Genome Medicine

REVIEW

Open Access

Machine learning integrative approaches to advance computational immunology

Fabiola Curion[1,2] and Fabian J. Theis[1,2,3*]

Abstract

The study of immunology, traditionally reliant on proteomics to evaluate individual immune cells, has been revolutionized by single-cell RNA sequencing. Computational immunologists play a crucial role in analysing these datasets, moving beyond traditional protein marker identification to encompass a more detailed view of cellular phenotypes and their functional roles. Recent technological advancements allow the simultaneous measurements of multiple cellular components—transcriptome, proteome, chromatin, epigenetic modifications and metabolites—within single cells, including in spatial contexts within tissues. This has led to the generation of complex multiscale datasets that can include multimodal measurements from the same cells or a mix of paired and unpaired modalities. Modern machine learning (ML) techniques allow for the integration of multiple "omics" data without the need for extensive independent modelling of each modality. This review focuses on recent advancements in ML integrative approaches applied to immunological studies. We highlight the importance of these methods in creating a unified representation of multiscale data collections, particularly for single-cell and spatial profiling technologies. Finally, we discuss the challenges of these holistic approaches and how they will be instrumental in the development of a common coordinate framework for multiscale studies, thereby accelerating research and enabling discoveries in the computational immunology field.

Background

The immune system is a complex network of cells that co-orchestrate a defence response to unrecognized perturbations, that potentially endanger the host's life. Throughout its evolution, this precise and lethal supertissue has developed into a dynamic system capable of monitoring the environment in search of potential threats, via a scattered and highly specialized network of sentinel cells that constantly senses and reacts to infections and harmful changes in the organisms' homeostatic equilibrium. The rigorous level of control that supervises its specialized activation is even more evident when, under a perceived threat, the immune system overreacts to pathogens or directs its attack against the host itself, eventually damaging healthy tissues and often degenerating into chronic diseases [1]. As immune cells are by design not bound to specific locations, they have adapted to operate in various microenvironments and under differing conditions of health and disease, making for an intriguing and constantly changing area of research. Studies have uncovered how the immune system plays an important role in complex diseases like neurological disorders [2, 3], diabetes [4] and cancer [5–7]. Given the involvement of the immune response in almost every aspect of an organism's life, it is perhaps not surprising that there is considerable interest in leveraging the immune system to devise patient-specific immunotherapies [8–10].

*Correspondence:
Fabian J. Theis
fabian.theis@helmholtz-munich.de
[1] Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany
[2] Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[3] School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany

The study of the immune system has focused on the evaluation of individual cells to identify characteristics that define their function. For a large part of the last two centuries [11], immunologists have built compendia of the cells that co-orchestrate immune responses by describing morphology, shape, tissue- and disease-specific occurrence and, more recently, phenotypic and molecular markers that delineate their lineage-membership and their evolution in time. Single-cell measurements are not new to immunology. Until recently, post-translational technologies measuring proteins like Fluorescence-Activated Cell Sorting (FACS) [12] and mass cytometry to single cells (CyTOF) [13] have been the preferred technologies to carry out individual cell characterization, thanks to selected surface or intracellular markers, whose expression levels are measured in thousands of single cells. However, with the advent of single-cell sequencing techniques, the field of immunology has advanced dramatically [11, 14–18].

The single-cell genomics revolution has impacted every field of biology and medicine [19]; in particular, the widespread adoption of barcoded droplet-based approaches [20] allowed a massive increase in the number of studies leveraging single-cell genomics. Since the launch of the international Human Cell Atlas (HCA) initiative [21], the global scientific community has leveraged single-cell genomics approaches to generate multiple atlases across tissues [22, 23], developmental stages [24] and diseases, with efforts such as the lung cell atlas [25] pioneering the field of integrative analysis of large collections of single cells. Similarly, recent studies have privileged the use of single-cell genomics to obtain a deep characterization of the immune system under healthy and disease conditions [19, 26, 27], such as infectious diseases [19, 26–28] a trend also boosted by the global COVID pandemic [29–31].

Recent technological advances in single-cell genomics allow measurements of multiple molecular read-outs: transcriptome, surface and intracellular proteome, chromatin, epigenetic modifications, immune repertoire and metabolites [32–34]. Both unspliced and spliced RNA transcripts are detected with standard scRNA-seq protocols [35], and furthermore, new protocols allowing lineage tracing, perturbation screenings with CRISPR-based transcriptional interference and sequencing of protein complexes [36] are offering new insights into dynamical properties of cells. Lastly, a series of spatial proteogenomics technologies have been developed, which combine microscopic imaging with gene expression [37], open chromatin [38] and proteome [39] while preserving the spatial location information.

In the past, scientists were used to analysing one modality at a time. Modelling the RNA expression provided a powerful means of identifying loci and genes contributing to disease [40]. Immune cell profiling was carried out with the help of a handful of protein markers in cell suspensions [41–43]. Likewise, the diagnosis of immune-mediated neurological disorders using structural imaging like magnetic resonance imaging (MRI) is non-trivial [44]. A typical immunology dataset could consist of several single-cell assays across multiple cellular readouts, bulk measurements, imaging, genetics and clinical data [45, 46]. As most phenotypes result from interactions where different biological layers are at play [47], multi-omics integrative studies provide comprehensive information on all these layers and unveil the hidden architecture behind a complex disease phenotype.

With the increasing size and complexity of these datasets, the future calls for approaches to generate comprehensive multimodal references onto which new datasets can be queried, to ensure fast knowledge transfer [48]. Integrating such multiscale datasets represents a new frontier for biomedical research. Broadly, the goal of machine learning (ML) integrative approaches is to generate a single representation of the various data sources, which can reduce the dimensions and preserve essential information from the input modalities such that the fused representation is more informative than the individual modalities [49, 50]. These embeddings form the foundation of the decision-making process: cell state identification, trajectory inference, molecular pathways and biomarker discovery and patient classification across complex collections of phenotypes.

The data types of multi-scale datasets are various: generally, single-cell technologies consist of sparse matrices with rows and columns indicating cells and features (genes, proteins, chromatin regions), while spatial profiling techniques [51] provide subcellular or cell-aggregates molecular profiling, as well as accompanying images of tissues. Sample-level (bulk) omics measurements, genetic and clinical data do not have cellular resolution but suffer from missing values to different extents [52–56]. These techniques differ greatly in their resolution and consequently, data formats (Fig. 1). Others have covered the topic of integration [57, 58], classifying integration methods based on the relationship and type anchors across the modalities, and introducing terminology for the type of integration such as vertical, horizontal, diagonal and mosaic [57]. Datasets which include both paired and unpaired measurements from different omics require mosaic integration approaches, a type of integration complicated by the limited number or the total lack of shared features between the omics to align. In this review, we will briefly review some of the methods that have found applications in immunological studies, or hold potential for application in such contexts,

highlighting key concepts for the integration (Additional file 1: Table S1). Given the increase availability of multi-scale, unpaired datasets, we will highlight approaches to integrate these data. Finally, we discuss challenges and future development of integrative machine learning approaches, intending to inform those who intend to build up computational expertise to enable multimodal data interpretation.

Integration of cell-based assays

Machine learning (ML) and multimodal integration are rapidly transforming immunological research by leveraging complex datasets from diverse sources [47, 57, 58]. Cross-technology integration of multimodal single-cell assays, such as CITEseq [59], with cytometry assays (FACS, CyTOF) can enable researchers to compound the domain knowledge accumulated with traditional proteomic techniques, to generate information-rich and interpretable references for immune studies.

Some methods originally developed for unimodal integration [68] are useful for integrating this type of multimodal data [47, 69, 70]. This first class of methods relies mostly on traditional linear models and has found wide applicability thanks to the intuitive interpretation of the linear manipulation of the data. On the other hand, a growing body of methods relies on Deep Learning (DL) techniques [71, 72], responding to the increasing complexity of multimodal data. Finally, given our focus on immunological applications, we include in this section methods developed for the integration of adaptive immune receptors (AIR) sequencing data with gene expression. Designed to work on paired data, these methods provide a much finer understanding of the adaptive immune system compared to any unimodal approach.

Linear models

Flavours of linear decompositions have been successful at mosaic integration by leveraging shared features across data modalities [57]. For example, *LIGER* performs integrative non-negative matrix factorization (iNMF) [73] on the shared features, to distinguish between omic-specific factors and shared factors, followed by the construction of a neighbourhood graph using only the shared factors. By including an unshared metagene matrix [74] to inform the factorization, the authors were able to improve the integration of unmatched data across several platforms. *UINMF* [74] extends the LIGER model by accounting for unshared features between the modalities.

CCA is a popular dimensionality reduction [75], identifying canonical covariate vectors that capture sources of variance that are shared between omics that do not necessarily share features [76]. The CCA values can be used to identify cells, or "anchors", with mutually similar profiles between the modalities, and correcting any systematic differences in expression levels between cells ensures their alignment. In [77], the authors leverage CCA to identify a rare subpopulation of CD11c-positive B cells, increasing upon COVID-19 infection, by integrating CyTOF and scRNAseq.

The same dataset is also used in *Bridge integration* [78]. With this approach, the authors characterized a very rare population of innate lymphoid cells, which were not identified in the CyTOF dataset, but correctly exhibited a CD25 + CD127 + CD161 + CD56 − immunophenotype. In this method, a multi-omic dictionary dataset is used as a bridge to translate between two experiments (the reference and the query) that have unpaired cells and features, but each share features with one of the individual assays of the bridge dataset. Horizontal integration of matching assays followed by a matrix factorization step allows generating a new set of shared features. Finally, these matrices undergo dimensionality reduction via Laplacian eigendecomposition and can be horizontally integrated.

CyCombine [79] integrates spectral flow cytometry, mass cytometry and CITESeq. After modality-specific preprocessing, which includes normalization or z-scaling of the expression of every marker in every batch, CyCombine clusters the cells into self-organizing maps (SOM) and applies a per-cluster batch correction method [80] to align the data and minimize technical noise. The authors were able to identify a relevant set of T and NKT cells increased in chronic lymphocytic leukaemia patients,

(See figure on next page.)

Fig. 1 **A** Multimodal immunological datasets can comprise multiple assays across different modalities and resolutions. The number of features measured in each assay ranges from tens to hundreds of thousands. Some of these assays (CITEseq [59]; Multiome [60]) collect joint information from the same cells or samples (modalities aligned vertically). Different assays may share subsets of features (CyTOF [13], FACS [12]) (modalities aligned horizontally). Sample level measurements do not have a cellular resolution (ATAC-seq [61], RNA-seq [62], mass spectrometry [63]; BCR and TCR sequencing [64]) but can be performed in parallel to single-cell assays; spatially resolved cells can be extracted from multiple platforms, sequencing or imaging-based (Spatial ATAC [38], Spatial CITESeq [65], Visium [66], MERFISH [67]). **B** Spatial profiling carries information about RNA expression at individual spatial barcodes (BC). Single-cell references can be leveraged to deconvolute spatial data inferring cell type proportions and gene expression at spatial locations. Histological sections are often an accompanying assay. They can be segmented to recover cell and subcellular structures, as well as general tissue properties such as the morphology of cells, the density of cells at specific locations and cell-to-cell interactions

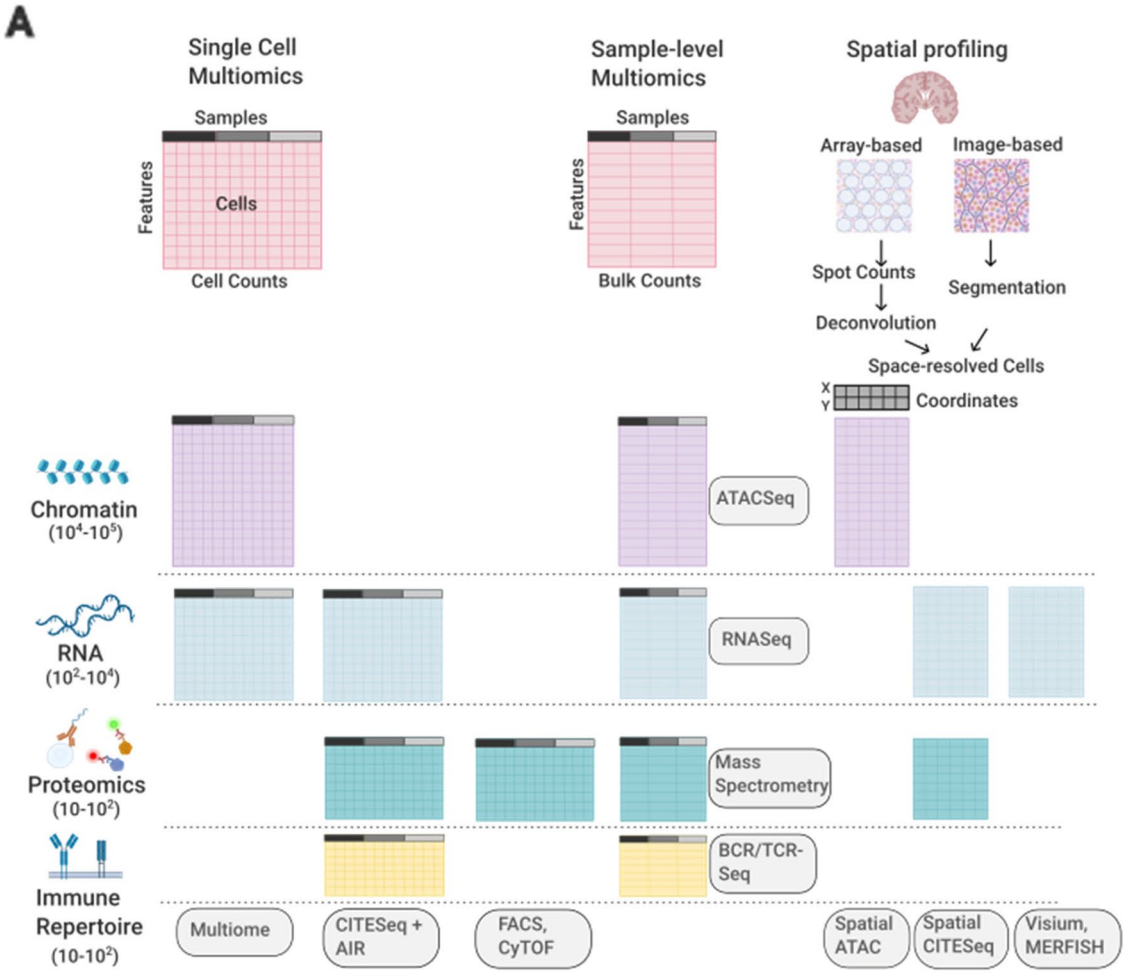

Fig. 1 (See legend on previous page.)

and used the newly generated multimodal embedding to identify a PD-1-positive subset of CD8 + and CD4 + effector memory T cells, classically associated with CLL progression. Nearest rank neighbours (*NRN*) is a method to project flow cytometry data on an ABseq (Surface proteins + Transcriptome) reference [81]. After scaling protein expression within each assay, Euclidean distances between the FACS cells (query) and the reference data are calculated, and then, the query cells are projected onto the reference by k-nearest neighbours search following the scMap [82] batch correction method. Mapping the functional data obtained by flow cytometry onto the genomics space, the authors inferred differentiation dynamics of the haematopoietic lineage differentiation, recovering an early primary erythromyeloid versus lymphomyeloid split.

MARIO [83] matches cells across modalities by performing pairwise matching based on shared features and then projecting the distinct features using CCA. The matching is refined using a convex combination of initial and refined matchings in the CCA space. This method can perform integration across several proteomics assays (Cytof, CODEX [84]) and RNA. The authors were able to generate a cross-species blood atlas under challenged with the influenza virus and Interferon (IFNγ) and to correctly identify key populations of macrophages that sustain the recruitment of immature neutrophils in a COVID-19 lung dataset. Other relevant methods include optimal-transport-based *SCOT* [85], *PAMONA* [86] and *Stabmap* [87], based instead on mosaic data topology (MDT) network, with nodes corresponding to each assay, and edges weighted by the number of shared features between them. *MATCHER* [88] assumes one underlying biological process generating a linear manifold of each modality, resulting in one distribution for each modality. It then aligns them by projecting these curves onto a reference line. Finally, *CoNGA* [89] was developed for the integration of the T cell repertoire of TCRs with gene expression. CoNGA aims to find a joint representation of TCR and RNA from the same cells by identifying the overlap between the similarity graphs constructed on each modality independently.

Deep learning approaches

One of the most widely used DL architectures for single-cell data is Autoencoders (AEs), neural networks that reduce dimensionality and/or noise from different types of data by combining an encoder and a decoder network [90, 91]. The encoder takes a raw data point from the input and maps it to a latent space of underlying factors, while the decoder controls the quality of the dimensionality reduction by reconstructing the original data from the latent representation. Variational Autoencoders (VAE) introduce Variational Inference to account for the irregularity of the latent space, returning for each encoded modality a distribution as opposed to a single point. Finally, the encoder and decoder can be distinct neural network architectures, such as Graph Neural Networks (GNN) [92], a type of neural network that can learn a latent representation from graph-structured data, or Generative Adversarial Networks (GAN) [93]. To align multimodal data, GANs train two neural networks, the generator (G) and the discriminator (D). The GAN model optimizes the integration by letting G and D compete against one another, respectively generating pseudo-data that resembles the real input, and discriminating between pseudo-data and real data (Fig. 2A).

MAGAN [94] aligns mass-spectrometry and scRNA-seq data, recovering shared immune cell populations in the presence of batch effects and enabling the imputation of proteins across the two technologies. MAGAN uses GANs to align data from different domains. Cells are mapped from one modality to the other minimizing a correspondence loss that measures the difference between points before and after the mapping. Autoencoders (AE) are a powerful architecture for multimodal data integration because multiple encoders can be used to learn efficient representations of the input data [95]. *TotalVI*, one of the first VAE frameworks proposed to integrate paired RNA and proteins from CITESeq assays, can be regarded as the reference method for more recent AE-based architectures tackling mosaic integration. TotalVI can denoise protein data and enable the classification of immune cells even in the presence of strong background antibody staining [96]. Some of the more complex VAE-based methods include *SCGLUE* [97], *Multigrate* [98] and *SCIM* [99]. In SCGLUE, the individual modalities encoders and a guidance graph that recapitulates features' relationships across omics are encoded into related factors, which are then concatenated to form a unified embedding, using an adversarial alignment discriminator. scGLUE was demonstrated by integrating chromatin and RNA assays, recovering cis-regulatory events specific to monocytes and B cells [100]. Multigrate combines the single modalities distribution using the Product of Experts (PoE) to obtain the unified latent distribution, allowing the flexibility to deal with both paired or unpaired data. Multigrate allows atlas-level integration of multimodal PBMC datasets including COVID patients, providing a framework for atlas building, patient classification, feature extraction and biomarker discovery. In SCIM, the authors demonstrate the integration of scRNA and CyTOF cells from a melanoma dataset [101], aligning the individual modalities' low-dimensional representations by simultaneously

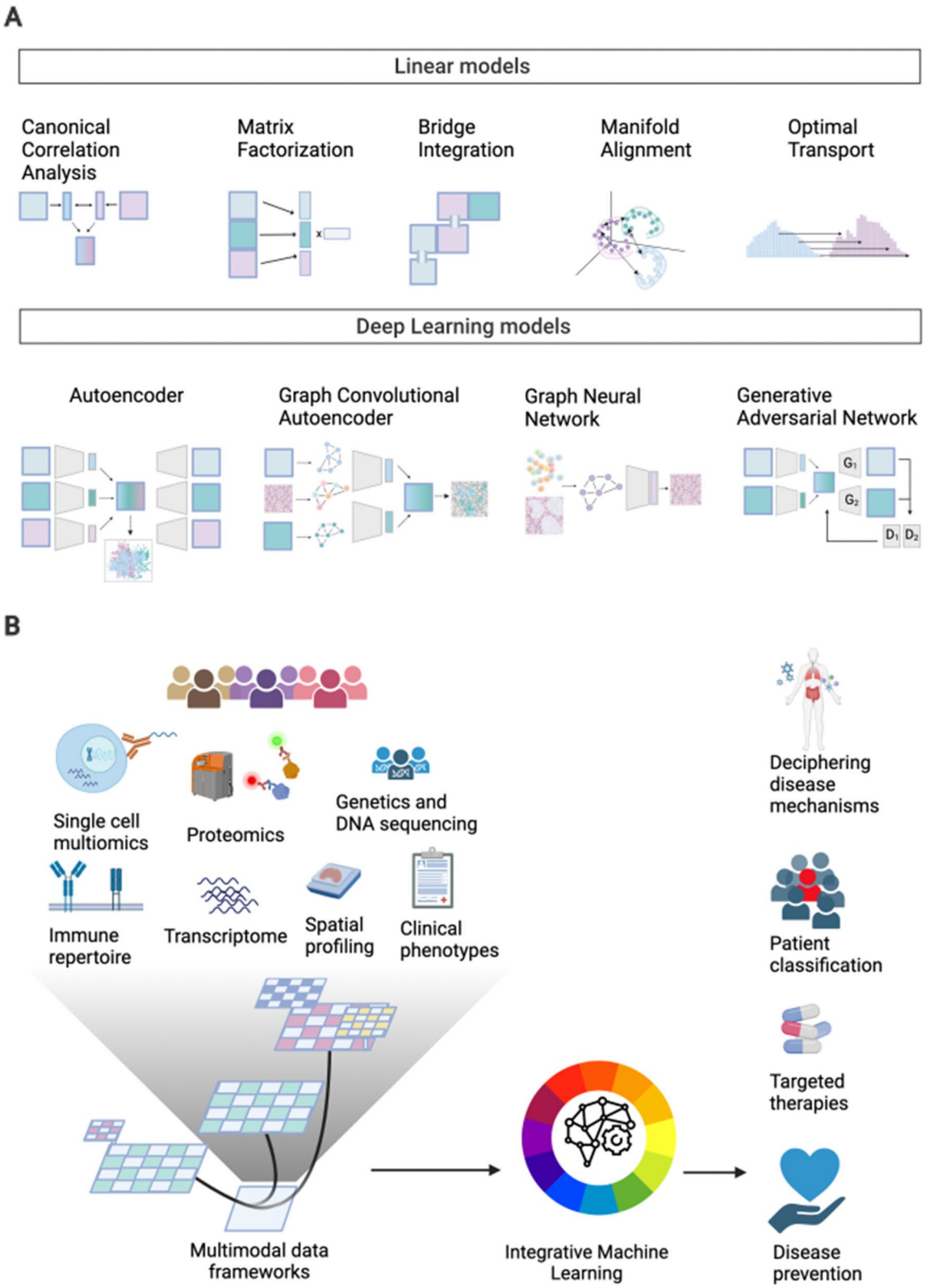

Fig. 2 **A** Schematic diagrams of representative integration approaches for multimodal data. **B** Multimodal datasets have different data formats and can be stored in dedicated data structures that allow to efficiently access and process each layer independently or jointly. These data infrastructures sit at the core of machine learning integrative methods, which in turn provide diverse biomedical insights

training the autoencoders and a discriminator network. The individual modalities' latent spaces are aligned using an adversarial loss, resulting in a joint cellular latent representation. *Cross-modal autoencoders* [102] adopt unimodal AE to integrate and translate between data modalities, with a discriminative objective function to combine the data distributions from the different modalities in the latent space. If prior knowledge of matching features or cell types between domains is available, an additional loss term in the objective function will encourage the alignment between specific markers or the anchoring of certain cells. Integrating single-cell RNA-seq and chromatin images, the authors identified distinct subpopulations of naive CD4 + T-cells that are poised for activation. *MIDAS* [103], a recent deep generative framework for the integration of mosaic data, was demonstrated to disentangle two rare unconventional T-cell populations [104, 105] and successfully predict the differentiation trajectory of myeloid cells in the bone marrow [106] from the inferred embedding. scBridge [107] and scJoint [108] leverage scRNA labels to annotate and integrate multimodal datasets, both developed to integrate ATAC and RNA. scBridge iteratively computes and aligns prototypes of unlabeled scATAC and labelled scRNA cells, gradually aligning batches of cells in the latent space for the two modalities. scBridge uses a confidence score to highlight novel cell types and allowed to distinguish a subset of naive CD4 + T cells that were not detected with scJoint in a T-cell stimulation multimodal dataset [109]. scJoint treats the integration as a domain adaptation problem, constraining the labelled and unlabeled cells to align in the latent space by minimizing a cosine similarity loss, and further optimizing the integration with joint training after label transfer by kNN. The scJoint embedding allowed the authors to distinguish a small population of natural killer T cells in the same multimodal dataset [109].

Finally, three methods use DL frameworks for the integration of AIR with gene expression. *Tessa* [110], a parametric Bayesian hierarchical model, takes as inputs the single-cell RNA and the embeddings of the TCR obtained by feeding a numerical representation of the TCR amino acid sequences to an autoencoder. This latent representation summarizes clusters of related clonotypes which are then correlated to the gene expression profile clusters. Similarly, *Benisse* [111] learns a sparse weighted graph from the embedding of the high-dimensional data of BCRs, under the supervision of gene expression, so that BCRs closer to each other in the latent space have similar BCR sequences and represent B cells with similar transcriptomic features. Finally, *mvTCR* [112] is VAE generating a joint representation from TCR and RNA data. This tool is currently the only one with single-cell resolution on the TCR clonotypes variation, taking as input the RNA and the amino acid sequence of the alpha and beta chains of the TCR from individual cells.

Methods for the integration of cells with spatial profiling data

Single-cell suspension sequencing methods have been largely successful within immunological research because most of the immune cells are not anchored in tissues and are therefore relatively easy to isolate. However, knowing the exact localization of immune cell compartments within solid tissues is important to elucidate cellular cross-talks in several diseases [113]. In cancer research, spatial molecular phenotyping of tumour micro-environment allows uncovering of mechanisms of immune escape [14, 114–116]; in autoimmune diseases affecting the central nervous system such as multiple sclerosis, failure of trophic and anti-inflammatory cellular communication was identified as clear features of the early stages of neurodegeneration [117].

Spatial profiling technologies capture biological patterns emerging in their context [51, 118, 119]. The landscape of spatial profiling methodologies is rapidly expanding [51] with applications in a variety of biological problems, including functional analysis of organs in multiple species, developmental processes and diseases [120, 121].

A distinction can be made across spatial techniques: (i) sequencing-based techniques for whole genomes, which do not have single-cell resolution [122, 123]; (ii) in situ sequencing techniques, which have increased resolution but smaller feature panels [124, 125]; (iii) in situ hybridization-based methods, which image smaller panels of RNAs or proteins at subcellular resolution, without sequencing readouts [38, 39, 66, 84, 126–132].

For sequencing-based methods, integration with single-cell data is needed to achieve at least two goals: prediction of the spatial distribution of features' expression, especially when the spatial profiling technique has low coverage, and deconvolution of multi-cellular spots into cell types [133]. Lastly, methods for the integration of scRNA or Protein datasets with imaging-based techniques are also gaining popularity: most require segmentation of images into cell-readouts, and extraction of informative morphological features linked to individual pixels, while other segmentation-free methods discover spatial domains by reconstructing spatial organization of features in local neighbourhoods (Fig. 1B).

Predicting spatial distribution of feature expression

Some of the methods employed for single-cell integration described in the previous section have been successfully

adapted to integrate spatial transcriptomics with single-cell data [38, 73, 74, 76, 83]. However as spatial profiling techniques gain in popularity alongside single-cell protocols, ad hoc methods have started to emerge, and best-performing methods that integrate spatial data with single cells are often based on DL architectures [134].

Building on the MAGAN architecture, *scMMGAN* [135] can integrate ST and scRNAseq and can recover spatial expression of genes associated with breast cancer progression [136]. *Tangram* aims to maximize the spatial alignment of scRNAseq onto a spatial reference using non-convex optimization in order to retrieve a probabilistic distribution of the feature expression in the spatial data [125]. Interestingly, despite the overall good performance in the experiments shown, Tangram struggled to spatially map immune cells when dealing with different cell-type compositions between the cell suspension and the ST assay, or divergent expression profiles in a cross-species integration.

gimVI [137] is based on the scvi framework [90] and uses a generative model to infer the spatial distribution of undetected transcripts. *stPlus* [138] learns a joint embedding of the spatial transcriptomic data and reference scRNA-seq data via an autoencoder, to then predict the expression of spatial features based on the cell embedding via a weighted k-nearest-neighbor (kNN) method. *SpaGE* [139] leverages the domain adaptation algorithm PRECISE [140] to align the sc and spatial datasets, computing linear latent factors on each dataset and finding gene combinations expressed in both datasets to obtain a joint representation of the data. On this embedding, a kNN algorithm is used to predict the expression of spatially unmeasured genes. OT [141] based methods, such as *novoSpaRc* [142], *SpaOTsc* [143] and *Moscot* [144], assume that single-cell suspensions can be mapped to a tissue space based on similarities between expression profiles. A probabilistic mapping that assigns each cell a distribution over locations on the physical space is computed, allowing to reconstruct spatial gene expression. Additionally, SpaOTsc uses the new coordinates to infer a cell-to-cell communication network based on patterns of ligand and receptor expression, and intercellular regulatory relationships between genes are reconstructed for each pair of genes at a given spatial distance.

Cell type deconvolution of spatial data

To improve the resolution of genomic-scale spatial technologies, methods were developed to estimate the abundance of given cell types at individual spots in histological sections. Many require an annotated scRNA-seq dataset with known cell-type markers to deconvolve the spatial data, adopting similar concepts applied for the deconvolution of bulk-RNAseq into single-cell profiles [145–147]. Indeed, methods like MuSiC [147] are often included in cell deconvolution benchmarks [148, 149], performing on par with methods designed ad-hoc for spatial data.

Cell2location [150] uses a reference single-cell dataset and the gene expression signature of the cell subpopulations in scRNA-seq data as input to infer gene expression at individual spatial locations into reference cell types. The authors were able to demonstrate cell2location ability to distinguish rare cell types, such as pre-germinal centre B cell population in a human lymphnode, and resolve the fine-grained immune cell types of the human gut. *RCTD* [151] employs supervised learning to estimate mixtures of cell types at each pixel. *SpatialDWLS* [149] first identifies the most likely cell types at each spot then a weighted-least-squares approach to infer cell type composition in the tissue.

SPOTlight [152], *DSTG* [153] and *CARD* [154] use the same pancreatic ductal adenocarcinoma (PDAC) dataset [155] with varying performances and results. SPOTlight was able to recover tumour-specific immune cell states in PDAC, applying a seeded non-negative matrix factorization (NMF) to obtain cell type-specific factors or topic profiles, then non-negative least squares (NNLS) regression is used to map each spot's transcriptome to a topic profile and determine the weights for each cell type that best fit each spot's topic profile by minimizing the residuals. Directly benchmarked against SPOTlight, DSTG additionally identifies spatial expression of marker genes associated with hypoxia and antigen presentation. DSTG leverages topological relations inside the data using graph-based convolutional networks to discover cell-type composition at spatial locations. CARD adapts the NMF framework to model spatial dependencies allowing a conditional autoregressive modelling on the columns of the inferred non-negative matrix. CARD could correctly infer global cell composition and gene expression of individual ST niches. *STRIDE* [156] trains a topic model on the scRNA-seq data to deconvolute cell types from spatial mixtures and was able to detect regulatory T lymphocytes at the interface between normal and tumour cells on a squamous cell carcinoma dataset [157]. *Stereoscope* [158] builds a probabilistic model to learn cell-type specific parameters on gene expression from scRNA-seq to obtain the cell mixtures in spatial data. *DestVI* [159] relies on variational inference to predict discrete cell-type-specific profiles and continuous latent variables of cell-states to describe the tissue architecture, and it is able to recover the interferon-induced changes in gene expression and cell-type composition of the spatial organization of lymph nodes upon bacterial infection. Recent methods can accomplish deconvolution without reference scRNA. *BayesTME* [160], a Bayesian generative model, can accurately infer cell type composition of ST data, identifying

immune cells at the interface with tumour cells in melanoma samples. The *SpatialGlue* [161] framework does not use a modality as a fixed reference, but learns a spatial proximity graph and a feature graph from each modality which are then the basis for the shared embedding, introducing within and across-modality attention aggregation layers to account for modality-dependent contributions. With this new strategy, the authors can reconcile the mismatched modality-dependent cell type annotations, correctly identifying the spatial distribution of B cells and T cells, and subpopulations of macrophages in individual spatial niches from a spleen [39] and thymus [162] multimodal datasets.

Integration of sc with imaging data

Integration of single cells with imaging data has the potential to complement the large body of histopathology and immunohistochemistry-based research with the mechanistic insights offered by single-cell sequencing data. After segmentation and quantification, images can be broken down into information that is complementary to single-cell sequencing data [121, 163] and integrated with standard approaches. For example, Tangram, novoSpaRc, SpaOTsc and Seurat's CCA described before have also the capacity to assign cells from scRNA-seq data to spatial locations in histological sections.

Alternative segmentation-free approaches have emerged that leverage the structural properties of tissue images to discover the spatial domains emerging from networks of dynamically interacting cells. Spatial data provides additional information that goes beyond nonmolecular features of cell representations, including morphology, the density of cells at individual locations and differential cell-to-cell communication in heterogeneous tissues. Integration of these with uncoupled single-cell readouts is possible [102, 164].

Methods like *SpaGCN* [165] and *Spage2vec* [166] rely on graph representation learning, a powerful deep learning framework that leverages relational information retained in local cell neighbourhoods [121]. SpaGCN demonstrates how incorporating histology information can improve the detection of cancer regions on the PDAC [155] dataset, using the inferred spatial domain to recover spatially variable genes associated with the disease [167, 168]. Similarly, *STlearn* [169] infers dynamic trajectories of biological processes in spatial data by integrating histology and gene expression and was able to detect immunoregulatory cell-to-cell interactions in breast cancer samples [170]. *SpaCell* [171] integrates tissue morphology and gene expression data to perform cell-type and disease-stage classification.

Given that the vast majority of datasets available come from unpaired data and a mixture of matched and unmatched samples, we anticipate that methods leveraging self-supervised learning such as contrastive learning and multiple-instance learning will be useful to learn models of biological diseases from high-dimensional data.

Methods for the integration of single cells within multiscale data collections

Datasets that result from collecting uncoupled modalities over time represent the largest body of data to analyse [172]. Strategies to integrate such datasets will need to make the heterogeneity a strength rather than a limitation. As single-cell datasets increase in size by profiling hundreds of patients, studies have now the power to link single-cell gene expression to genetic variation [173] or other data modalities profiled in bulk. In the context of integrating multiscale datasets, the difference in dimensionality and feature profiled represents the main challenge to obtaining a unified embedding [174]. The hierarchical nature of biological systems requires incorporating this structure in model design in the form of informative priors [175, 176]. Manifold alignment techniques could in principle provide an effective way of finding a low-dimensional embedding of multiscale data collections that preserves any known correspondences between them [177]. Similarly, strategies adopting cross-modal autoencoders can map heterogeneous data to a shared embedding and can learn holistic representations of cell and entire patient's physiological states [102, 164]. Finally, linear models like tensor decomposition [178] can prove powerful for the identification of pathways that are connected to disease progression and highlight key biomarkers for pharmacological treatment [45].

When the same patients are profiled across multiple platforms, approaches that focus on integrating data leveraging the common sample axis by summarizing the single cell expression into cell types are especially valuable for patient classification tasks and genetic association studies. Genome-wide association study (GWAS) have pinpointed risk genes and genetic variants for complex diseases. Variants that result in a shift in gene expression (*expression quantitative trait loci*, eQTLs) offer a handle for the interpretation of disease and tissue-dependent mechanisms of gene regulation [179]. Combining GWAS insights with single-cell readouts allows to link genetic and expression variations in individual cell types, uncovering cellular regulatory circuits at unprecedented resolution. Provided sufficient cell numbers per patient within individual cell types and appropriate sequencing coverage [180], strategies are emerging to discover eQTL from single cells [181], or more often, resorting to pseudobulking cell type expression profiles. Most notably, the large single-cell datasets generated in this context

have prioritized the analysis of PBMCs to identify links between risk loci of autoimmune disease and cell-type specific gene expression [182, 183]. Other studies proposed integrating genetics and single-cell readouts by combining GWAS summary statistics with individual cell types' gene expression [184–186] or with gene programs discovered in single cells. Those approaches led to the identification of immune cell regulation pathways in a host of immune-related diseases, including COVID-19, ulcerative colitis and asthma [25, 187], but also re-confirming immune components of neurological autoimmune diseases like AD and MS [187, 188]. In the future, as large datasets of organs become available [25], it will become easier to distil the tissue-specific genetic variation that is also associated with morbidities.

Integrative machine learning: challenges and opportunities

Multimodal integration of multiscale data collections promises a holistic approach to understanding complex disease mechanisms [174, 189–191]. Several challenges remain:

Selection of modalities, feature extraction and interpretation

Not all modalities contribute the same quantity and quality of information to the final biological question. Without prioritization of assays and analyses, researchers may produce costly experiments to only get stuck interpreting individual omics. Furthermore, as the data increases in size and complexity, the "curse of dimensionality" becomes a serious threat to modelling and thus impedes efficiently leveraging multimodal datasets [192, 193]. A hypothesis-driven approach would help scientists formulate and prioritize central questions, enabling them to define the relevance of each modality for the focal point of research [194]. Therefore, integration methods that allow quantifying and controlling for the contribution of individual modalities to the shared latent embedding may provide an intuitive framework for prioritizing assays. Feature extraction from individual omics is a way of selecting informative features, such that redundant features' information and noise can be minimized. Statistical tests to rank the importance of the features should take into account the dependencies between assays or are adjusted for multiple testing across the different modalities [195, 196]. Approaches like WNN [70] and TotalVI [96] although developed for paired integration include ways of quantifying the contribution of each modality to the final cell type prediction, either by associating weights to individual cells in each modality or directly quantifying how much variation is retained in joint latent representation.

Combining traditional factor models with DL methods allows to dissect the impact of covariates of interest on the individual modalities [197]. Methods that adopt mechanisms such as attention [198] or Shapley Values [199] to define features, modalities and cell types contributing to biologically relevant pathways will effectively provide a more interpretable framework to enable biomarker discovery.

Generating a common reference

The exponential growth of published references resulted in a babel of annotations and nomenclatures. The lack of shared annotation systems still represents a major drawback for immunologists and is effectively hindering the full exploitation of these complex datasets to identify actionable targets for study and therapy.

Building multimodal references may finally provide the framework to advance computational immunology, speeding up the process of cell annotation [22, 23, 200, 201]. Datasets that include flow or mass cytometry assays should be included in such multimodal references, to generate a resource that experienced immunologists will trust and computational immunologists can build upon. Similarly, deposited references generated by integrating single-cell proteomics and transcriptomics assays can be queried with any new unimodal dataset sharing at least a subset of the features [77, 78], without the need for time-consuming independent analysis of the new data.

Data infrastructure

Multimodal datasets are heterogeneous, including sparse or dense matrices, images and genomic regions, and after processing, alternative views of the data like dimensionality reduction and relational data such as graphs and ligand-receptor connections can be generated. These modalities require collecting all information in one container which allows fast access across the different layers and links the coupled modalities by their respective handle (a cell, or a patient sampled across modalities) (Fig. 2B). Data management infrastructures are emerging that respond to this need [202–206], and we anticipate they will define the foundational core for ML-integrative approaches moving forward. Finally, pipelines leveraging multimodal data containers [207, 208] offer a systematic approach for both customization and reproducibility and will speed up data processing while ensuring a stable foundation for new scientific discoveries.

Multimodal data, multidisciplinary teams

Generation, integration and interpretation of multimodal data calls for a diversity of expertise ranging from sample collection, data processing, storage and finally data analysis [209]. Multidisciplinary teams can tackle

these challenges in a multitude of ways because they can count on the rich background of unique team members. Beyond the obvious advantages, building a multidisciplinary team takes time and careful consideration, dedicated support and infrastructure [210, 211] and requires creating a culture where individual experts can feel comfortable in sharing their expertise with colleagues of different disciplines [212].

Integrating multimodal, multiscale data collections has emerged as an effective approach to address complex disease mechanisms, inform the prioritization of assays and speed up the process of biomarker discovery and cell annotation, ultimately facilitating the identification of actionable targets for study and therapy.

Conclusions and future directions

Using multimodal technologies, immunologists have gathered a wealth of multiscale measurements. ML approaches to integrate and interpret these data will be instrumental to understanding the mechanisms sustaining the fine regulation of the immune system in health and disease. To ensure that a vast audience of computational immunologists can benefit from these methods, it will be essential for developers to ensure the interpretability of the methods, availability of benchmarks across a wealth of conditions and well-documented use cases. Linear models will often be the preferred choice for their intuitive interpretation, especially in low-sample regimens. When data is not a limiting factor, DL and, more recently, generative AI methods, a new class of DL models with unprecedented abilities to generate new data that mimics real-world distributions, have started to show promise in numerous fields beyond their initial applications in image and text generation. These models have already proven powerful with traditionally complex tasks in immunological research, including the unbiased classification of the adaptive immune cells [89, 110–112] using positional sequence modelling [213, 214], protein structure prediction [215, 216], gene expression prediction from DNA sequences [217] and forecasting viral escape for pandemic preparedness [218]. In the future, these methods will allow to generate synthetic models of immune system behaviour under various conditions, offering insights for potential therapeutic interventions without the need for extensive laboratory experiments. Integrative methods will inform the design of new multiscale datasets, including environmental and lifestyle factors measurements, to study the role of the immune system in complex multifactorial diseases. In this outlook, computational immunologists will be pivotal in advancing scientific progress by leveraging integrative approaches to develop personalized immunological interventions. This includes the design of vaccines, immunotherapies and treatment plans that are finely tuned to individual immune system profiles, thereby enhancing the precision and effectiveness of medical treatments. With this review, we have offered a short overview of the emerging challenges and opportunities of multimodal integration applied to multiscale datasets. Recent scientific successes have started to reward those who invested in generating multimodal datasets, engineering software and fostering collaboration in multidisciplinary teams. We expect that the growing body of research on this topic will empower researchers and encourage many others to embrace the multimodal revolution.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-024-01350-3.

Additional file 1: Table S1. Integration methods cited in the review.

Acknowledgements
We wish to thank Dr Hannah Spitzer, Lennard Halle, Dr Luke Zappia, Karin Hrovatin, Lukas Heumos, Lilly May and Dr Ignacio Ibarra for revising the manuscript and for feedback on figures, and Dr Carlos Talavera-López for inspiring discussions on the topic of computational immunology. Figures were created with BioRender.com.

Authors' contributions
FC and FJT conceived the study and wrote the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. FC is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) –SFB- TRR 338/1 2021 –452881907.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Cellarity and has ownership interest in Dermagnostix GmbH and Cellarity. The remaining author declares no competing interests.

Received: 29 June 2023 Accepted: 23 May 2024
Published online: 11 June 2024

References

1. Kotas ME, Medzhitov R. Homeostasis, inflammation, and disease susceptibility. Cell. 2015;160:816–27.
2. Bettcher BM, Tansey MG, Dorothée G, Heneka MT. Peripheral and central immune system crosstalk in Alzheimer disease - a research prospectus. Nat Rev Neurol. 2021;17:689–701.

3. Meltzer A, Van de Water J. The role of the immune system in autism spectrum disorder. Neuropsychopharmacology. 2017;42:284–98.
4. Donath MY, Dinarello CA, Mandrup-Poulsen T. Targeting innate immune mediators in type 1 and type 2 diabetes. Nat Rev Immunol. 2019;19:734–46.
5. Hiam-Galvez KJ, Allen BM, Spitzer MH. Systemic immunity in cancer. Nat Rev Cancer. 2021;21:345–59.
6. Quah HS, Cao EY, Suteja L, Li CH, Leong HS, Chong FT, et al. Single cell analysis in head and neck cancer reveals potential immune evasion mechanisms during early metastasis. Nat Commun. 2023;14:1680.
7. Kwok AJ, Allcock A, Ferreira RC, Cano-Gamez E, Smee M, Burnham KL, et al. Neutrophils and emergency granulopoiesis drive immune suppression and an extreme response endotype during sepsis. Nat Immunol. 2023;24:767–79.
8. Lutgens E, Atzler D, Döring Y, Duchene J, Steffens S, Weber C. Immunotherapy for cardiovascular disease. Eur Heart J. 2019;40:3937–46.
9. Esfahani K, Roudaia L, Buhlaiga N, Del Rincon SV, Papneja N, Miller WH Jr. A review of cancer immunotherapy: from the past, to the present, to the future. Curr Oncol. 2020;27:S87-97.
10. Abdelfattah N, Kumar P, Wang C, Leu J-S, Flynn WF, Gao R, et al. Single-cell analysis of human glioma and immune cells identifies S100A4 as an immunotherapy target. Nat Commun. 2022;13:767.
11. Kaufmann SHE. Immunology's coming of age. Front Immunol. 2019;10:684.
12. Picot J, Guerin CL, Le Van KC, Boulanger CM. Flow cytometry: retrospective, fundamentals and recent instrumentation. Cytotechnology. 2012;64:109–30.
13. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem. 2009;81:6813–22.
14. Pittet MJ, Michielin O, Migliorini D. Clinical relevance of tumour-associated macrophages. Nat Rev Clin Oncol. 2022;19:402–21.
15. Gomes T, Teichmann SA, Talavera-López C. Immunology driven by large-scale single-cell sequencing. Trends Immunol. 2019;40:1011–21.
16. Ginhoux F, Yalin A, Dutertre CA, Amit I. Single-cell immunology: Past, present, and future. Immunity. 2022;55:393–404.
17. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018;18:35–45.
18. Efremova M, Vento-Tormo R, Park J-E, Teichmann SA, James KR. Immunology in the era of single-cell technologies. Annu Rev Immunol. 2020;38:727–57.
19. Rood JE, Maartens A, Hupalowska A, Teichmann SA, Regev A. Impact of the human cell Atlas on medicine. Nat Med. 2022;28:2486–96.
20. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.
21. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. Elife. 2017;6:e27041. https://doi.org/10.7554/eLife.27041.
22. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376:eabl5197.
23. Eraslan G, Drokhlyansky E, Anand S, Fiskin E, Subramanian A, Slyper M, et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. Science. 2022;376:eabl4290.
24. Suo C, Dann E, Goh I, Jardine L, Kleshchevnikov V, Park J-E, et al. Mapping the developing human immune system across organs. Science. 2022;376:eabo0510.
25. Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, et al. An integrated cell atlas of the lung in health and disease. Nat Med. 2023;29:1563–77.
26. Delorey TM, Ziegler CGK, Heimberg G, Normand R, Yang Y, Segerstolpe Å, et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature. 2021;595:107–13.
27. Luo G, Gao Q, Zhang S, Yan B. Probing infectious disease by single-cell RNA sequencing: progresses and perspectives. Comput Struct Biotechnol J. 2020;18:2962–71.
28. Yang W, Liu L-B, Liu F-L, Wu Y-H, Zhen Z-D, Fan D-Y, et al. Single-cell RNA sequencing reveals the fragility of male spermatogenic cells to Zika virus-induced complement activation. Nat Commun. 2023;14:2476.
29. Melms JC, Biermann J, Huang H, Wang Y, Nair A, Tagore S, et al. A molecular single-cell lung atlas of lethal COVID-19. Nature. 2021;595:114–9.
30. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med. 2020;26:681–7.
31. Edahiro R, Shirai Y, Takeshima Y, Sakakibara S, Yamaguchi Y, Murakami T, et al. Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. Nat Genet. 2023;55:753–67. https://doi.org/10.1038/s41588-023-01375-1.
32. Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. Nat Methods. 2020;17:11–4.
33. Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nat Biotechnol. 2021;39:236–45.
34. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24(8):494–515.
35. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560:494–8.
36. Vistain L, Van Phan H, Keisham B, Jordi C, Chen M, Reddy ST, et al. Quantification of extracellular proteins, protein complexes and mRNAs in single cells by proximity sequencing. Nat Methods. 2022;19:1578–89.
37. Yue L, Liu F, Hu J, Yang P, Wang Y, Dong J, et al. A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. Comput Struct Biotechnol J. 2023;21:940–55.
38. Llorens-Bobadilla E, Zamboni M, Marklund M, Bhalla N, Chen X, Hartman J, et al. Solid-phase capture and profiling of open chromatin by spatial ATAC. Nat Biotechnol. 2023;41:1085–8. https://doi.org/10.1038/s41587-022-01603-9.
39. Ben-Chetrit N, Niu X, Swett AD, Sotelo J, Jiao MS, Stewart CM, et al. Integration of whole transcriptome spatial profiling with protein markers. Nat Biotechnol. 2023;41:788–93. https://doi.org/10.1038/s41587-022-01536-3.
40. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
41. Mair F, Tyznik AJ. High-dimensional immunophenotyping with fluorescence-based cytometry: a practical guidebook. Methods Mol Biol. 2019;2032:1–29.
42. Maby P, Corneau A, Galon J. Phenotyping of tumor infiltrating immune cells using mass-cytometry (CyTOF). Methods Enzymol. 2020;632:339–68.
43. Gadalla R, Noamani B, MacLeod BL, Dickson RJ, Guo M, Xu W, et al. Validation of CyTOF against flow cytometry for immunological studies and monitoring of human cancer clinical trials. Front Oncol. 2019;9:415.
44. Shadmani G, Simkins TJ, Assadsangabi R, Apperson M, Hacein-Bey L, Raslan O, et al. Autoimmune diseases of the brain, imaging and clinical review. Neuroradiol J. 2022;35:152–69.
45. COMBAT Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. Cell. 2022;185:916-38.e58.
46. Tietscher S, Wagner J, Anzeneder T, Langwieder C, Rees M, Sobottka B, et al. A comprehensive single-cell map of T cell exhaustion-associated immune environments in human breast cancer. Nat Commun. 2023;14:98.
47. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20:257–72.
48. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol. 2021. https://doi.org/10.1038/s41587-021-01001-7.
49. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.
50. Dugourd A, Kuppe C, Sciacovelli M, Gjerga E, Gabor A, Emdal KB, et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol Syst Biol. 2021;17:e9730.
51. Moffitt JR, Lundberg E, Heyn H. The emerging landscape of spatial profiling technologies. Nat Rev Genet. 2022;23:741–59.

52. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: assessing the impact of missing data in electronic health records. J Biomed Inform. 2023;139:104269.
53. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics. 2012;13(Suppl 16):S5.
54. Song M, Greenbaum J, Luttrell J 4th, Zhou W, Wu C, Shen H, et al. A review of integrative imputation for multi-omics datasets. Front Genet. 2020;11:570255.
55. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–76.
56. Oh S, Kang DD, Brock GN, Tseng GC. Biological impact of missing-value imputation on downstream analyses of gene expression profiles. Bioinformatics. 2011;27:78–86.
57. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021. https://doi.org/10.1038/s41587-021-00895-7.
58. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf Fusion. 2019;50:71–91.
59. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14:865–8.
60. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361:1380–5.
61. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8.
62. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
63. Urban PL. Quantitative mass spectrometry: an overview. Philos Trans A Math Phys Eng Sci. 2016;374. https://doi.org/10.1098/rsta.2015.0382.
64. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, et al. Next-generation sequencing of T and B cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. Immunity. 2020;53:442-55.e4.
65. Liu Y, DiStasio M, Su G, Asashima H, Enninful A, Qin X, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. Nat Biotechnol. 2023;41:1405–9.
66. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353:78–82.
67. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015;348:aaa6090.
68. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2021;19:41–50.
69. 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;21:111.
70. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-87.e29.
71. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
72. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389–403.
73. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177:1873-87.e17.
74. Kriebel AR, Welch JD. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat Commun. 2022;13:780.
75. Wang H-T, Smallwood J, Mourao-Miranda J, Xia CH, Satterthwaite TD, Bassett DS, et al. Finding the needle in a high-dimensional haystack: canonical correlation analysis for neuroscientists. Neuroimage. 2020;216:116745.
76. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
77. Repapi E, Agarwal D, Napolitani G, Sims D, Taylor S. Integration of single-cell RNA-Seq and CyTOF data characterises heterogeneity of rare cell subpopulations. F1000Res. 2022;11:560.
78. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Satija R. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42(2):293–304.
79. Pedersen CB, Dam SH, Barnkob MB, Leipold MD, Purroy N, Rassenti LZ, et al. cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. Nat Commun. 2022;13:1698.
80. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
81. Triana S, Vonficht D, Jopp-Saile L, Raffel S, Lutz R, Leonce D, et al. Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. Nat Immunol. 2021;22:1577–89.
82. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15:359–62.
83. Zhu B, Chen S, Bai Y, Chen H, Liao G, Mukherjee N, et al. Robust single-cell matching and multimodal analysis using shared and distinct features. Nat Methods. 2023;20:304–15.
84. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell. 2018;174:968-81.e15.
85. Demetci P, Santorella R, Sandstede B, Noble WS, Singh R. SCOT: single-cell multi-omics alignment with optimal transport. J Comput Biol. 2022;29:3–18.
86. Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics. 2021. https://doi.org/10.1093/bioinformatics/btab594.
87. Ghazanfar S, Guibentif C, Marioni JC. Stabilized mosaic single-cell data integration using unshared features. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-023-01766-z.
88. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 2017;18:138.
89. Schattgen SA, Guion K, Crawford JC, Souquette A, Barrio AM, Stubbington MJT, et al. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). Nat Biotechnol. 2022;40:54–63.
90. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
91. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10:390.
92. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: a review of methods and applications. AI Open. 2020;1:57–81.
93. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y. Generative adversarial nets. Adv Neural Informa Process Syst. 2014;27.
94. Amodio M, Krishnaswamy S. MAGAN: aligning biological manifolds. In: Dy JG, Krause A, editors. In Proc. 35th International Conference on Machine Learning. PMLR; 2018. p. 215–23.
95. Suzuki M, Nakayama K, Matsuo Y. Joint multimodal learning with deep generative models. arXiv [stat.ML]. 2016. Available from: http://arxiv.org/abs/1611.01891.
96. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 2021;18:272–82.
97. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022. https://doi.org/10.1038/s41587-022-01284-4.
98. Lotfollahi M, Litinetskaya A, Theis FJ. Multigrate: single-cell multi-omic data integration. Available from: https://icml-compbio.github.io/2021/papers/WCBICML2021_paper_44.pdf. [cited 2021 Sep 22].
99. Stark SG, Ficek J, Locatello F, Bonilla X, Chevrier S, Singer F, et al. SCIM: universal single-cell matching with unpaired feature sets. Bioinformatics. 2020;36:i919-27.

100. Satoh J-I, Asahina N, Kitano S, Kino Y. A comprehensive profile of ChIP-Seq-based PU.1/Spi1 target genes in microglia. Gene Regul Syst Bio. 2014;8:127–39.
101. Irmisch A, Bonilla X, Chevrier S, Lehmann K-V, Singer F, Toussaint NC, et al. The Tumor Profiler Study: integrated, multi-omic, functional tumor profiling for clinical decision support. Cancer Cell. 2021;39:288–93.
102. Yang KD, Belyaeva A, Venkatachalapathy S, Damodaran K, Katcoff A, Radhakrishnan A, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat Commun. 2021;12:31.
103. He Z, Hu S, Chen Y, An S, Zhou J, Liu R, et al. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. Nat Biotechnol. 2024. https://doi.org/10.1038/s41587-023-02040-y.
104. Godfrey DI, Uldrich AP, McCluskey J, Rossjohn J, Moody DB. The burgeoning family of unconventional T cells. Nat Immunol. 2015;16:1114–23.
105. Overgaard NH, Jung J-W, Steptoe RJ, Wells JW. CD4+/CD8+ double-positive T cells: more than just a developmental stage? J Leukoc Biol. 2015;97:31–8.
106. Murre C. Defining the pathways of early adult hematopoiesis. Cell Stem Cell. 2007;1:357–8.
107. Li Y, Zhang D, Yang M, Peng D, Yu J, Liu Y, et al. scBridge embraces cell heterogeneity in single-cell RNA-seq and ATAC-seq data integration. Nat Commun. 2023;14:6045.
108. Lin Y, Wu T-Y, Wan S, Yang JYH, Wong WH, Wang YXR. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40:703–10.
109. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 2021. https://doi.org/10.1038/s41587-021-00927-2.
110. Zhang Z, Xiong D, Wang X, Liu H, Wang T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat Methods. 2021;18:92–9.
111. Zhang Z, Chang WY, Wang K, Yang Y, Wang X, Yao C, et al. Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse. Nat Mach Intell. 2022;4:596–604.
112. Drost F, An Y, Dratva LM, Lindeboom RGH, Haniffa M, Teichmann SA, et al. Integrating T-cell receptor and transcriptome for large-scale single-cell immune profiling analysis. bioRxiv. 2022. p. 2021.06.24.449733. Available from: https://www.biorxiv.org/content/10.1101/2021.06.24.449733v2. [cited 2023 May 3].
113. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome Med. 2022;14:68.
114. Fu T, Dai L-J, Wu S-Y, Xiao Y, Ma D, Jiang Y-Z, et al. Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response. J Hematol Oncol. 2021;14:98.
115. Hirz T, Mei S, Sarkar H, Kfoury Y, Wu S, Verhoeven BM, et al. Dissecting the immune suppressive human prostate tumor microenvironment via integrated single-cell and spatial transcriptomic analyses. Nat Commun. 2023;14:663.
116. Sorin M, Rezanejad M, Karimi E, Fiset B, Desharnais L, Perus LJM, et al. Single-cell spatial landscapes of the lung tumour immune microenvironment. Nature. 2023;614:548–54.
117. Kaufmann M, Schaupp A-L, Sun R, Coscia F, Dendrou CA, Cortes A, et al. Identification of early neurodegenerative pathways in progressive multiple sclerosis. Nat Neurosci. 2022;25:944–55.
118. Elmentaite R, Domínguez Conde C, Yang L, Teichmann SA. Single-cell atlases: shared and tissue-specific cell types across human organs. Nat Rev Genet. 2022;23:395–410.
119. Karimi E, Yu MW, Maritan SM, Perus LJM, Rezanejad M, Sorin M, et al. Single-cell spatial immune landscapes of primary and metastatic brain tumours. Nature. 2023;614:555–63.
120. Lein E, Borm LE, Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. Science. 2017;358:64–9.
121. Palla G, Fischer DS, Regev A, Theis FJ. Spatial components of molecular tissue biology. Nat Biotechnol. 2022;40:308–18.
122. Elmentaite R, Kumasaka N, Roberts K, Fleming A, Dann E, King HW, et al. Cells of the human intestinal tract mapped across space and time. Nature. 2021;597:250–5.
123. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. Cancer Res. 2018;78:5970–9.
124. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science. 2018;362. https://doi.org/10.1126/science.aau5324.
125. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods. 2021;18:1352–62.
126. Giesen C, Wang HAO, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat Methods. 2014;11:417–22.
127. Radtke AJ, Kandov E, Lowekamp B, Speranza E, Chu CJ, Gola A, et al. IBEX: A versatile multiplex optical imaging approach for deep phenotyping and spatial analysis of cells in complex tissues. Proc Natl Acad Sci U S A. 2020;117:33455–65.
128. Walch A, Rauser S, Deininger S-O, Höfler H. MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. Histochem Cell Biol. 2008;130:421–34.
129. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, et al. Multiplexed ion beam imaging of human breast tumors. Nat Med. 2014;20:436–42.
130. Saka SK, Wang Y, Kishi JY, Zhu A, Zeng Y, Xie W, et al. Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nat Biotechnol. 2019;37:1080–90.
131. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods. 2019;16:987–90.
132. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363:1463–7.
133. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22:627–44.
134. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods. 2022;19:662–70.
135. Amodio M, Youlten SE, Venkat A, San Juan BP, Chaffer CL, Krishnaswamy S. Single-cell multi-modal GAN reveals spatial patterns in single-cell data from triple-negative breast cancer. Patterns (N Y). 2022;3:100577.
136. Hussein YR, Bandyopadhyay S, Semaan A, Ahmed Q, Albashiti B, Jazaerly T, et al. Glut-1 Expression Correlates with Basal-like Breast Cancer. Transl Oncol. 2011;4:321–7.
137. Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv [cs.LG]. 2019. Available from: http://arxiv.org/abs/1905.02269.
138. Shengquan C, Boheng Z, Xiaoyang C, Xuegong Z, Rui J. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. Bioinformatics. 2021;37:i299-307.
139. Abdelaal T, Mourragui S, Mahfouz A, Reinders MJT. SpaGE: spatial gene enhancement using scRNA-seq. Nucleic Acids Res. 2020;48:e107.
140. Mourragui S, Loog M, van de Wiel MA, Reinders MJT, Wessels LFA. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. Bioinformatics. 2019;35:i510-9.
141. Villani C. Optimal transport: old and new, vol. 338. Berlin: Springer; 2009. p. 23.
142. Nitzan M, Karaiskos N, Friedman N, Rajewsky N. Gene expression cartography. Nature. 2019;576:132–7.
143. Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat Commun. 2020;11:2084.
144. Klein D, Palla G, Lange M, Klein M, Piran Z, Gander M, et al. Mapping cells through time and space with moscot. bioRxiv. 2023. p. 2023.05.11.540374. Available from: https://www.biorxiv.org/content/10.1101/2023.05.11.540374. [cited 2023 Jun 5].

145. Aliee H, Theis FJ. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. Cell Syst. 2021;12:706-15.e4.
146. Erdmann-Pham DD, Fischer J, Hong J, Song YS. Likelihood-based deconvolution of bulk gene expression data using single-cell references. Genome Res. 2021;31:1794–806.
147. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10:380.
148. Coleman K, Hu J, Schroeder A, Lee EB, Li M. SpaDecon: cell-type deconvolution in spatial transcriptomics with semi-supervised learning. Commun Biol. 2023;6:378.
149. Dong R, Yuan G-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. Genome Biol. 2021;22:145.
150. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell 2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40:661–71.
151. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40:517–26.
152. Elosua-Bayes M, Nieto P, Mereu E, Gut I, Heyn H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res. 2021;49:e50.
153. Song Q, Su J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbaa414.
154. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. Nat Biotechnol. 2022;40:1349–59.
155. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol. 2020;38:333–42.
156. Sun D, Liu Z, Li T, Wu Q, Wang C. STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. Nucleic Acids Res. 2022;50:e42.
157. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. Cell. 2020;182:1661–2.
158. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Commun Biol. 2020;3:565.
159. Lopez R, Li B, Keren-Shaul H, Boyeau P, Kedmi M, Pilzer D, et al. DestVI identifies continuums of cell types in spatial transcriptomics data. Nat Biotechnol. 2022;40:1360–9.
160. Zhang H, Hunter MV, Chou J, Quinn JF, Zhou M, White RM, et al. BayesTME: an end-to-end method for multiscale spatial transcriptional profiling of the tissue microenvironment. Cell Syst. 2023;14:605-19.e7.
161. Long Y, Ang KS, Liao S, Sethi R, Heng Y, Zhong C, et al. Integrated analysis of spatial multi-omics with SpatialGlue. bioRxiv. 2023. p. 2023.04.26.538404. Available from: https://www.biorxiv.org/content/biorxiv/early/2023/05/02/2023.04.26.538404. [cited 2024 Jan 29].
162. Liao S, Heng Y, Liu W, Xiang J, Ma Y, Chen L, et al. Integrated spatial transcriptomic and proteomic analysis of fresh frozen tissue based on stereo-seq. bioRxiv. 2023. p. 2023.04.28.538364. Available from: https://www.biorxiv.org/content/10.1101/2023.04.28.538364v1. [cited 2024 Jan 29].
163. Schapiro D, Sokolov A, Yapp C, Chen Y-A, Muhlich JL, Hess J, et al. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nat Methods. 2022;19:311–5.
164. Radhakrishnan A, Friedman SF, Khurshid S, Ng K, Batra P, Lubitz SA, et al. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. Nat Commun. 2023;14:2436.
165. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021;18:1342–51.
166. Partel G, Wählby C. Spage2vec: Unsupervised representation of localized spatial gene expression signatures. FEBS J. 2021;288:1859–70.
167. Li D, Ni X-F, Tang H, Zhang J, Zheng C, Lin J, et al. KRT17 functions as a tumor promoter and regulates proliferation, migration and invasion in pancreatic cancer via mTOR/S6k1 pathway. Cancer Manag Res. 2020;12:2087–95.
168. Lee J, Lee J, Kim JH. Identification of matrix metalloproteinase 11 as a prognostic biomarker in pancreatic cancer. Anticancer Res. 2019;39:5963–71.
169. Pham D, Tan X, Balderson B, Xu J, Grice LF, Yoon S, et al. Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. Nat Commun. 2023;14:7739.
170. Lin W, Xu D, Austin CD, Caplazi P, Senger K, Sun Y, et al. Function of CSF1 and IL34 in macrophage homeostasis, inflammation, and cancer. Front Immunol. 2019;10:2019.
171. Tan X, Su A, Tran M, Nguyen Q. SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells. Bioinformatics. 2020;36:2293–4.
172. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, et al. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit Med. 2019;2:115.
173. Cuomo ASE, Nathan A, Raychaudhuri S, MacArthur DG, Powell JE. Single-cell genomics meets human genetics. Nat Rev Genet. 2023. https://doi.org/10.1038/s41576-023-00599-5.
174. Schaffer LV, Ideker T. Mapping the multiscale structure of biological systems. Cell Syst. 2021;12:622–35.
175. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell. 2020;38:672-84.e6.
176. Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34:660–8.
177. Huang J, Sheng J, Wang D. Manifold learning analysis suggests strategies to align single-cell multimodal data of neuronal electrophysiology and transcriptomics. Commun Biol. 2021;4:1308.
178. Hore V, Viñuela A, Buil A, Knight J, McCarthy MI, Small K, et al. Tensor decomposition for multiple-tissue gene expression experiments. Nat Genet. 2016;48:1094–100.
179. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16:197–212.
180. Cuomo ASE, Alvari G, Azodi CB, single-cell eQTLGen consortium, McCarthy DJ, Bonder MJ. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. Genome Biol. 2021;22:188.
181. Cuomo ASE, Heinen T, Vagiaki D, Horta D, Marioni JC, Stegle O. Cell RegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. Mol Syst Biol. 2022;18:e10663.
182. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science. 2022;376:eabf3041.
183. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science. 2022;376:eabf1970.
184. Calderon D, Bhaskar A, Knowles DA, Golan D, Raj T, Fu AQ, et al. Inferring relevant cell types for complex traits by using single-cell gene expression. Am J Hum Genet. 2017;101:686–99.
185. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228–35.
186. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219.
187. Jagadeesh KA, Dey KK, Montoro DT, Mohan R, Gazal S, Engreitz JM, et al. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nat Genet. 2022;54:1479–92.
188. Ma Y, Qiu F, Deng C, Li J, Huang Y, Wu Z, et al. Integrating single-cell sequencing data with GWAS summary statistics reveals CD16+monocytes and memory CD8+T cells involved in severe COVID-19. Genome Med. 2022;14:16.
189. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022;22:114–26.
190. Cappuccio A, Tieri P, Castiglione F. Multiscale modelling in immunology: a review. Brief Bioinform. 2016;17:408–18.

191. Abedi V, Hontecillas R, Carbo A, Philipson C, Hoops S, Bassaganya-Riera J. Chapter 8 - Multiscale modeling: concepts, technologies, and use cases in immunology. In: Bassaganya-Riera J, editor. Computational immunology Models and Tools. 1st ed. Cambridge: Elsevier Academic Press; 2016. p. 145–73.
192. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28:1773–84.
193. Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med. 2021;4:153.
194. Bonaguro L, Schulte-Schrepping J, Ulas T, Aschenbrenner AC, Beyer M, Schultze JL. A guide to systems-level immunomics. Nat Immunol. 2022;23:1412–23.
195. Garali I, Adanyeguh IM, Ichou F, Perlbarg V, Seyer A, Colsch B, et al. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. Brief Bioinform. 2018;19:1356–69.
196. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020;21:630–44.
197. Lotfollahi M, Klimovskaia Susmelj A, De Donno C, Hetzel L, Ji Y, Ibarra IL, et al. Predicting cellular responses to complex perturbations in high-throughput screens. Mol Syst Biol. 2023;19:e11517.
198. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv [cs.CL]. 2017. Available from: http://arxiv.org/abs/1706.03762.
199. Shapley L. A value for n-person games. contributions to the theory of games ii (1953) 307-317. In Classics in game theory. Princeton: Princeton University Press; 2020. p. 69–79.
200. García-Alonso L, Zondervan KT, Vento-Tormo R. A novel resource to study endometriosis at the single-cell level. Nat Rev Endocrinol. 2023;19:256–7 Nature Publishing Group.
201. Garcia-Alonso L, Lorenzi V, Mazzeo CI, Alves-Lopes JP, Roberts K, Sancho-Serra C, et al. Single-cell roadmap of human gonadal development. Nature. 2022;607:540–7.
202. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19:171–8.
203. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
204. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol. 2023;41:604–6. https://doi.org/10.1038/s41587-023-01733-8.
205. Dries R, Zhu Q, Dong R, Eng CHL, Li H, Liu K, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. 2021;22:78.
206. Marconato L, Palla G, Yamauchi KA, Virshup I, Heidari E, Treis T, et al. SpatialData: an open and universal data framework for spatial omics. Nat Methods. 2024. https://doi.org/10.1038/s41592-024-02212-x.
207. Rich-Griffin C, Curion F, Thomas T, Agarwal D, Theis FJ, Dendrou CA. Pan-pipes: a pipeline for multiomic single-cell data analysis. bioRxiv. 2023. p. 2023.03.11.532085. Available from: https://www.biorxiv.org/content/10.1101/2023.03.11.532085v1. [cited 2023 Mar 20].
208. Curion F, Wu X, Heumos L, André MMG, Halle L, Ozols M, et al. hadge: a comprehensive pipeline for donor deconvolution in single-cell studies. Genome Biol. 2024;25:109.
209. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31.
210. Woolston C. Why science needs more research software engineers. Nature. 2022.
211. Disis ML, Slattery JT. The road we must take: multidisciplinary team science. Sci Transl Med. 2010;2:22cm9.
212. Wang S, Marr LC, Contreras LM, Theis FJ, Nurse P. The challenges in finding your home as a multidisciplinary scientist. Cell. 2022;185:2623–5.
213. Pertseva M, Gao B, Neumeier D, Yermanos A, Reddy ST. Applications of machine and deep learning in adaptive immunity. Annu Rev Chem Biomol Eng. 2021;12:39–62.
214. Sidhom J-W, Oliveira G, Ross-MacDonald P, Wind-Rotolo M, Wu CJ, Pardoll DM, et al. Deep learning reveals predictive sequence concepts within immune repertoires to immunotherapy. Sci Adv. 2022;8:eabq5089.
215. Abanades B, Wong WK, Boyles F, Georges G, Bujotzek A, Deane CM. ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. Commun Biol. 2023;6:575.
216. Ruffolo JA, Chu L-S, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun. 2023;14:2389.
217. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. bioRxiv. 2023. p. 2023.08.30.555582. Available from: https://www.biorxiv.org/content/10.1101/2023.08.30.555582v1. [cited 2024 Feb 16].
218. Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Ritter D, et al. Learning from prepandemic data to forecast viral escape. Nature. 2023;622:818–25.

Publisher's Note