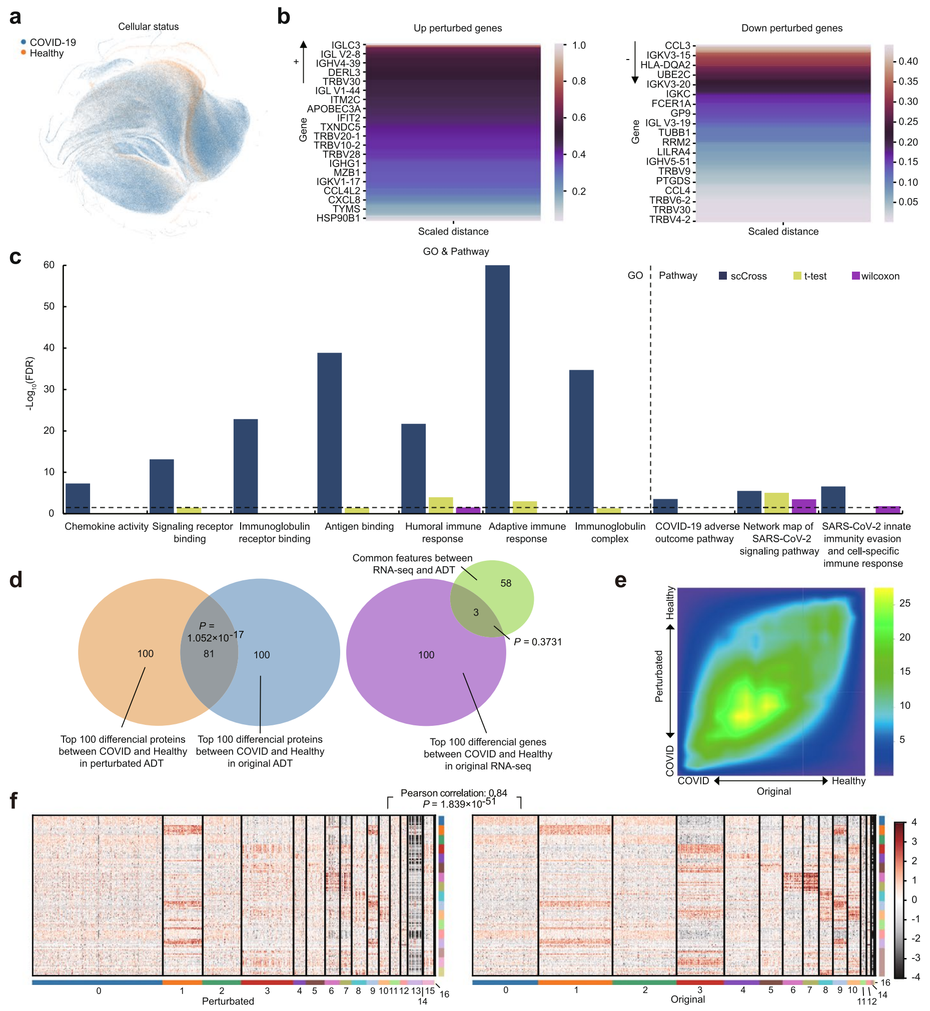


Fig. 6 scCross empowers in silico multi-omic perturbation as demonstrated in the COVID-19 dataset. **a** UMAP visualization of COVID-19 scRNA-seq and ADT data clusters. **b** Signature genes for in silico perturbation bridging COVID-19 and healthy states. **c** Pathways and GO terms of the top differential genes detected by scCross (down-perturbation gene findings), with -log(*p*-value) indicating significance, compared with *t*-test and wilconxon. **d** Gene comparison between scCross-derived perturbed protein data and common-gene based on scRNA-seq to protein translation. **e** RRHO plot comparing crossmodal perturbed and original COVID-19 protein data against healthy samples. **f** Heatmap aligning original and scCross-derived perturbed COVID-19 protein data

damage and respiratory issues in COVID-19 patients [60–62]. Furthermore, the GO and pathway outcomes from the down-perturbation gene findings, showcased in Fig. 6c, unveiled significant associations with immune responses. Aspects like signaling receptor binding, immunoglobulin receptor binding, antigen binding, humoral immune response, adaptive immune response, and immunoglobulin complex are highlighted. These observations align well with the recognized impact of immune responses and chemokine activations in the context of COVID-19 [61]. The performance on the COVID-19 adverse outcome pathway [63], Network map of SARS-CoV-2 signaling pathway [64], and SARS-CoV-2 innate immunity evasion and cell-specific immune response pathway [65] further validates scCross's potential in pinpointing signature genes between disease conditions within the same modality.

The aforementioned intra-modality perturbation analysis also enables us to identify critical cell subpopulations, reflecting known mechanisms in COVID-19. By applying in silico perturbations to repress highly variable genes in the COVID-19 group for each cell population (with more than 100 cells) and calculating the average distance divergences before and after perturbation for the top 10 genes, we found that NK_56hi cells were the most significantly affected (Additional file 1: Fig. S16). NK_56hi cells, crucial in COVID-19 development, exhibit superior cytokine production and cytotoxicity compared to their circulating counterparts [66]. Their presence indicates an essential role in early antiviral responses and modulation of adaptive immune responses. Additionally, a 2021 study showed that CD56bright NK cells in COVID-19 patients have a distinct transcriptional profile with upregulated proinflammatory genes and pathways, highlighting their involvement in the hyperactive immune response and antiviral immunity [67].

To further demonstrate our method's cross-generation-powered cross-modal in silico perturbation between two different biological conditions (healthy controls vs. COVID-19 patients), we trained our model on healthy control single-cell multi-omics data (RNA and protein). We then applied this model to the RNA modality of COVID-19 patients, enabling it to infer the disease-perturbed COVID-19 protein modality in silico. This approach showcases the model's ability to perform cross-generation even under different cellular conditions. The results of this more challenging scenario, depicted in Fig. 6d-f, affirm that our method is capable of effectively inferring the protein modality under disease perturbation, highlighting its robustness and versatility in in silico cross-modal perturbation studies.

Figure 6d delves into scCross's potential in performing cross-modality in silico perturbations. By specifically perturbing the gene expression of signature genes that potentially transition healthy cells to a COVID-19-like state in the healthy scRNA-seq data and utilizing cross-generation, we generate perturbed protein data (ADT) across modalities to simulate the COVID condition. This perturbed ADT, when juxtaposed with the healthy protein data, yielded differential proteins (List A). This list was then contrasted with the actual differential protein list (List B) obtained from single-cell ADT comparisons between healthy and COVID datasets. The profound overlap between the two lists showcases the accuracy of cross-modal in silico perturbations using scCross. Only three of the top 100 differentially expressed genes in scRNA-seq corresponded with those identified in single-cell protein measurements. This misalignment underscores the pitfalls of the naive strategy, which assumes that a gene differential at the RNA level is similarly differential at the protein level. Such discrepancies emphasize the efficacy and critical importance of scCross in these analyses.

Figure 6e, with its RRHO plot, further magnifies the similarity between the cross-modal perturbed ADT and the actual ADT COVID dataset. The heatmap in Fig. 6f underlines this alignment, with a p-value of 1.839×10^{-51} (via the Wilcoxon test) cementing the proximity between the perturbed protein data and actual single-cell protein measurements. Conclusively, the data presents a compelling case for scCross's ability to manage in silico perturbations across modalities, reiterating its value in intricate multi-omic cellular state explorations.

Empowered by its cross-modal generation and perturbation capabilities, scCross reveals novel biological insights that would otherwise be unattainable. For instance, the cross-modal generation capability of scCross enables the identification of critical marker proteins from single-cell RNA-seq data that would otherwise be missed without the matched single-cell protein modality measurement, such as CD38 and CLEC12A (Additional file 2: Table S1). CD38, involved in NAD metabolism, plays a pivotal role in COVID-19 pathogenesis and is upregulated in endothelial cells during SARS-CoV-2 infection [68, 69]. Notably, CLEC12A, identified through our in silico cross-modal perturbation analysis, could not be detected even with ground-truth single-cell multi-omics measurements of both healthy and COVID-19 samples. CLEC12A is linked to COVID-19 by potentially disrupting viral spike protein entry and supporting CD4/CD28 T cell survival. Targeting CLEC12A and other C-type lectins may enhance immune responses against COVID-19. These findings underscore scCross's power in revealing novel biological insights and therapeutic targets, thereby highlighting its potential to advance our understanding of complex diseases and uncover new avenues for treatment [70, 71].

Discussion

The ever-evolving domain of single-cell analyses, complemented by multi-omics strategies, continues to redefine our understanding of intricate cellular dynamics. Yet, seamlessly integrating these multi-omics dimensions has remained a challenge. In this context, our scCross method emerges, revolutionizing the way we perceive and amalgamate single-cell multi-omics data. The rigorous testing and evaluations underscore its superior performance, enabling more accurate and comprehensive multi-omics data integration.

But beyond mere integration, scCross introduces the concept of cross-modal generation. This capability allows researchers to derive data from one modality using another, once the model has been trained on a reference single-cell multi-omics dataset encompassing both source and target modalities. Given the often prohibitive cost and experimental challenges of single-cell multi-omics data acquisition, such functionality is crucial. Often, scientists have extensive data in one modality (e.g., RNA-seq) but limited or no profiling in others. With scCross, the knowledge acquired from a comprehensive multi-omics dataset can be channeled to generate data in other modalities from a given single modality, filling gaps in our understanding and providing a more complete view of cellular states.

Further accentuating its ability, scCross showcases an array of functionalities tailored for the demands of modern cellular research. Its memory efficiency stands out, ensuring that even datasets with millions of cells are integrated with finesse, making it apt for the vast cell atlases being assembled. Simultaneously, its prowess extends to its capacity for single-cell multi-omics data simulation. Through judicious sampling from the joint latent space and adjusting omic-decoders, it can simulate matched single-cell datasets across various modalities for the same consistent set of cells. This serves not only for data augmentation but also as a benchmarking tool, given the known ground truth in these simulated datasets. Building on these capabilities, scCross offers the powerful in silico perturbation functionality. As evidenced in cases like COVID-19 cellular dynamics, it offers a fresh perspective to formulate and probe potential intervention strategies to modulate cellular states. Once primed with a reference multi-omics dataset, it can be harnessed on single-modality data, enabling intricate multi-omic in silico explorations.

Moreover, the power of in silico intra-modal and cross-modal perturbation lies in its ability to reveal novel biological insights that would otherwise be unattainable. This includes the identification of novel biomarkers and critical cell populations associated with diseases, providing additional layers of understanding and potential therapeutic targets.

The scCross method presents significant potential for the single-cell research community. Its distinct functionalities and reliable performance position it as a potentially valuable tool for researchers in the field of single-cell multi-omics analyses. scCross facilitates the integration of different modalities, supports comprehensive data generation, and enables detailed simulations and perturbations. These capabilities could contribute substantially to advancing the study of complex biological systems.

Conclusions

scCross has the potential to significantly influence the field of single-cell multi-omics research. Its ability to bridge modalities effectively, along with capabilities in cross-modal generation, multi-omics simulation, and augmentation, may enhance both the speed and depth of biomedical research. Additionally, its suitability for in silico explorations offers practical applications beyond academic research, potentially assisting in the design of targeted experiments for cellular state manipulations and possibly hastening the translation of biomedical research into real-world applications.

Methods

Single-cell data processing

The single-cell data employed in our study follows specific preprocessing pipelines tailored to each modality. Initial procedures, such as cell calling, quality control, and other relevant steps, can follow standard single-cell data processing methods and are often performed by the original data sources [20, 36, 38, 39]. Within our workflow, these matrices are further transformed to produce cell vectors optimized for neural network input. Besides, these matrices' feature sets are denoted as $\mathcal{S}_k$, where k ranges from 1 to K. K symbolizes the number of distinct omics data types collected. For example, in scRNA-seq, $\mathcal{S}_k$ represents a gene list, while in scATAC-seq, it represents a set of chromatin regions. Profiling matrices from the k^{th} modality is represented as $\mathcal{X}_k$. Individual cells from this modality are given by $\mathbf{x}_k^{(n)}$, and N_k indicates the sample size.

For single-cell transcriptomic data, specifically scRNA-seq, we adopt Scanpy [12] (v.1.8.2). The cell-by-gene expression matrix experiences normalization, log-transformation, and scaling. Dimensionality reduction follows, with PCA, as realized in Scanpy [72], being our method of choice. On the epigenetic data front, modalities like scATAC-seq and snmC-seq demand distinct processes. SnmC-seq data transitions through a sequence of normalization, log-transformation, scaling, and PCA, utilizing epiScanpy [73] (v.0.4.0). Given the inherent sparsity of the scATAC-seq data matrix, dimensionality reduction employs the LSI (Latent Semantic Indexing) function of scCross, grounded in the latent semantic indexing algorithm [17, 74].

To facilitate the calculations of gene sets matrix and common gene MNN priors, it is crucial to have a consistent gene-centric representation across different modalities. Due to the inherent nature of MNN priors, the single-cell data from diverse modalities needs to be translated into activities associated with genes. To achieve this uniform representation, we employ the geneactivity function in epiScanpy [73] (v.0.4.0), transforming various omics data types (like scATAC-seq and snmC-seq) into gene activity. The gene-formed data (gene expression/gene activity score) of each modality k is represented with $\dot{\mathbf{x}}_{\mathbf{k}}$. It is obvious that $\dot{\mathbf{x}}_{\mathbf{k}} = \mathbf{x}_{\mathbf{k}}$ when k^{th} modality is scRNA-seq.

It's noteworthy that our model's applicability isn't confined to the discussed modalities. For any modality, the transformation of each cell into a real-valued vector is a must. Often, normalization is required to address technical nuances, such as variations in library size among cells. For modalities not tackled in this work, we advocate adhering to the modality-specific standard single-cell preprocessing protocol when generating the input vectors.

Curating gene sets from Gene Ontology and pathway terms

In our study, gene sets $\mathcal{GS}$ are tapped as an avenue to integrate functional biological priors, enhancing the representation learning of cells across different modalities. This, in turn, bolsters the effectiveness of single-cell data integration and generation efforts, as these tasks are fundamentally rooted in effective cell embeddings.

We utilize a rich collection of gene sets, comprising 7481 sets from the GO Biological Process ontology (denoted as c5.go.bp in MSigDB [75]), 2922 gene sets curated from pathway databases (c2.cp in MSigDB [75]), and 335 sets of transcription factor targets sourced from [76]. Gene sets related to biological processes bear the prefix "GOBP," while those derived from pathway databases are assigned prefixes echoing their respective sources, such as KEGG [77], WP [78], and REACTOME [79].

To effectively harness the gene sets, $\mathcal{GS}$, each gene set is characterized based on the genes it contains. In the context of scRNA-seq data, a gene set is represented by the expression levels of its constituent genes. This provides a direct quantification of the gene set's activity within a cell, crucial for grasping cellular states and functions. However, with non-transcriptomic modalities, such as scATAC-seq, where direct gene expression values are absent, we resort to the geneactivity function from epiScanpy [73] (v.0.4.0). This tool calculates a gene activity matrix, serving as an effective surrogate for gene expression. Using this matrix, we can obtain the binary matrix $\tilde{\mathbf{x}}_k$, which is the correlation of genes in gene expression/activity matrix and genesets in the databases we mentioned above. We utilize the approach of [80] where we set the correlation of a gene to a geneset to 1 if the gene is contained in the geneset, 0 otherwise. Then, we obtain the cell by geneset matrix with the function:

$$\mathcal{GS}_k = \dot{\mathbf{x}}_k \cdot \tilde{\mathbf{x}}_k \tag{1}$$

where $\mathcal{GS}_k$ means the cell by geneset feature matrix of modality k. These geneset scores, once extracted, are integrated with the processed input features, contributing to solid single-cell representation learning and multi-omics integration (Additional file 1: Fig. S17).

Deep generative neural network structure

To learn cell embeddings from omics data, we employ a variational autoencoder (VAE).

Our model initially projects $\mathbf{x}_k$ into a low-dimensional latent space, $\mathbf{z}_k$, utilizing a variational approach. The encoder function is formulated as:

$$\begin{aligned} &q\left(\mathbf{z}_k \mid \mathbf{x}_k, \mathcal{GS}_k; \phi_{en_k}\right) = \\ N\Big(&\mathbf{z}_k; \mathrm{MLP}_{en_{k,\mu}}\left(\mathrm{CONCATE}(\mathrm{GCN}(\mathbf{x}_k, \mathcal{G}_k), \mathcal{GS}_k); \phi_{en_k}\right), \\ &\mathrm{MLP}_{en_{k,\sigma^2}}\left(\mathrm{CONCATE}(\mathrm{GCN}(\mathbf{x}_k, \mathcal{G}_k), \mathcal{GS}_k); \phi_{en_k}\right)\Big) \end{aligned} \tag{2}$$

Sampling directly from this distribution during the training process can be problematic for backpropagation because it introduces stochasticity. To overcome this, we use the reparameterization trick, which involves expressing $\mathbf{z}_k$ as a deterministic function of $\mathrm{MLP}_{en_{k,\mu}}\left(\mathrm{CONCATE}(\mathrm{GCN}(\mathbf{x}_k, \mathcal{G}_k), \mathcal{GS}_k); \phi_{en_k}\right)$ and $\mathrm{MLP}_{en_{k,\sigma^2}}\left(\mathrm{CONCATE}(\mathrm{GCN}(\mathbf{x}_k, \mathcal{G}_k), \mathcal{GS}_k); \phi_{en_k}\right)$, and a noise variable ϵ drawn from a standard normal distribution:

$$\begin{aligned} \mathbf{z}_k &= \mathrm{MLP}_{en_{k,\mu}}\left(\mathrm{CONCATE}(\mathrm{GCN}(\mathbf{x}_k, \mathcal{G}_k), \mathcal{GS}_k); \phi_{en_k}\right) \\ &+ \mathrm{MLP}_{en_{k,\sigma^2}}\left(\mathrm{CONCATE}(\mathrm{GCN}(\mathbf{x}_k, \mathcal{G}_k), \mathcal{GS}_k); \phi_{en_k}\right) \odot \epsilon, \sim N(0,1) \end{aligned} \tag{3}$$

Here, $q(\mathbf{z}_k \mid \mathbf{x}_k, \mathcal{GS}_k; \phi_{en_k})$ describes the encoder part of the VAE for the k^{th} modality, CONCATE means concatenating two matrices together. It uses a graph convolutional network (GCN) to model the latent space $\mathbf{z}_k$ given the observed data $\mathbf{x}_k$.

$$\mathcal{G}_k = \mathrm{KNN}(\mathbf{x}_k) \tag{4}$$

The graph $\mathcal{G}_k$ is constructed using the k-nearest-neighbor algorithm and serves as the adjacency matrix for the GCN.

The data likelihoods $p(\mathbf{x}_k \mid \mathbf{z}_k; \theta_k)$, which represent the data decoders, are constructed using a multilayer perceptron. The specific formulation of the data likelihood depends on the distribution of the omics data. In the case of count-based scRNA-seq and scATAC-seq data, our model utilizes the negative binomial (NB) distribution. The NB distribution is defined as follows:

$$p\left(\mathbf{x}_k \mid \mathbf{z}_k; \theta_k, \phi_{de_k}\right) = \prod_{i \in \mathcal{S}_k} \mathrm{NB}\left(\mathbf{x}_{\mathrm{k}}^{(\mathrm{i})}; \mu^{(\mathrm{i})}, \theta^{(\mathrm{i})}, \phi_{\mathrm{de}_\mathrm{k}}\right) \tag{5}$$

$$\begin{aligned} &\mathrm{NB}\left(\mathbf{x}_{\mathrm{k}}^{(\mathrm{i})}; \mu^{(\mathrm{i})}, \theta^{(\mathrm{i})}, \phi_{\mathrm{de}_\mathrm{k}}\right) = \\ &\frac{\Gamma\left(\mathbf{x}_k^{(i)} + \theta^{(i)}\right)}{\Gamma\left(\theta^{(i)}\right)\Gamma\left(\mathbf{x}_k^{(i)} + 1\right)} \left(\frac{\mu^{(i)}}{\theta^{(i)} + \mu^{(i)}}\right)^{\mathbf{x}_k^{(i)}} \left(\frac{\theta^{(i)}}{\theta^{(i)} + \mu^{(i)}}\right)^{\theta^{(i)}} \end{aligned} \tag{6}$$

In the equation above, $\mu, \theta \in \mathbb{R}_+^{|\mathcal{S}_k|}$ are the mean and dispersion of the negative binomial distribution. Additionally, $\alpha \in \mathbb{R}_+^{|\mathcal{S}_k|}, \beta \in \mathbb{R}^{|\mathcal{S}_k|}$ are scaling and bias factors. The Hadamard product is denoted by $\odot$, $\mathrm{Softmax}^{(\mathrm{i})}$ represents the i^{th} dimension of the softmax output, and $\sum_{v \in \mathcal{S}_k} \mathbf{x}_k^{(v)}$ provides the total count in the cell ensuring the library size of reconstructed data matches the original.

$$\mu^{(i)} = \mathrm{Softmax}^{(\mathrm{i})}\left(\alpha \odot \mathrm{MLP}_{de_k}\left(\mathbf{z}_k, \phi_{de_k}\right) + \beta\right) \cdot \sum_{v \in \mathcal{S}_k} \mathbf{x}_k^{(v)} \tag{7}$$

The set of learnable parameters in this context is $\theta_k = \{\alpha, \beta, \theta\}$. Meanwhile, ϕ_{de_k} denotes the set of learnable parameters in the multilayer perceptron decoder of the k^{th} omics layer, MLP_{de_k}. The flexibility of this model also allows for the incorporation of many other distributions.

Finally, the optimization process aims to maximize the evidence lower bound:

$$p(\mathbf{z}_k) = N\left(\mathbf{z}_k; [0]^m, [1]^m\right) \tag{8}$$

$$\begin{gathered} \mathcal{L}'_{\mathcal{X}}(\theta, \phi_{en}, \phi_{de}) = \\ \sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}_k \sim p_{\text{ data }}(\mathbf{x}_k)} \left[\begin{matrix} \mathbb{E}_{\mathbf{z}_k \sim q\left(\mathbf{z}_k | \mathbf{x}_k, \mathcal{GS}_k; \phi_{en_k}\right)} \log p\left(\mathbf{x}_k \mid \mathbf{z}_k; \theta_k, \phi_{de_k}\right) \\ -\mathrm{KL}\left(q\left(\mathbf{z}_k \mid \mathbf{x}_k, \mathcal{GS}_k; \phi_{en_k}\right) \| p(\mathbf{z}_k)\right) \end{matrix} \right] \end{gathered} \tag{9}$$

This equation represents the loss function that the VAE optimizes, where m means the length of the latent space vector which is 50 and $\theta = \bigcup_{k=1}^{K} \theta_k$, $\phi_{en} = \bigcup_{k=1}^{K} \phi_{en_k}$, $\phi_{de} = \bigcup_{k=1}^{K} \phi_{de_k}$ are the combined sets of the encoder and decoder parameters, respectively. It incorporates both the data likelihood and a Kullback-Leibler (KL) divergence term to enforce a meaningful latent space for each of the K modalities.

Multi-omics data integration, crossmodal generation and simulation

To harmonize the latent embeddings produced from different omics modalities, we implement a cross-modal bidirectional alignment method utilizing multi-layer perceptrons (MLPs):

$$\begin{gathered} q\left(\mathbf{z} \mid \mathbf{z}_k; \phi_{al\_en_k}\right) = \\ N\left(\mathrm{z}; \mathrm{MLP}_{\mathrm{al\_en_k},\mu}\left(\mathbf{z}_\mathrm{k}; \phi_{\mathrm{al\_en_k}}\right), \mathrm{MLP}_{\mathrm{al\_en_k},\sigma^2}\left(\mathbf{z}_\mathrm{k}; \phi_{\mathrm{al\_en_k}}\right)\right) \end{gathered} \tag{10}$$

Employing reparameterization trick, the above equation can be written as follows:

$$\mathrm{z} = \mathrm{MLP}_{\mathrm{al\_en_k},\mu}\left(\mathrm{z_k}; \phi_{\mathrm{al\_en_k}}\right) + \mathrm{MLP}_{\mathrm{al\_en_k},\sigma^2}\left(\mathrm{z_k}; \phi_{\mathrm{al\_en_k}}\right) \odot \epsilon, \epsilon \sim \mathrm{N}(0, 1) \tag{11}$$

This reparameterized form shows the process of encoding $\mathrm{z_k}$ to z. It reflects that latent variable z is sampled from the normal distribution showed in the formula above.

$$\begin{gathered} p\left(\mathbf{z}_k \mid \mathbf{z}; \phi_{al\_de_k}\right) = \\ N\left(z_k; \mathrm{MLP}_{al\_de_k,\mu}\left(\mathbf{z}; \phi_{al\_de_k}\right), \mathrm{MLP}_{al\_de_k,\sigma^2}\left(\mathbf{z}; \phi_{al\_de_k}\right)\right) \end{gathered} \tag{12}$$

Similarly, this can also be reparameterized as:

$$\mathrm{z_k} = \mathrm{MLP}_{\mathrm{al\_de_k},\mu}\left(\mathrm{z}; \phi_{\mathrm{al\_de_k}}\right) + \mathrm{MLP}_{\mathrm{al\_de_k},\sigma^2}\left(\mathrm{z}; \phi_{\mathrm{al\_de_k}}\right) \odot \epsilon, \epsilon \sim \mathrm{N}(0, 1) \tag{13}$$

In these equations, $\mathrm{MLP}_{al\_en_k}$ denotes the aligner encoder, which converts the latent embedding of the k^{th} omics layer to a collective latent space. On the other hand, $\mathrm{MLP}_{al\_de_k}$ represents the aligner decoder, which translates the unified latent space embedding back to the latent space of the k^{th} omics modality. This alignment process is key to converting single-cell multi-omics data into a shared latent representation, denoted as **z**. By establishing such a common latent space, we facilitate an effective integration of disparate omics data.

The architecture of scCross operates on a two-step variational autoencoder (VAE) framework to encode omics layers into a merged space. Inverting this methodology, any encoded data in this unified space can be reverted to any particular omics layer's latent representation using a dual-step decoding procedure. The benefit here is the cost-efficiency: rather than requiring *k*! distinct encoder/decoder combinations to synchronize any two modalities, this method necessitates only a couple. This streamlined approach facilitates multi-omics simulations across any number of omics layers. Nevertheless, the method's success heavily depends on the precise integration of data in the shared latent space. If discrepancies emerge in this unified space between modalities, the generated data will mirror those differences.

By leveraging a bi-directional aligner, the cross-modal generation of data from modality *j* to modality *h* can be conceptualized with the following equation:

$$\mathbf{x}_{h\_cross} \sim p\left(\mathbf{x}_h \mid p\left(\mathbf{z}_h \mid q\left(\mathbf{z} \mid q\left(\mathbf{z}_j \mid \mathbf{x}_j, \mathcal{GS}_j; \phi_{en_j}\right); \phi_{al\_en_j}\right); \phi_{al\_de_h}\right); \theta_h, \phi_{de_h}\right) \tag{14}$$

where x_{h_cross} means cross generated x_h from x_j. A visual representation of the described procedure is illustrated in Additional file 1: Fig. S1a. The process begins with the input vector from modality *j*, which undergoes dimensionality reduction and mapping to the joint latent space *z*. Subsequently, the decoder associated with modality *h* is employed to reconstruct *z* into the data vector for modality *h*.

Notably, when *j* is not equal to *h*, this process facilitates cross-modal generation, allowing the transformation of information from one modality to another. Conversely, when *j* is equal to *h*, it becomes a mechanism for self-augmentation within the same modality. A visual representation of the described procedure is illustrated in Additional file 1: Fig. S1b.

Similarly, single-cell data simulations tailored to specific cellular states for a particular modality *h* are described as:

$$\mathbf{x}^s_{h\_simu} \sim p\left(\mathbf{x}_h \mid p\left(\mathbf{z}_h \mid \mathbf{z}^{s'}; \phi_{al\_de_h}\right); \theta_h, \phi_{de_h}\right) \tag{15}$$

$$\mathbf{z}^{s'} \sim q\left(\mathbf{z} \mid q\left(\mathbf{z}_1 \mid \mathbf{x}^s_1, \mathcal{GS}^s_1; \phi_{en_h}\right), \cdots, q\left(\mathbf{z}_K \mid \mathbf{x}^s_K, \mathcal{GS}^s_K; \phi_{en_h}\right); \phi_{al\_en_h}\right) \tag{16}$$

In this context, $\mathbf{x}^s_{h\_simu}$ means simulated data of *h* modality for specific cellular states, $\mathbf{x}^s_k$ pertains to the feature matrices of specific cellular states for k^{th} modality, $\mathcal{GS}^s_k$ is the gene set matrices for k^{th} modality of specific cellular states, $\mathbf{z}^{s'}$ represents the sample result of $q\left(\mathbf{z} \mid q\left(\mathbf{z}_1 \mid \mathbf{x}^s_1, \mathcal{GS}^s_1; \phi_{en_h}\right), \cdots, q\left(\mathbf{z}_K \mid \mathbf{x}^s_K, \mathcal{GS}^s_K; \phi_{en_h}\right); \phi_{al\_en_h}\right)$. To simulate single-cell multi-omics for a certain cell population. We first sample in the joint latent space to produce latent embedding vectors to represent cells from the specific cell population. The number of samples determines the number of simulated cells. The latent embedding z_h will then be converted back into the input vector x_h via specific modality-specific decoder. A visual representation of the described procedure is illustrated in Additional file 1: Fig. S1c.

The model's fitting procedure seeks to maximize the evidence lower bound expressed as:

$$\mathcal{L}_{\mathcal{X}}(\theta,\phi) = \sum\nolimits_{k=1}^{K} \mathbb{E}_{\mathbf{x}_k \sim p_{\text{ data }}(\mathbf{x}_k)} \left[\begin{array}{c} \mathbb{E}_{\mathbf{z}_k \sim q(\mathbf{z}_k \mid \mathbf{x}_k, \mathcal{GS}_k; \phi_k)} \log p(\mathbf{x}_k \mid \mathbf{z}_k; \theta_k) \\ -\mathrm{KL}(q(\mathbf{z}_k \mid \mathbf{x}_k, \mathcal{GS}_k; \phi_k) \| p(\mathbf{z}_k)) \\ -\mathrm{KL}\Big(q\big(\mathbf{z} \mid \mathbf{z}_k; \phi_{al\_en_k}\big) \| \frac{1}{K} \sum\nolimits_{l=1}^{K} q(\mathbf{z}_l \mid \mathbf{x}_l, \mathcal{GS}_l; \phi_l)\Big) \end{array} \right] \tag{17}$$

Here, the set of the encoder and decoder parameters are expressed as $\phi = \bigcup_{k=1}^{K} \{\phi_{en_k}, \phi_{de_k}, \phi_{al\_en_k}, \phi_{al\_de_k}\}$ and $\theta = \bigcup_{k=1}^{K} \theta_k$, respectively.

High-quality data integration and generation via generative adversarial learning

In the endeavor to ensure precise alignment of diverse omics data, we harness the potential of a generative adversarial learning strategy, a method that showcases its prowess in numerous prior studies [29, 81].

Central to our model is a discriminator, denoted as $\mathrm{D_z}$, equipped with a K-dimensional softmax output, designed specifically for the seamless alignment of various omics layers [38, 82]. Functioning based on the latent space embeddings of cells, denoted as z, $\mathrm{D_z}$ endeavors to identify the omics layer origin of the cells. The training of this discriminator is steered by the minimization of the multi-class classification cross entropy:

$$\mathcal{L}_{\mathrm{D_z}}(\phi,\psi) = -\frac{1}{K} \sum\nolimits_{k=1}^{K} \mathbb{E}_{\mathbf{x}_k \sim p_{\text{data}}(\mathbf{x}_k)} \mathbb{E}_{\mathbf{z}_k \sim q\left(\mathbf{z}_k | \mathbf{x}_k, \mathcal{GS}_k; \phi_{en_k}\right)} \mathbb{E}_{\mathbf{z} \sim q\left(\mathbf{z} | \mathbf{z}_k; \phi_{al\_en_k}\right)} \log \mathrm{D}_{\mathrm{z_k}}(\mathbf{z}; \psi) \tag{18}$$

Here, $\mathrm{D_{z_k}}$ symbolizes the *k*th dimension of the discriminator output, while ψ refers to the suite of parameters, amenable to learning, within the discriminator. Contrarily, the data encoders are trained with the ambition to deceive this discriminator, a strategic move to fortify the alignment of cell embeddings stemming from distinct omics layers.

Furthermore, to bolster the generation of cross-modality data, we present unique discriminators for each modality. This approach aids in refining the reconstruction of the data:

$$\begin{aligned} &\mathcal{L}_{\mathrm{D_{ge}}}(\phi,\theta,\delta) = \\ &-\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}_k \sim p_{\text{data}}(\mathbf{x}_k)} \mathbb{E}_{\mathbf{z}_k \sim q\left(\mathbf{z}_k | \mathbf{x}_k, \mathcal{GS}_k; \phi_{en_k}\right)} \mathbb{E}_{\mathbf{z} \sim q\left(\mathbf{z} | \mathbf{z}_k; \phi_{al\_en_k}\right)} \\ &\mathbb{E}_{\hat{\mathbf{z}}_k \sim p\left(\hat{\mathbf{z}}_k | \mathbf{z}; \phi_{al\_de_k}\right)} \mathbb{E}_{\hat{\mathbf{x}}_k \sim p\left(\hat{\mathbf{x}}_k | \hat{\mathbf{z}}_k; \theta_k, \phi_{de_k}\right)} \log \mathrm{D}_{\mathrm{ge_k}}((\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{\mathrm{k}}); \delta_{\mathrm{k}}) \end{aligned} \tag{19}$$

$$\mathcal{L}_{\mathrm{D}}(\phi,\psi,\theta,\delta) = \mathcal{L}_{\mathrm{D_z}}(\phi,\psi) + \mathcal{L}_{\mathrm{D_{ge}}}(\phi,\theta,\delta) \tag{20}$$

Within this context, $\delta = \bigcup_{k=1}^{K} \delta_k$ encapsulates the collection of learnable parameters in the discriminators. Additionally, $\mathrm{D_{ge_k}}$ represents the output from the k^{th} discriminator, tailored to reinforce the congruence between the model-generated data $\hat{\mathbf{x}}_k$ and the real data $\mathbf{x}_k$ for the k^{th} omics layers. The terms $\hat{\mathbf{z}}_k$ and $\hat{\mathbf{x}}_k$ depict the data generated by the aligner decoder and VAE decoder for the k^{th} modality, respectively.

Common gene MNN prior to align cells in similar cellular states

For an effective alignment across various modalities, it is essential to harness common genes that span across modalities, such as gene expression in scRNA-seq and gene activity in scATAC-seq and snmC-seq. These genes serve as a pivotal anchor, ensuring consistency and coherence across modalities. While previous omics data integration methods have been successful in achieving a semblance of agreement in the data distribution, the main challenge remains. The true test of alignment is whether embeddings of the same or similar cells across different modalities can be represented identically or similarly within the shared latent space.

Here, the effective multi-modal alignment is achieved by utilizing the mutual nearest neighbors (MNN) approach, which has been shown to effectively align cells of analogous cellular states across different omics data [83]. First, each modality is represented by a matrix, $\dot{\mathbf{x}}$, comprised of cells by common genes. Subsequently, the MNN-corrected common genes matrix from each modality is processed using PCA. This produces a common gene prior, G, which encapsulates all modalities:

$$G = \mathrm{PCA}(\mathrm{MNN\_CORRECT}(\dot{\mathbf{x}}_1, \ldots, \dot{\mathbf{x}}_k)). \tag{21}$$

The alignment loss function, $\mathcal{L}_{\mathrm{MNN}}$, is then formulated to minimize the difference between cosine similarities of latent embeddings of cells and their counterparts based on common gene priors, ensuring consistency in the joint space:

$$\mathcal{L}_{\mathrm{MNN}} = \sum_{h \neq j}^{K} \left\| \left(\mathbf{z}'_h \cdot \mathbf{z}'^{\mathsf{T}}_j \right) - \lambda_G \left(G_h \cdot G_j^{\mathsf{T}} \right) \right\|^2 \tag{22}$$

where $\mathbf{z}'_h \sim q\left(\mathbf{z} \mid \mathbf{z}_h; \phi_{al\_en_h}\right)$, $\mathbf{z}'_j \sim q\left(\mathbf{z} \mid \mathbf{z}_j; \phi_{al\_en_j}\right)$ and λ_G are the cell embedding in the shared space for the modality h and j, respectively, generally set to 1, assists in scaling the difference between the two subtractions. In this equation, G_h and G_j refer to the common gene priors for modality h and j corresponding to the cells' embedding $\mathbf{z}'_h$ and $\mathbf{z}'_j$, respectively.

Besides, it is important to note that despite this transformation for training purposes, both the input and output data of our method in all functions remain in their original feature forms. For instance, in the scenario of cross-generating from scRNA-seq to scATAC-seq, the resulting scATAC-seq data retains its original peak features, not the transformed genes. The MNN pairs computed based on the converged gene matrix (e.g., from the ATAC-seq peaks) serve only as anchors to align cells from two modalities better. Nevertheless, the cells coming from each modality (e.g., single-cell ATAC-seq) are still represented by their original feature vector (e.g., peak vector) to avoid information loss. This approach ensures that the integrity and specificity of the original feature data are preserved throughout the integration process.

Overall training strategy

The training process for scCross unfolds over two distinct stages. Initially, the primary objective is the isolated training of a variational autoencoder (VAE) for each modality. This approach enables us to capture and understand the unique biological information inherent within each modality before proceeding to their integration. The loss function for this first stage is represented as:

$$\max_{\theta,\phi,\psi,\delta} \mathcal{L}'_{\mathcal{X}}(\theta,\phi_{en},\phi_{de}) \tag{23}$$

Transitioning to the second stage, the trained VAE models from the preliminary stage serve as a foundation. At this juncture, bi-directional aligners are incorporated, merging the modalities into a unified latent space. Discriminators also come into play during this stage, contributing to the integration of multi-omics data. The associated training losses for this stage are detailed as:

$$\min_{\psi,\delta} \lambda_D \mathcal{L}_{\mathrm{D}}(\phi,\psi,\theta,\delta) \tag{24}$$

$$\max_{\theta,\phi,\psi,\delta} \mathcal{L}_{\mathcal{X}}(\theta,\phi) + \lambda_D \mathcal{L}_{\mathrm{D}}(\phi,\psi,\theta,\delta) - \lambda_{MNN} \mathcal{L}_{\mathrm{MNN}} \tag{25}$$

In these equations, the terms λ_D and λ_{MNN} dictate the relative contributions of the discriminator and the common gene MNN prior, respectively. To optimize the training of the scCross model, we utilize the stochastic gradient descent technique. Notably, the RMSprop optimizer, devoid of a momentum term, is selected to enhance stability during adversarial training.

Single-cell multi-omics data integration

As highlighted as among the best for single-cell multi-omics data integration in recent reviews [84, 85], We selected scglue [29], Seurat v4 [28], and uniPort [30] for our performance comparison. These methods were specifically chosen because they accommodate both matched and unmatched datasets, similar to our approach. For instance, scglue [29] is noted for its superior performance compared to other integration methods such as UnionCom [86], Pamona [87], and several others. Seurat v4 [28] is renowned for its widespread use and has demonstrated enhanced performance over methods like totalVI [88] and MOFA+ [89]. uniPort [30] has been shown to excel in multi-omic data integration over alternatives including scglue [29] and Harmony [9], among others.

To further enhance our benchmarking, we have incorporated additional methods such as scDART [32] and sciCAN [31] due to their proven effectiveness in existing benchmarks. scDART [32], for example, outperforms LIGER [90], Seurat v3 [91], and UnionCom [86]. Similarly, sciCAN [31] demonstrates superior integration capabilities compared to methods like LIGER [90], Harmony [9], and ArchR [92]. We also included Harmony [9] to broaden the scope of our comparison.

We evaluate the agreement between the cell type labels and the Leiden algorithm [93] clusters obtained from the integrated dataset for each omic separately using Normalized Mutual Information (NMI). NMI measures the overlap between two clusterings, with scores ranging from 0 (no overlap) to 1 (perfect agreement). We perform the Leiden clustering [93] with a resolution of 1 for all methods and datasets. We use the scikit-learn [94] (v.0.22.1) implementation of NMI. The NMI scores in our benchmarking were individually calculated for each omic and then averaged across all omics to obtain a comprehensive evaluation.

The Rand index is another metric that compares the overlap of two clusterings. We compare the cell type labels with the Leiden clustering [93] computed on the integrated dataset for each omic separately. An ARI of 0 or 1 corresponds to uncorrelated clustering or a perfect match, respectively. We perform the Leiden clustering [93] with a resolution of 1 for all methods and datasets. We use the scikit-learn [94] (v.0.22.1) implementation of the ARI. The ARI scores in our benchmarking were individually calculated for each omic and then averaged across all omics to obtain a comprehensive evaluation.

In our work, there are two kinds of ASW. One is cell type ASW, another is omics layer ASW. Cell type ASW is used to evaluate the cell type resolution. It can metrix model's clustering ability. Cell type ASW is defined as in a recent benchmark study [29, 40]:

$$\text{cell type ASW} = \frac{1}{2}\left(\frac{\sum_{i=1}^{N} s_{\text{cell type}}^{(i)}}{N} + 1\right) \tag{26}$$

where $s_{\text{cell type}}^{(i)}$ is the cell type silhouette width for the i^{th} cell, and N is the total number of cells. Cell type ASW has a range of 0 to 1. Higher values indicate better cell type clustering resolution.

Omics layer ASW is used to evaluate the integration mixing ability among omics layers. It is defined as in a recent benchmark study[29, 40]:

$$\text{omics layer ASW} = \frac{1}{\mathrm{M}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \text{omics layer ASW}_{\mathrm{j}} \tag{27}$$

$$\text{omics layer ASW}_{\mathrm{j}} = \frac{1}{\mathrm{N_j}}\sum_{\mathrm{i}=1}^{\mathrm{N_j}} 1 - \left|\mathrm{s}_{\text{omics layer}}^{(\mathrm{i})}\right| \tag{28}$$

where $s_{\text{omics layer}}^{(i)}$ is the omics layer silhouette width for the i^{th} cell, N_j is the number of cells in cell type j, and M is the total number of cell types. Omics layer ASW has a range of 0 to 1. Higher values indicate better integration mixing ability.

The graph connectivity metric [29, 40] is used to estimate the ability to mix different omics data:

$$\mathrm{GC} = \frac{1}{\mathrm{M}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \frac{\left|\mathrm{LCC_j}\right|}{\mathrm{N_j}} \tag{29}$$

where $\mathrm{LCC_j}$ is the number of cells in the largest connected component of the cell k-nearest neighbors graph ($K = 15$) for cell type j, N_j is the number of cells in cell type j and M is the total number of cell types. Graph connectivity has a range of 0 to 1. Higher values of graph connectivity indicate better mixing ability.

The FOSCTTM metric [29, 41] is used to evaluate the single-cell level alignment accuracy. It is computed on two datasets with known cell-to-cell pairings. Suppose that each dataset contains N cells, and that the cells are sorted in the same order, that is, the i^{th} cell in the first dataset is paired with the i^{th} cell in the second dataset. Denote x and y as the cell embeddings of the first and second dataset, respectively. The FOSCTTM is then defined as:

$$\begin{aligned} \mathrm{FOSCTTM} &= \frac{1}{2N}\left(\sum_{i=1}^{N}\frac{n_1^{(i)}}{N}+\sum_{i=1}^{N}\frac{n_2^{(i)}}{N}\right) \\ n_1^{(i)} &= \left|\left\{t \mid d\left(\mathbf{x}^{(t)},\mathbf{y}^{(i)}\right) < d\left(\mathbf{x}^{(i)},\mathbf{y}^{(i)}\right)\right\}\right| \\ n_2^{(i)} &= \left|\left\{t \mid d\left(\mathbf{x}^{(i)},\mathbf{y}^{(t)}\right) < d\left(\mathbf{x}^{(i)},\mathbf{y}^{(i)}\right)\right\}\right| \end{aligned} \tag{30}$$

where $n_1^{(i)}$ and $n_2^{(i)}$ represent the number of cells in the first and second datasets that are closer to the i^{th} cell than their true matches in the opposite dataset. The Euclidean distance, denoted as d, is used to calculate the distance between cells. FOSCTTM ranges from 0 to 1, and lower values indicate higher accuracy.

To train the atlas level large-scale data, we made several adaptations to our training process. We employed the Leiden algorithm [93] to cluster the original scRNA-seq and scATAC-seq data and obtained the cluster numbers. Then, we aggregated and averaged the cells within each cluster to create a meta cell representing that cluster. The common gene MNN prior for cells within each cluster was calculated based on these meta cells. To boost training efficiency, we utilized the MNN loss, calculated based on the meta cells instead of all cells, for the second stage of fine-tuning scCross (loss defined in Eq. 22). Finally, we utilized Scanpy [12] (v.1.8.2) to generate the UMAP visualization.

In the triple omics integration, as the snmC-seq data's distribution is close to zero-inflated log-normal (ZILN) distribution, we use that distribution in our data decoder:

$$p\left(\mathbf{x}_k \mid \mathbf{z}_k; \theta_k, \phi_{de_k}\right) = \prod_{i\in\mathcal{S}_k} \mathrm{ZILN}\left(\mathbf{x}_k^{(i)}; \mu^{(i)}, \sigma^{(i)}, \delta^{(i)}, \phi_{de_k}\right) \tag{31}$$

$$\begin{gathered} \mathrm{ZILN}\left(\mathbf{x}_k^{(i)}; \mu^{(i)}, \sigma^{(i)}, \delta^{(i)}, \phi_{de_k}\right) = \\ \begin{cases} \frac{1-\delta^{(i)}}{\mathbf{x}_k^{(i)}\sigma^{(i)}\sqrt{2\pi}} \exp\left(-\frac{\left(\log \mathbf{x}_k^{(i)} - \mu^{(i)}\right)^2}{2{\sigma^{(i)}}^2}\right), & \mathbf{x}_k^{(i)} > 0 \\ \delta^{(i)}, & \mathbf{x}_k^{(i)} = 0 \end{cases} \end{gathered} \tag{32}$$

In the equation above, $\mu \in \mathbb{R}^{|\mathcal{S}_k|}, \sigma \in \mathbb{R}_+^{|\mathcal{S}_k|}, \delta \in (0,1)^{|\mathcal{S}_k|}$ are the log-scale mean, log-scale standard deviation and zero-inflation parameters of the zero-inflated log-normal distribution, respectively. Additionally, $\alpha \in \mathbb{R}_+^{|\mathcal{S}_k|}, \beta \in \mathbb{R}^{|\mathcal{S}_k|}$ are scaling and bias factors. The set of learnable parameters in this context is $\theta_k = \{\sigma, \delta, \alpha, \beta\}$ and ϕ_{de_k}.

To further verify our model's triple omics integration ability, we employ a nearest neighbor-based label transfer approach using the ATAC-seq dataset as the reference and test the cell similarity of cells with the same label in different omics by markers. For each cell in the scRNA-seq and snmC-seq datasets, we identify the five closest neighbors in the ATAC-seq dataset within the shared embedding space. We achieve label assignment through a majority voting system based on these neighbors. To confirm the accuracy of our alignment, we test for significant overlap in cell type marker genes across modalities. Initially, we convert features from all omics layers into gene-centric representations with geneactivity function in epiScanpy [73]. Subsequently, we pinpoint cell type markers for each omics layer. We adopt a one-versus-rest Wilcoxon rank-sum test to ascertain these markers, using a significance threshold of FDR $<$ 0.05. The significance of overlapping markers among the three omics datasets is gauged using the three-way Fisher's exact test [53]. Visualization of the significant markers, as well as a comparison of FDR with the scglue method [29], is facilitated by the UpSet plot [95].

Cross-modal generation of single-cell data

Cross-modal generation of single-cell data serves diverse application scenarios. Primarily, our methodology focuses on training our model using a reference multi-omics dataset encompassing both source and target modalities. Subsequently, this trained model facilitates the generation of the missing modality in an untrained dataset. An alternative application stems from cases with limited data for one modality. To demonstrate our method's prowess, we train it on such partial datasets and leverage this for cross-modal generation.

Initially, we employ the unmatched mouse cortex dataset [38] to train our model. Post this preliminary training, we execute fine-tuning using the scATAC-seq component of the matched brain dataset [36]. We then project the scATAC-seq data from this matched dataset to its corresponding scRNA-seq, using the latter's original data as a benchmark. To showcase the capability of our model in handling datasets with limited modalities, we design a simulation using the matched mouse cortex dataset [36]. The dataset is bifurcated: 20% is employed for model training, and the remaining serves for cross-generation and validation. This simulation essentially aims at highlighting how our model, when trained on single-cell multi-omics data with a restricted modality, can upscale and extract valuable insights from the limited data. After the training phase, we employ our model to cross-generate the scRNA-seq data from the available scATAC-seq modality, thereby demonstrating the model's ability to effectively leverage the trained insights from the limited dataset.

The robustness of our cross-modal generation technique is underscored by a comprehensive downstream analysis. The generated data is first clustered using the Leiden algorithm [93]. With the assistance of Scanpy [12], we discern the top 100 differentially expressed genes (DEgenes) for each cell type in both the generated and original scRNA-seq datasets. These DEgenes not only facilitate the transfer of cell type annotations from the original to the generated datasets but also serve as potential biomarker genes. Cell types with the majority of shared genes are aligned. Pathway enrichment further enriches our validation process. For each cell type, pathway associations rooted in the top 100 DEgenes are pinpointed via the ToppGene portal [56]. Concurrently, to decipher the intricate interplay between cells, CellChat [96] is employed for detailed cell-cell interaction analysis. To gauge the accuracy and resemblance between our generated and the original scRNA-seq data, we deploy an array of validation metrics. Pearson correlation [97] coefficients provide insights into decomposition, cell-cell interactions, and pathway analyses.

Single-cell multi-omic data simulation

Our model's training is grounded on the matched mouse cortex dataset [36]. Specifically, for our simulation exercise, we choose the cellular type Ast as the target cellular status. Leveraging our model, we generate data sets at onefold (1X) and fivefold (5X) of the original count of Ast cells' single-cell multi-omics data. These data sets are subsequently visualized using both Scanpy [12] (v.1.8.2) and epiScanpy [73].

With the assistance of Scanpy [12], we discern the top 100 differentially expressed genes (DEgenes) for Ast cells in both the 5X simulated and original scRNA-seq data. GO and pathway enrichment in both data further enriches our validation process. GO and pathway associations rooted in the top 100 DEgenes are pinpointed via the ToppGene portal [56].

To assess the level of correlation between the differential gene signatures of the original Ast cells and the simulated 1X Ast data against a background, we employ the RRHO methodology [98].

To further harness and validate the simulated 1X Ast single-cell multi-omics data, we input it into various integration techniques, notably our scCross and the externally established scglue [29]. The outcomes from these integrative analyses serve to demonstrate the adeptness of the scCross model in generating high-quality matched single-cell multi-omics data for specific cellular states, which can be potentially leveraged to benchmark other single-cell multi-omics data integration methods.

In silico perturbations and explorations

In this section, we aim to showcase the in silico perturbation pipeline of scCross, using the single-cell multi-omics dataset of COVID-19 as an exemplar. This dataset, comprising both scRNA-seq and ADT data (CITE-Seq), serves as a suitable backdrop for our demonstration, given its comprehensive nature that facilitates deeper insights into the disease [20]. The preliminary step in our pipeline involves visualizing the scRNA-seq data of the COVID-19 dataset. We utilized the Scanpy package for this purpose [12] (v.1.8.2), generating a UMAP plot depicted in Fig. 6. Concurrently, the same package aided in the identification of the highly variable genes. With these genes in focus, the scCross model, once trained on matched scRNA-seq and single-cell protein datasets [20], embarked on the in silico perturbation process. In our study, virtual perturbations involved modulating the gene expression levels in the dataset, not by simply setting the expression to zero but through a nuanced approach. For upregulation, we incremented the original counts of selected genes by enhancing the expression by 50% of the original counts, and for downregulation, we reduced the expression by 50% of the original counts. This approach allowed us to simulate the potential effects of gene activation or suppression within the cellular environment, which is to imitate the real perturbing situation of wet lab perturbation studies [99, 100]. Each highly variable gene underwent a virtual perturbation on COVID-19 patient cells, and the model gauged and scored its influence on bridging the latent space gap between contrasting disease states, namely healthy and COVID-19 conditions. The influence of gene perturbations was quantitatively assessed by comparing the perturbed data with control or healthy datasets within the joint latent space produced by scCross. We specifically measured the cosine distance between the perturbed and control matrices. This metric provided a robust method to evaluate how perturbations alter the cell state, effectively pinpointing critical disease-associated markers. The culmination of this process results in a ranked list of genes, with top scorers spotlighted as potential RNA bio-markers for COVID-19.

To further visualize the ramifications of these perturbations, we employed the Seaborn Python package [101] (v.0.11.2). This involved computing the cosine distances in the latent space between COVID-19 and healthy cells during both upward and downward perturbations (scaled by ±1 in log space) of each gene. The resultant visualization, available in Fig. 6b, exclusively displays genes that effectively narrowed the latent space divide.

To further analyze the genes obtained from perturbation, we perform pathway analysis using the ToppGene website [56] with the down-perturbed genes that make the latent space closer. Additionally, we use the *t*-test [102] and Wilcoxon [103] tests, which are general methods for differential gene expression analysis, to generate the same number of DEgenes as our approach between COVID-19 patient cells and healthy cells. The pathway analysis is also performed using the ToppGene website [56]. We select several pathways highly associated with COVID-19 to demonstrate the effectiveness of our approach in identifying disease biomarkers via in silico perturbations (Fig. 6c).

Furthermore, we execute cross-generation of single-cell protein data using the scRNA-seq dataset. For this procedure, we perturb the relevant genes identified in the prior analysis within the healthy scRNA-seq data. This gene perturbation leads to the creation of perturbed COVID-19 scRNA-seq data, which is subsequently harnessed for cross-generation in order to produce a generated dataset representing single-cell COVID-19 protein data.

To evaluate the correlation between the similarity of differential protein expressions in the original COVID-19 protein data and the perturbed, generated COVID-19 protein data against the backdrop of healthy protein data, we employ RRHO [98] as a visualization technique.

In order to validate the degree of resemblance between the up-perturbed and cross-generated single-cell protein data and the original dataset, we leverage the Scanpy package [12] (v.1.8.2). This comparison is performed within a comprehensive landscape encompassing all cells (Fig. 6f). Pearson correlation [97], along with its associated *p*-values, is adopted to quantitatively affirm the similarity between these datasets.

Beyond this validation, our method also allows for quantitative measurement of distance shifts in another modality space, providing an alternative way to assess the influence of perturbed genes. By perturbing highly variable genes in samples from any condition and generating cross-modality data, we can calculate the distance between the perturbed matrices and the original matrices for the unperturbed data. This enables us to score and identify the genes that cause the most significant shifts across modalities. These genes could potentially serve as important markers for intervention candidates.

Batch effect correction

Batch effect intro modalities can be a hard problem to resolve when integrating multimodal datasets. Assuming $b \in 1, 2, \ldots, B$ is the batch index, where B is the total number of batches in all modalities. To better address the problem, we utilize the batch effect correction method in [29], which transforms the traditional parameters α and β to batch-specific parameters α_b and β_b in VAE data decoder:

$$p\left(\mathbf{x}_k \mid \mathbf{z}_k, b; \theta_k, \phi_{de_k}\right) = \prod_{i \in \mathcal{S}_k} \mathrm{NB}\left(\mathbf{x}_{\mathrm{k}}^{(\mathrm{i})}; \mu^{(\mathrm{i})}, \theta_{\mathrm{b}}^{(\mathrm{i})}, \phi_{\mathrm{de_k}}\right) \tag{33}$$

$$\mathrm{NB}\left(\mathbf{x}_{\mathrm{k}}^{(\mathrm{i})};\mu^{(\mathrm{i})},\theta_{\mathrm{b}}^{(\mathrm{i})},\phi_{\mathrm{de}_{\mathrm{k}}}\right)= \frac{\Gamma\left(\mathbf{x}_k^{(i)}+\theta_b^{(i)}\right)}{\Gamma\left(\theta_b^{(i)}\right)\Gamma\left(\mathbf{x}_k^{(i)}+1\right)}\left(\frac{\mu^{(i)}}{\theta_{b^{(i)}}+\mu^{(i)}}\right)^{\mathbf{x}_k^{(i)}}\left(\frac{\theta_b^{(i)}}{\theta_b^{(i)}+\mu^{(i)}}\right)^{\theta_b^{(i)}} \tag{34}$$

$$\mu^{(i)}=\mathrm{Softmax}^{(\mathrm{i})}\left(\alpha_b\odot\mathrm{MLP}_{de_k}\left(\mathbf{z}_k,\phi_{de_k}\right)+\beta_b\right)\cdot\sum_{\nu\in\mathcal{S}_k}\mathbf{x}_k^{(\nu)} \tag{35}$$

where $\theta\in\mathbb{R}_+^{B*|\mathcal{S}_k|}$, $\alpha\in\mathbb{R}_+^{B*|\mathcal{S}_k|}$, $\beta\in\mathbb{R}^{B*|\mathcal{S}_k|}$, and $\theta_b,\alpha_b,\beta_b$ are the b^{th} row of θ,α,β. Other VAE decoders in different distributions can also be extended in similar ways. These formulas represent the batch-specific scaling and bias factors out of MLP networks. The inclusion of parameters α_b and β_b alongside the Multi-Layer Perceptron (MLP) is motivated by the formulas above, wherein each batch in training receives specific α_b and β_b values to adjust the data within that batch. These parameters, α_b and β_b, can vary for each batch b during training, allowing for the correction of batch effects inherent in the original dataset. Given that data within a dataset may originate from different samples and experimental batches, the presence of batch effects can hinder data integration. By incorporating α_b and β_b differences between batches can be corrected within the data matrix during the training process. Acting as training parameters, α_b and β_b automatically scale and bias the data within each batch towards a unified common latent space. Without these parameters, batch effects would persist, potentially disrupting both integration and cross-generation processes.

Implementation details

In our model, we employ linear dimensionality reduction techniques such as PCA (Principal Component Analysis) [72] for scRNA-seq data, gene information, and gene set information. For scATAC-seq data, we use LSI (Latent Semantic Indexing) [74] as the first transformation layers of the data encoders (while the decoders were still fitted in the original feature spaces). These canonical methods effectively reduce the model size and allow for modular input, enabling the use of advanced dimensionality reduction or batch effect correction methods as preprocessing steps for scCross integration.

To balance the losses in our model, we determine the value of lambda (λ) for loss balancing. The lambda selection process can be found in Additional file 1: Fig. S6a–b (please refer to the supplementary materials of the corresponding publication).

Our model consists of two steps of training. In the first step, each omic layer's VAE is trained separately for each modality. This step aims to obtain well-preserved bio-information in the latent embeddings (represented as $\mathbf{z}_k$) without interference from other modalities. In the second step, integration is performed by training all modalities' $\mathbf{z}_k$ together to achieve a unified latent space represented as $\mathbf{z}$. This step enables the integration of different omic layers in the scCross model. The set of the best hyperparameters is presented in Additional file 1: Fig. S18a. The stability of scCross is demonstrated through comprehensive parameter and data corruption studies, as shown in Additional file 1: Fig. S18a–b. These analyses highlight the model's robustness. Our experiments with the hyperparameter λ_p, which balances the VAE and GAN losses, found that a setting of 0.05 consistently outperformed other tested values and is therefore set as the default in our model. Similarly, for the "MNN" hyperparameter λ_{MNN}, which adjusts the balance between VAE and MNN losses, a setting of 0.04 yielded the best results and has been adopted as the standard setting. The model exhibits a maximum performance variance of approximately 0.2 for λ_p and 0.3 for λ_{MNN} across various settings, demonstrating consistent stability under parameter adjustments, as detailed in Additional file 1: Fig. S18a. Furthermore, it maintains robust performance with minimal impact, even with a data corruption rate of 50% (i.e., half of the data was corrupted), as shown in Additional file 1: Fig. S18b. The loss curve of scCross in Additional file 1: Fig. S18c clearly demonstrates satisfactory convergence over the course of the training period. These stability tests were conducted using the matched mouse cortex dataset and are applicable to other datasets. As detailed in Additional file 1: Fig. S6a–b, scCross demonstrates efficient consumption of computational resources in terms of both time and memory. This efficiency is particularly critical as the scale of single-cell data increases. Beyond a dataset size of 10,000 cells, our method either matches or surpasses the performance of all benchmarked tools in terms of computational efficiency. These comparisons were conducted on the Compute Canada platform, utilizing a hardware environment of 1 x NVIDIA P100 Pascal GPU and 2 x Intel E5-2650 v4 Broadwell CPUs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03338-z.

Additional file 1: Supplementary Figures.

Additional file 2: Table S1. Top 5 differentially expressed proteins/genes for ground-truth COVID-19 protein samples, cross-generated COVID-19 protein samples with perturbation, and ground-truth COVID-19 scRNA-seq samples, each compared to healthy controls. Table S2. Detailed information for datasets including Matched mouse cortex, Matched mouse atherosclerotic plaque immune cells, Matched mouse lymphonodus, Unmatched mouse cortex, Human cell atlas and COVID-19.

Additional file 3: Table S3. Detailed benchmarking information for the performance of scCross against other integration methods.

Additional file 4: Review history.

Review history

The review history is available as Additional file 4.

Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

JD and XY conceived the study; JD, XY, and HW designed the analyses and experiments; XY performed the analyses, experiments and visualizations; XY and JD prepared the first draft; JD, XY, HW, and KKM reviewed and edited the draft; KKM contributed the data GSE240753.

Funding

This work is partially supported by grants awarded to JD from the Canadian Institutes of Health Research (CIHR) (Grant No. PJT-180505), the Fonds de recherche du Québec - Santé (FRQS) (Grant Nos. 295298 and 295299), and the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. RGPIN2022-04399), along with support from the Meakins-Christie Chair in Respiratory Research. The study also benefits from partial funding to HW through grants from the National Natural Science Foundation of China (Grant Nos. 62272278 & 61972322), the National Key Research and Development Program (Grant No. 2021YFF0704103), and the Fundamental Research Funds of Shandong University. Additionally, KKM has received partial support for this research from the CIHR (Grant No. PJT-166142).

Availability of data and materials

Our algorithm is designed to be highly adaptable and can effectively handle datasets from different experimental setups without the need for matched single-cell data across these modalities. Specifically, our method supports the integration and cross-modal analysis of various single-cell omics datasets, irrespective of whether they originate from the same cell populations or from split samples where cells are apportioned into different sequencing workflows (such as one portion for scRNA-seq and another for scATAC-seq). The datasets used in this study have been obtained from various sources. The matched mouse cortex dataset [36, 104] is obtained from the Gene Expression Omnibus (GEO) repository under the following accession numbers GSE126074. The matched mouse atherosclerotic plaque immune cells dataset [105, 106] is got from the Gene Expression Omnibus (GEO) repository under the following accession numbers GSE240753. The matched mouse lymphonodus dataset [37, 107] is downloaded from the Gene Expression Omnibus (GEO) repository under the following accession numbers GSE140203. The unmatched mouse cortex scRNA-seq dataset [38, 108] is publicly available on the Dropviz website at http://dropviz.org. The unmatched mouse cortex snmC-seq dataset [39, 109] is publicly available and obtained from the Gene Expression Omnibus (GEO) repository under the following accession numbers GSE97179. The unmatched mouse cortex scATAC-seq dataset is downloaded from the 10X Genomics website at https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k. The human cell atlas dataset is obtained from the Gene Expression Omnibus (GEO) repository under the following accession numbers GSE156793 for scRNA-seq [43, 110] and GSE149683 for scATAC-seq [44, 111]. Lastly, the COVID-19 dataset [20, 112] used in the study is obtained from Sanger's website at https://covid19.cog.sanger.ac.uk/submissions/release1/haniffa21.processed.h5ad. Detailed information for those datasets can be found in Additional file 2: Table S2 and detailed average benchmarking information for our integration evaluation across the methods section is available in Additional file 3: Table S3.

The scCross framework is implemented in the "scCross" Python package, which is available under the MIT license at Github with the link https://github.com/mcgilldinglab/scCross [113] and Zenodo with the link https://doi.org/10.5281/zenodo.12552875 [114]. Examples of each dataset utilized in our study can be found in our Github repository.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 December 2023 Accepted: 16 July 2024
Published online: 29 July 2024

References

1. Xue R, Zhang Q, Cao Q, Kong R, Xiang X, Liu H, et al. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. Nature. 2022;612(7834):41–147.
2. Blanchard JW, Akay LA, Davila-Velderrain J, von Maydell D, Mathys H, Davidson SM, et al. APOE4 impairs myelination via cholesterol dysregulation in oligodendrocytes. Nature. 2022;611(7937):769–79.
3. Niño JLG, Wu H, LaCourse KD, Kempchinsky AG, Baryiames A, Barber B, et al. Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. Nature. 2022;611(7937):810–17.
4. Finkbeiner C, Ortuño-Lizarán I, Sridhar A, Hooper M, Petter S, Reh TA. Single-cell ATAC-seq of fetal human retina and stem-cell-derived retinal organoids shows changing chromatin landscapes during cell fate acquisition. Cell Rep. 2022;38(4):110294.
5. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015;526(7571):131–5.
6. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol. 2018;14(4):e8046.
7. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci. 2015;18(1):145–53.
8. Heath JR, Ribas A, Mischel PS. Single-cell analysis tools for drug discovery and development. Nat Rev Drug Discov. 2016;15(3):204–16.
9. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96.
10. Lin Y, Wu TY, Wan S, Yang JY, Wong WH, Wang Y. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40(5):703–10.
11. Jin S, Zhang L, Nie Q. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol. 2020;21:1–19.
12. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
13. Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat Methods. 2019;16(5):397–400.
14. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18(11):1333–41.
15. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10(11):1096–8.
16. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):1–12.

17. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348(6237):910–4.
18. Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. A rapid and robust method for single cell chromatin accessibility profiling. Nat Commun. 2018;9(1):1–9.
19. Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science. 2017;357(6351):600–4.
20. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med. 2021;27(5):904–16.
21. Wen H, Ding J, Jin W, Wang Y, Xie Y, Tang J. Graph neural networks for multimodal single-cell data integration. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining ACM. New York: Association for Computing Machinery; 2022. p. 4153–63. https://dl.acm.org/doi/abs/10.1145/3534678.3539213.
22. Cao Y, Fu L, Wu J, Peng Q, Nie Q, Zhang J, et al. Integrated analysis of multimodal single-cell data with structural similarity. Nucleic Acids Res. 2022;50(21):e121–e121.
23. Li G, Fu S, Wang S, Zhu C, Duan B, Tang C, et al. A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data. Genome Biol. 2022;23(1):1–23.
24. Lynch AW, Theodoris CV, Long HW, Brown M, Liu XS, Meyer CA. MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. Nat Methods. 2022;19(9):1097–108.
25. Xiong L, Tian K, Li Y, Ning W, Gao X, Zhang QC. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. Nat Commun. 2022;13(1):6118.
26. Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. Nat Commun. 2022;13(1):7705.
27. Zhao J, Wang G, Ming J, Lin Z, Wang Y, Wu AR, et al. Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets. Nat Comput Sci. 2022;2(5):317–30.
28. Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87.
29. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022;40(10):1458–66.
30. Cao K, Gong Q, Hong Y, Wan L. A unified computational framework for single-cell data integration with optimal transport. Nat Commun. 2022;13(1):7419.
31. Xu Y, Begoli E, McCord RP. sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network. NPJ Syst Biol Appl. 2022;8(1):33.
32. Zhang Z, Yang C, Zhang X. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. Genome Biol. 2022;23(1):139.
33. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. 2019;20(1):1–25.
34. Steinley D. Properties of the hubert-arable adjusted rand index. Psychol Methods. 2004;9(3):386.
35. Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3(Dec):583–617.
36. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019;37(12):1452–7.
37. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell. 2020;183(4):1103–16.
38. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, et al. Molecular diversity and specializations among the cells of the adult mouse brain. Cell. 2018;174(4):1015–30.
39. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol. 2018;36(5):428–31.
40. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41–50.
41. Singh R, Demetci P, Bonora G, Ramani V, Lee C, Fang H, et al. Unsupervised manifold alignment for single-cell multi-omics data. In: Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics ACM. New York: Association for Computing Machinery; 2020. p. 1–10. https://dl.acm.org/doi/abs/10.1145/3388440.3412410.
42. Cui Z, Liao Y, Xu T, Wang Y. GeneFormer: Learned Gene Compression using Transformer-based Context Modeling. 2022. arXiv preprint arXiv:221208379.
43. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. Science. 2020;370(6518):eaba7721.
44. Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, et al. A human cell atlas of fetal chromatin accessibility. Science. 2020;370(6518):eaba7612.
45. Jouanneau J, Moens G, Bourgeois Y, Poupon M, Thiery J. A minority of carcinoma cells producing acidic fibroblast growth factor induces a community effect for tumor progression. Proc Natl Acad Sci. 1994;91(1):286–90.
46. Li M, Zhang T, Chen Y, Smola AJ. Efficient mini-batch training for stochastic optimization. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining ACM. New York: Association for Computing Machinery; 2014. p. 661–70. https://dl.acm.org/doi/abs/10.1145/2623330.2623612.
47. King A, Burrows T, Hiby S, Bowen J, Joseph S, Verma S, et al. Surface expression of HLA-C antigen by human extravillous trophoblast. Placenta. 2000;21(4):376–87.
48. Tantbirojn P, Crum C, Parast M. Pathophysiology of placenta creta: the role of decidua and extravillous trophoblast. Placenta. 2008;29(7):639–45.
49. Champion H, Innes BA, Robson SC, Lash GE, Bulmer JN. Effects of interleukin-6 on extravillous trophoblast invasion in early human pregnancy. Mol Hum Reprod. 2012;18(8):391–400.
50. Kettenmann H, Hanisch UK, Noda M, Verkhratsky A. Physiology of microglia. Physiol Rev. 2011;91(2):461–553.
51. Loane DJ, Byrnes KR. Role of microglia in neurotrauma. Neurotherapeutics. 2010;7(4):366–77.

52. Perry VH, Nicoll JA, Holmes C. Microglia in neurodegenerative disease. Nat Rev Neurol. 2010;6(4):193–201.
53. Wang M, Zhao Y, Zhang B. Efficient test and visualization of multi-set intersections. Sci Rep. 2015;5(1):16923.
54. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. 2018. arXiv preprint arXiv:180203426.
55. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet. 2021;22(2):71–88.
56. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37(suppl_2):W305–11.
57. Koehler RC, Roman RJ, Harder DR. Astrocytes and the regulation of cerebral blood flow. Trends Neurosci. 2009;32(3):160–9.
58. Pardridge WM. Drug transport across the blood-brain barrier. J Cereb Blood Flow Metab. 2012;32(11):1959–72.
59. Sweeney MD, Ayyadurai S, Zlokovic BV. Pericytes of the neurovascular unit: key functions and signaling pathways. Nat Neurosci. 2016;19(6):771–83.
60. Xu ZS, Shu T, Kang L, Wu D, Zhou X, Liao BW, et al. Temporal profiling of plasma cytokines, chemokines and growth factors from mild, severe and fatal COVID-19 patients. Signal Transduct Target Ther. 2020;5(1):100.
61. Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, et al. COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. Nat Biotechnol. 2020;38(8):970–9.
62. Lee JS, Park S, Jeong HW, Ahn JY, Choi SJ, Lee H, et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. Sci Immunol. 2020;5(49):eabd1554.
63. Vinken M. COVID-19 and the liver: an adverse outcome pathway perspective. Toxicology. 2021;455:152765.
64. Rex D, Dagamajalu S, Kandasamy RK, Raju R, Prasad TK. SARS-CoV-2 signaling pathway map: A functional landscape of molecular mechanisms in COVID-19. J Cell Commun Signal. 2021;15(4):601–8.
65. SARS-CoV-2 innate immunity evasion and cell-specific immune response | WikiPathways. https://www.wikipathways.org/pathways/WP5039.html.
66. Deng X, Terunuma H, Nieda M. Exploring the Utility of NK Cells in COVID-19. Biomedicines. 2022;10(5):1002.
67. Leem G, Cheon S, Lee H, Choi SJ, Jeong S, Kim ES, et al. Abnormality in the NK-cell population is prolonged in severe COVID-19 patients. J Allergy Clin Immunol. 2021;148(4):996–1006.
68. Horenstein AL, Faini AC, Malavasi F. CD38 in the age of COVID-19: a medical perspective. Physiol Rev. 2021;101(4):1457–86.
69. Zeidler JD, Kashyap S, Hogan KA, Chini EN. Implications of the NADase CD38 in COVID pathophysiology. Physiol Rev. 2022;102(1):339–41.
70. Wang Y. Activating organ's immunizing power against COVID–19–Learning from SARS. Research & reviews: J Biol. 2024;12(1):68–78.
71. Kattner S, Müller J, Glanz K, Manoochehri M, Sylvester C, Vainshtein Y, et al. Identification of two early blood biomarkers ACHE and CLEC12A for improved risk stratification of critically ill COVID-19 patients. Sci Rep. 2023;13(1):4388.
72. Roweis ST. EM algorithms for PCA and SPCA. In: Jordan MI, Kearns MJ, Solla SA, editors. Advances in neural information processing systems 10. Cambridge: MIT Press; 1998. p. 626–32.
73. Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colomé-Tatché M. EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun. 2021;12(1):5228.
74. Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval ACM. New York: Association for Computing Machinery; 1999. p. 50–7. https://dl.acm.org/doi/pdf/10.1145/312624.312649.
75. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
76. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. Reconstructing dynamic regulatory maps. Mol Syst Biol. 2007;3(1):74.
77. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2007;36(suppl_1):D480–4.
78. Kerr M. The human complement system: assembly of the classical pathway C3 convertase. Biochem J. 1980;189(1):173–81.
79. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55.
80. Ding J, Aronow BJ, Kaminski N, Kitzmiller J, Whitsett JA, Bar-Joseph Z. Reconstructing differentiation networks and their regulation from time series single-cell expression data. Genome Res. 2018;28(3):383–95.
81. Dincer AB, Janizek JD, Lee SI. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. Bioinformatics. 2020;36(Supplement_2):i573–82.
82. Cao ZJ, Wei L, Lu S, Yang DC, Gao G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. Nat Commun. 2020;11(1):1–13.
83. Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7.
84. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023;24(8):550–72.
85. Fouché A, Zinovyev A. Omics data integration in computational biology viewed through the prism of machine learning paradigms. Front Bioinforma. 2023;3:1191961.
86. Cao K, Bai X, Hong Y, Wan L. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics. 2020;36(Supplement_1):i48–56.
87. Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics. 2022;38(1):211–9.
88. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multiomic data with totalVI. Nat Methods. 2021;18(3):272–82.

89. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:1–17.
90. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protocol. 2020;15(11):3632–62.
91. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
92. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet. 2021;53(3):403–11.
93. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):1–12.
94. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
95. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33(18):2938–40.
96. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using Cell Chat. Nat Commun. 2021;12(1):1–20.
97. Cohen I, Huang Y, Chen J, Benesty J, Benesty J, Chen J, et al. Pearson correlation coefficient. Noise reduction speech process. 2009. p. 1–4. https://www.nature.com/articles/s41467-023-40155-7. https://link.springer.com/article/10.1186/s12920-023-01543-6.
98. Plaisier SB, Taschereau R, Wong JA, Graeber TG. Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. Nucleic Acids Res. 2010;38(17):e169.
99. Susman MW, Karuna EP, Kunz RC, Gujral TS, Cantu AV, Choi SS, et al. Kinesin superfamily protein Kif26b links Wnt5a-Ror signaling to the control of cell and tissue behaviors in vertebrates. Elife. 2017;6:e26509.
100. Myers KS, Riley NM, MacGilvray ME, Sato TK, McGee M, Heilberger J, et al. Rewired cellular signaling coordinates sugar and hypoxic responses for anaerobic xylose fermentation in yeast. PLoS Genet. 2019;15(3):e1008037.
101. Waskom ML. Seaborn: statistical data visualization. J Open Source Softw. 2021;6(60):3021.
102. Kim TK. T test as a parametric statistic. Korean J Anesthesiol. 2015;68(6):540–6.
103. Cuzick J. A Wilcoxon-type test for trend. Stat Med. 1985;4(1):87–90.
104. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. GEO. 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126074.
105. Makhani K, Yang X, Dierick F, Subramaniam N, Gagnon N, Ebrahimian T, et al. Unveiling the impact of arsenic toxicity on immune cells in atherosclerotic plaques: insights from single-cell multi-omics profiling. bioRxiv. 2023:2023.11.23.568429. [cited 2024 July 21]. Available from: https://www.biorxiv.org/content/10.1101/2023.11.23.568429.
106. Makhani K, Yang X, Dierick F, Subramaniam N, Gagnon N, Ebrahimian T, et al. Unveiling the impact of arsenic toxicity on immune cells in atherosclerotic plaques: insights from single-cell multi-omics profiling. GEO. 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE240753.
107. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell GEO. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140203.
108. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, et al. Molecular diversity and specializations among the cells of the adult mouse brain. Dropviz. 2018. Accessed 3 Nov 2021. http://dropviz.org.
109. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. GEO. 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97179.
110. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, et al. A human cell atlas of fetal gene expression. GEO. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156793.
111. Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, et al. A human cell atlas of fetal chromatin accessibility. GEO. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149683.
112. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. Sanger. 2021. https://covid19.cog.sanger.ac.uk/submissions/release1/hania21.processed.h5ad.
113. Xiuhui Yang KKM, Wu H, Ding J. scCross: a deep generative model for unifying single-cell multi-omics with seamless integration, cross-modal generation, and in-silico exploration. Github. 2024. https://github.com/mcgillding lab/scCross. Accessed 21 July 2024.
114. Xiuhui Yang KKM, Wu H, Ding J. scCross: a deep generative model for unifying single-cell multi-omics with seamless integration, cross-modal generation, and in silico exploration. Zenodo. 2024. https://doi.org/10.5281/zenodo.12552875.

Publisher's Note