Briefings in Bioinformatics, 2026, 27, bbag608
https://doi.org/10.1093/bib/bbag608
Published: 15 November 2026
Problem Solving Protocol

OXFORD

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

Integration of single-cell multi-omic data with graph-based topic modelling

Gabriele Malagoli[1,2], Filippo Valle[3], Andreina Tirabassi[1,4], Annalisa Marsico[1], Loredana Martignetti[5], Michele Caselle[3,*,‡], Maria Colomé-Tatché[1,2,6,*,‡]

[1]Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
[2]Biomedical Center, Division of Physiological Chemistry, Faculty of Medicine, Ludwig-Maximilians-Universität Munich, Großhaderner Str. 9, 82152 Planegg-Martinsried, Germany
[3]Physics Department, University of Turin and INFN, Via Pietro Giuria 1, 10125 Torino, Italy
[4]TUM School of Computation, Information and Technology (CIT), Boltzmannstraße 3, 85748 Garching bei München, Germany
[5]Institut Curie, Inserm U1331, Mines ParisTech, PSL Research University, 26 rue d'Ulm, 75248 Paris cedex 05, France
[6]Hospital del Mar Research Institute (HMRIB), C/Doctor Aiguader 88, E-08003, Barcelona, Spain
*Corresponding authors. Maria Colomé-Tatché, Hospital del Mar Research Institute (HMRIB), C/Doctor Aiguader 88, E-08003, Barcelona, Spain. E-mail: mcolome@researchmar.net; Michele Caselle, Physics Department, University of Turin and INFN, Via Pietro Giuria 1, 10125 Torino, Italy. E-mail: michele.caselle@unito.it.
‡Michele Caselle and Maria Colomé-Tatché are joint senior authors.

Abstract

Recent advances in single-cell biology enable the profiling of multiple molecular layers, such as the transcriptome, epigenome, and surface proteins, within a single cell. Tackling the complexity of these data from multiple perspectives allows researchers to gain the most comprehensive insights into cellular biological properties. Here, we propose a graph-based topic modelling method called bionSBM. Our method leverages well-known community-detection methods for multipartite graphs and the interpretability of topic modelling to cluster and explain high-dimensional, sparse, and noisy single-cell matrices. We applied our algorithm to paired single-cell multi-omics data, such as 10X Multiome, SHARE-seq, and CITE-seq. We showed that it achieves superior performance compared with state-of-the-art methods for ground-truth label retrieval, with high specificity and distinct biological interpretability.

Keywords **single-cell multi-omics, topic modelling, stochastic block model, multipartite networks, gene regulatory programs, data integration**

Background

The emergent phenotypes of biological systems result from complex molecular interactions. These interactions play a role at multiple levels of gene expression regulation: genome, epigenome, transcriptome, and proteome. Characterizing such regulatory behaviour is crucial to studying these processes. It is an essential step towards the development of drugs and targeted, personalized medicine strategies [1], as well as for other translational applications such as drug discovery and validation [2] or tumour microenvironment profiling [3]. The latest advances in single-cell sequencing technologies, such as 10X Multiome and SHARE-seq [4], enable profiling of gene expression and chromatin accessibility at the single-cell level. Moreover, technologies such as CITE-seq [5] allow profiling not only of gene expression but also of surface molecules using antibody-derived tags (ADTs). All these techniques are opening new avenues for understanding complex gene regulatory mechanisms [6].

The development of experimental protocols poses a new challenge for the scientific community: algorithms and downstream pipelines must be adapted to integrate paired and multi-omic data. In particular, when analysing these data, one is typically interested in their modular structure. The main goal is to identify groups of molecular features that act in a coordinated way, as well as groups of cells with similar molecular patterns. Given the multimodal nature of the data, the focus can be on gene groups, ATAC-seq peaks, ADTs, or any features captured by the measurements. Moreover, it is of interest to link these modular structures across omics layers and use them to infer cell-level clusters, a highly non-trivial task. Several algorithms have been proposed in recent years to address this task. A popular approach is deep learning. Deep learning methods, such as autoencoders [7], perform effective dimensionality reduction but are sensitive to differences in input scale, a key characteristic of multimodal data [8]. More recent approaches have extended representation learning to increasingly heterogeneous settings, including contrastive learning for single-cell mosaic integration in ACE [9] and graph neural networks combined with contrastive learning for spatial multi-omics in SpaMosaic [10]. Network-based integration methods provide

Received: April 30, 2026. **Revised:** October 8, 2026. **Accepted:** October 13, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

another strategy and differ substantially in how omics measurements are represented as graphs. For example, NIC learns modality-specific cell–cell similarity networks from transcriptomic and epigenomic profiles and integrates them through joint matrix factorization [11], whereas other multi-network approaches construct gene–gene networks from protein interactions and methylation data [12], or from differential co-expression and co-methylation patterns [13].

Here, we focus instead on a class of algorithms based on topic modelling. Topic modelling algorithms were initially developed in natural language processing (NLP) to automatically identify 'topics' in a corpus of texts, using only word frequencies as input. They can group sets of words that 'speak about the same theme' and, in this way, identify the texts' topics. There is an obvious analogy between this task and the one we are interested in. If we map texts to cells, and the frequency of words to gene expression levels, chromatin accessibility at ATAC-seq peaks, or surface protein levels, the same approach described above can be used to identify and characterize cell populations based on their patterns of molecular measurements. In this analogy, topics are groups of genes/open genomic loci/surface proteins that act cooperatively, and these regulomes and pathways can be used to characterize different cell populations. The rationale for using topic modelling methods for single-cell data is that the cell transcriptome exhibits statistical properties similar to those of words in text [14]. One of the most interesting features of topic modelling algorithms is that they output probability distributions over cells for each topic, which better describe the complexity of biological systems than a naive, deterministic assignment in which cells are simply assigned to a group (or cluster). The topics, on the other hand, are not only lists of loci; they are described as probability distributions over all genomic features.

The most popular topic models are based on Latent Dirichlet Allocation (LDA) [15], Non-Negative Matrix Factorization (NMF) [16], or Stochastic Block Models (SBM) [17]. LDA approaches the problem of topic modelling by assuming a Dirichlet prior for the distribution of words within topics and of topics within documents (in our setting, corresponding to the distributions of genes, peaks, and ADTs within topics, and of topics within cells). The choice of this prior is mostly for analytical and computational reasons, lacking a real biological motivation [18]. Since the Dirichlet distribution is unimodal, regardless of the choice of hyperparameters, it assumes a single shape for the distribution of words across topics and topics across documents, which is a strong homogeneity assumption that is difficult to reconcile with the heterogeneity of biological systems [19]. On the other hand, NMF has two main limitations: its linearity limits the ability to capture complex interactions, which are typical in biological systems, and it tends to find steep boundaries between clusters, which can be problematic when dealing with highly heterogeneous data, as in single-cell sequencing [18].

We will therefore concentrate on the following in the third category of topic models, based on SBM [17, 19, 20]. We present a graph-based method, bionSBM, for integrating and interpreting paired single-cell multi-omic data [Fig. 1]. In contrast to approaches that first derive cell–cell or gene–gene similarity networks, bionSBM directly represents each modality as a cell–feature bipartite layer, in which each non-zero molecular measurement defines a weighted edge between a cell and the corresponding feature. These layers share the cell nodes and are jointly modelled as a single multipartite network. bionSBM can cluster cells based on any modality a user deems relevant (peaks, ADTs, messenger RNAs (mRNAs), the complete transcriptome, etc.), treating them symmetrically *a priori*, meaning no modality-specific weights or scaling coefficients are introduced. It does not impose a unimodal prior; instead, it assumes a uniform prior to fit the topics-cells and features-topics distributions. In [21], it was shown that a uniform prior can yield better results even with non-uniform synthetic latent variables. Moreover, in [19], it was shown that the prior assumptions of different algorithms can affect the inferred results; therefore, a uniform assumption is the one with as little information as possible. This algorithm is part of a long-standing effort to apply SBM methods to biomedical data, which started with the construction of a hierarchical version of SBM-based topic modelling (hSBM) [18] for gene expression data, then expanded to bulk tissue multi-omics data (nSBM) [22], to single cell gene expression data [23] and finally, in the present paper, to single cell multi-omics data. In the following text, we first describe the main features of bionSBM, then compare it with the topic modelling integration algorithms ShareTopic [24] and Mowgli [25], as well as COBOLT [26], MultiVI [27], and totalVI [28].

Results

The bionSBM model

bionSBM leverages a class of algorithms known as SBMs [17]. In particular, bionSBM is based on a remarkably efficient hierarchical version of SBM, developed in [19, 20], known as hierarchical stochastic block modelling (hSBM), which was later extended to n-partite SBM (nSBM) for multipartite graphs.

The idea behind hSBM is to maximize the posterior probability of the model given the data, namely $P(model \mid data)$. In a Bayesian framework, this is done by assuming a model $P(model)$ and maximizing the likelihood $P(data \mid model)$. hSBM assumes a uniform prior on $P(model)$ and performs statistical model selection based on the principle of minimum description length (DL) (Methods). In general terms, this class of algorithms seeks the most compact possible representation of the data.

In the context of single-cell omics data, the observations correspond to molecular measurements, such as single-cell RNA sequencing (scRNA-seq) or single-cell ATAC-sequencing (scATAC-seq) matrices [Fig. 1A], and the model is a hierarchy of partitions of both molecular features and cells. In previous work, hSBM was extended to nSBM [22] to accommodate a data structure with multiple sources. Now, the bionSBM algorithm is efficiently implemented to handle a multipartite network built from single-cell data, where nodes represent different entities, such as genes, peaks, or proteins, and cells [Fig. 1B]. Genes and peaks are connected to cells based on their molecular measurement levels, using weighted edges. bionSBM allows multiple feature types to be associated with the same data point (i.e. cell) [Fig. 1B]. This enables the concurrent modelling of multiple modalities in a single inference run [Fig. 1B, C]. bionSBM is integrated into the single-cell Python universe [29] by handling *AnnData* [30] for single-modality data and *Muon* [31] objects for multi-modal data.

In addition, bionSBM distinguishes itself from nSBM in terms of time and scalability by efficiently creating large sparse graphs, which are characteristic of single-cell experiments [Supplementary Fig. S1 A, B]. A key advantage of bionSBM over other algorithms is that, because it operates on graphs, preprocessing does not require scaling or harmonizing values across omics data. Moreover, features with heterogeneous statistical characteristics (counts, continuous values, or binary measurements) originating from different omics

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

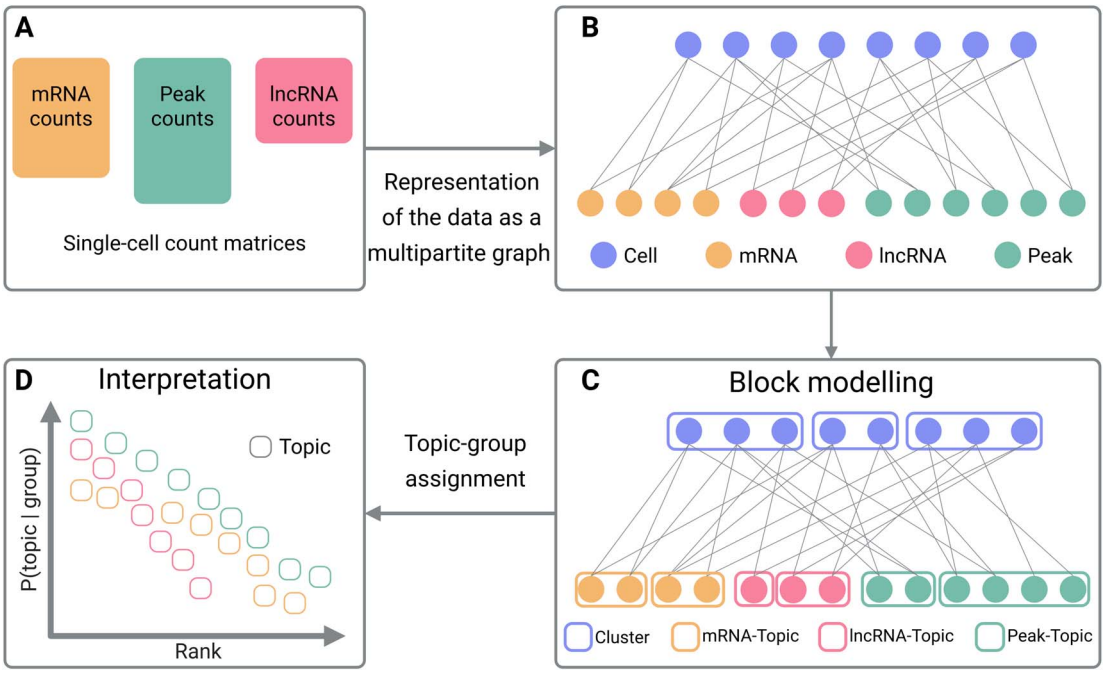

Figure 1 Schematic pipeline of the bionSBM algorithm. **A** Data are input as count matrices. **B** Input matrices are used to build a multipartite network: features are connected to the cell branch if the feature is expressed/open in the cell with weighted edges, but features are not connected. **C** bionSBM finds the best block model to group the cells into clusters and the features into topics. **D** The explanation for the cell grouping can then be easily obtained from the topic distribution probabilities within each cluster. Figure created with BioRender.

layers and preprocessing pipelines can be incorporated directly into the multipartite graph, as the inference procedure does not require cross-modality scaling or harmonization. The output of bionSBM is a hierarchical clustering of cells, while providing modality-specific topic assignments [Fig. 1D]. In other words, the terms 'clusters" and 'topics" assume a different meaning than in LDA-based approaches: here, a cluster is a set of cells, like in a community detection method, but a topic is a set of features. This allows the creation of groups of cells (clusters) [Fig. 1C], groups of peaks (peaks-topics) [Fig. 1C], groups of genes (gene-topics) [Fig. 1C], or any other kind of feature, for the same multipartite graph. This richer representation allows the disentangling of distinct feature sets across modalities. It is important to note that bionSBM requires that nodes of a given type be non-mixed, so that blocks contain only homogeneous nodes; hence, each group will contain either cells, genes, peaks, etc.

Model benchmarking

We compared the performance and results of bionSBM with those of ShareTopic [24], Mowgli [25], COBOLT [26], MultiVI [27], and totalVI [28] on six different data sets. We chose them to represent a variety of situations, including mouse and human genomes, from fully developed to differentiating cells. The data sets show a heterogeneous complexity, ranging from a few thousand to 40 000 cells and from about 10 to 35 cell types. The number of genes per experiment is one of the most homogeneous characteristics across data sets, because the species fix it [Supplementary Fig. S2]. Four of the data sets considered (PBMC from the 10x database [32], BMMCMultiOme [33], HSPC [34], MouseSkin [4]) consist of gene expression and chromatin openness data (scRNA-seq and scATAC-seq). Among these, the MouseSkin data set has been profiled using the SHARE-seq technology [4], whereas the other three are based on the 10X Multiome platform. Two additional data sets, Spleen [28] and BMMCCite [35], are CITE-seq experiments that combine gene expression and surface protein data. This ensemble of diverse data was chosen to test the efficacy of the methods across different contexts, ensuring that performance is not driven by technological, species-, or omics-level biases (see Supplementary Table S1, Methods).

For benchmarking, ShareTopic can handle only scATAC-seq and scRNA-seq pairs; therefore, the two Cite-seq data sets are excluded from this method. For the other four data sets, we could compare all the algorithms. For each data set, we explored different options for inputting transcriptional features. In a previous study [23], we showed that cell classification improves when analysing the expression of protein-coding genes and long non-coding RNAs (lncRNAs) as separate omics layers using the aforementioned nSBM approach. Hence, we created three transcriptomic count matrices: the total Gene Expression (GEX), which contains all the possible transcripts measured by an scRNA-seq experiment; the 'mRNA' matrix, containing only protein-coding genes; and the one named 'lncRNAs', which consists only of lncRNAs.

We tested the quality of the results on the cell layer separately (to evaluate the algorithm's ability to recover cell types) and on the feature layers (to assess topic quality across different feature spaces). For each test setting, we ran the algorithms multiple times to explore the stochasticity of the results and the impact of various combinations of transcriptional features. All the algorithms were executed 25 times per data set and feature-space combination, and we measured runtime and memory usage to provide an accurate report of the required resources [Supplementary Fig. S1 C, D].

All the methods embed the input spaces into a low-dimensional representation. Mowgli and ShareTopic create a single low-dimensional topic space, where each topic is a mixture of features across all omics, making interpretation more laborious because the algorithm mixes

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

features from different omics; the same applies to COBOLT factors. Instead, bionSBM treats each -omic separately in its own output. Therefore, when using the term 'topic' for bionSBM, we will specify which -omic we refer to (mRNA-topic, peaks-topic, etc.), while for ShareTopic, Mowgli, and COBOLT, the topics will be a mixture of these. MultiVI and totalVI do not create topics or factors.

bionSBM better recapitulates cell type in all the designed experiments and across technologies

We first evaluated the performance of the algorithms in recovering different cell types across the data sets. All the methods require the number of desired topics (or the latent dimension of the number of factors) as input. In contrast, bionSBM automatically infers the optimal granularity of each level of the hierarchy. The same applies to the number of clusters. In this work, for comparability, this hyperparameter was set to match the number of annotated cell types, thereby constraining the model to the expected biological heterogeneity. Notably, varying the number of topics/factors/latent dimensions of the comparator methods had little impact on performance [Supplementary Fig. S3].

To measure agreement between the clusters obtained with the algorithms and the ground-truth annotations across the six data sets, we show the NMI/NMI∗ (Methods). The larger the NMI/NMI∗ ratio, the better the agreement. Since this measure compares the clustering results to a random assignment of cells to clusters, it is affected by the characteristics of the specific data set, and the results should not be used to compare data sets (for instance, the number of cell types varies widely across data sets). Since bionSBM builds a hierarchy of partitions with different granularities, we suggest the user focus mainly on the one with the highest concordance with the known annotation, which is the one we report in Fig. 2. To make the effect of this *post hoc* selection explicit, Supplementary Fig. S4 reports NMI/NMI∗ across all inferred hierarchy levels for every dataset and experiment. The results show that high performance is often maintained across adjacent levels, highlighting that the choice of the level does not significantly affect the results. bionSBM shows performance similar to or better than the other methods [Fig. 2]. The PBMC data set is the one in which bionSBM stands out more significantly [Fig. 2A], whereas its performance on HSPC, BMMCMultiOme, and MouseSkin is less pronounced [Fig. 2B, C, D]. On the other hand, bionSBM achieves very high NMI/NMI∗ values on the CITE-seq dataset, even better than totalVI, which is tailored to these specific data sets [Fig. 2 E, F]. It is interesting to notice that the Cite-seq data also have the highest number of cell types, reflecting superior performance on more complex data.

Graph-based topic modelling signatures are highly specific and clearly distinct

An important feature of algorithms in the topic modelling framework is their probabilistic output, encoded in the two probability distributions $P(topic \mid cell)$ and $P(feature \mid topic)$, where feature is the biological feature (mRNA, lncRNA, peak, ADT) and topic is the appropriate topic family (mRNA-topic, etc). It is then possible to associate to each group of cells its most important topic simply using $t^*(group) = \arg\max_t P_c(topic \mid group)$, where P_c is a corrected version of the $P(topic \mid cell)$ (Methods). Moreover, one can assess the quality of this topic assignment using two standard metrics: specificity and distinctiveness. The former metric measures how group-specific the dominant topic is (i.e. the topic with the highest probability), while the latter captures how well topics are separated by quantifying the in-group and out-of-group topic probabilities (Methods). Both metrics take values in the range [0, 1], with 1 representing the maximum performance. We will assess them, excluding the AE-based methods, since they do not produce factors or topics. As mentioned above, bionSBM embeds each feature (mRNAs, lncRNAs, scATAC-seq peaks, etc.) in separate spaces, whereas Mowgli, ShareTopic, and COBOLT create a set of topics that mix features. Hence, the specificity and distinctiveness of bionSBM are computed for each -omic space, while for the other methods, the topic space is mixed.

bionSBM generally achieves high topic specificity [Fig. 3], higher than the other methods. The specificity is high across all data sets and modalities, indicating that the method is robustly able to explain each cell type in each modality. In agreement with the classification results [Fig. 2], bionSBM achieves a particularly high specificity in the BMMCCite and Spleen data sets [Fig. 3E, F], which are the most complex ones in terms of cell types [Supplementary Table S1]. The algorithms achieve similar values across all data sets for their distinctiveness metric [Fig. 4], with good performance in general, around 0.75.

Together, these two metrics show that bionSBM's signatures are highly specific and exhibit strong in-group enrichment relative to out-of-group probabilities; thus, they can be used to assign biological explanations to cell-type groups, making the method easy to interpret and translational. All the methods achieve similar levels of distinctiveness, meaning that all topic models analyzed assign distinct topics per cell type, but bionSBM outperforms the others in specificity, assigning topics to clusters more precisely.

bionSBM retrieves meaningful chromatin-transcript associations

To assess whether the model learns meaningful chromatin-gene associations, we examined their interplay within and across topics. We quantified the correlation [36] between peak and mRNA topics, as well as the enrichment of transcription factor (TF) motifs and their targets across both modalities, using different approaches (Methods).

Distance correlation (dCor) was computed between mRNA and peak topics to quantify their statistical dependence. For each pair of mRNA and peak topics (t, u), dCor was evaluated using the two corresponding topic vectors across cells, capturing arbitrary dependencies between modalities. This results in a matrix of mRNA–peak topic correlations. The correlation was compared with a null model obtained by independently permuting topic values across cells, thereby preserving marginal distributions (Methods). Final results are reported as Z-scores comparing the observed and null mean dCors (see the Methods section). All data sets show consistent positive Z-scores [Fig. 5A], indicating that the inferred topics encode shared cross-modal structure across cells. This supports the interpretation that chromatin accessibility and transcriptional programs are coordinated at the level of latent topics.

To further explore the chromatin-transcriptome interplay, we tested whether mRNA topics are enriched for genes targeted by TFs within peak topics (identified via motif analysis with HOMER [37]). We quantified, for each mRNA–peak topic pair, the enrichment of TF targets within the mRNA-topic gene sets relative to a null model in which the peak-associated TFs were kept fixed and the mRNA gene sets were randomized while matching their size, mean expression, detection

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

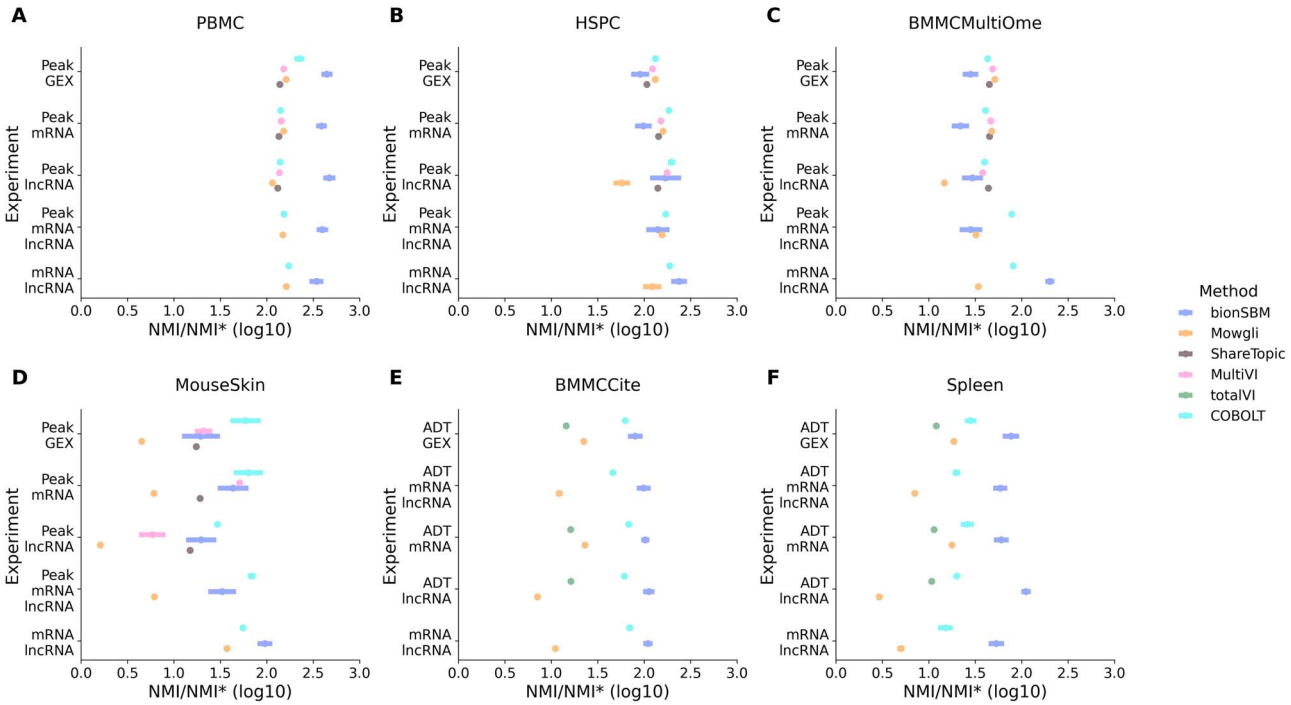

Figure 2 **A–F** NMI/NMI∗ score between each algorithm partition and ground truth annotation for the six data sets. Each dot represents the mean value, and the error bar represents three times the uncertainty on the mean. The higher the values, the closer the clustering is to the ground truth annotations.

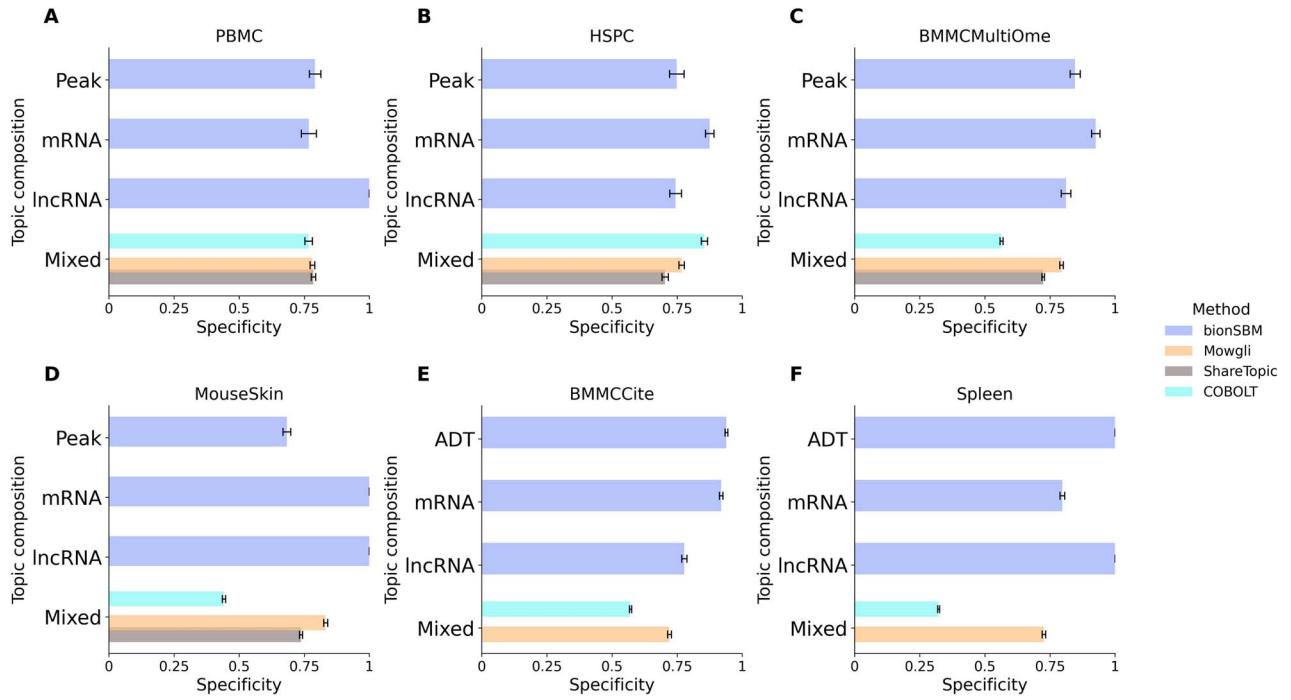

Figure 3 **A–F** specificity, defined as the fraction of cell groups assigned distinct dominant topics, for the six data sets. Each bar represents the mean value over multiple runs, and the error bar represents three times the uncertainty on the mean.

coverage, and genomic length (Methods). We obtained an enrichment score matrix for each topic pair and selected the most enriched pairs, assuming that each peak-topic targets primarily one mRNA-topic. We report the results as Z-scores relative to the null model (Methods). The PBMC, HSPC, and BMMCMultiOme data sets show very strong enrichment, whereas the MouseSkin one shows a milder one [Fig. 5B]. Overall, this outcome suggests that bionSBM identifies latent, coordinated ATAC and GEX topics whose gene composition is consistent with known transcriptional regulatory programs.

Finally, for each mRNA topic, we quantified the overlap between its gene set and the set of known TF interactors. We summed, for all TFs present in the topic, the number of shared genes between the

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

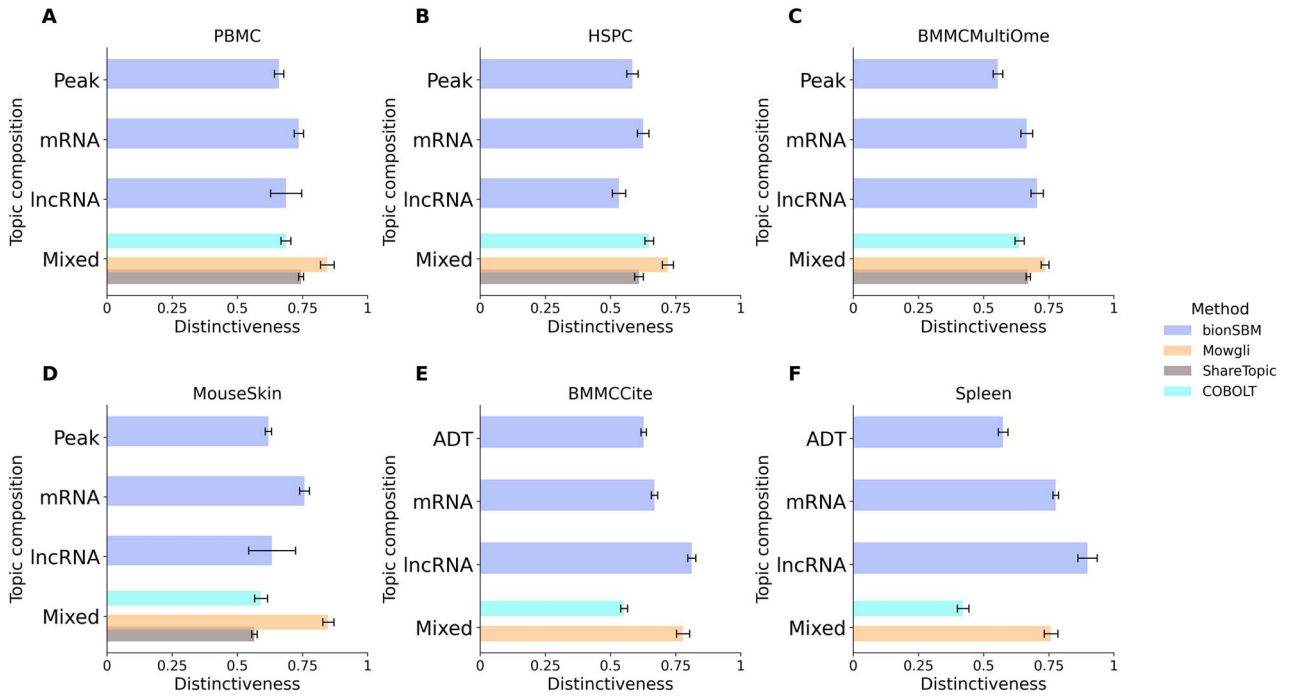

Figure 4 **A**–**F** distinctiveness (a measure of how well topics are separated by quantifying the in-cluster topic probability and the out-of-cluster probability) of the topics for the six data sets. Each bar represents the mean value over multiple runs, and the error bar represents three times the uncertainty on the mean.

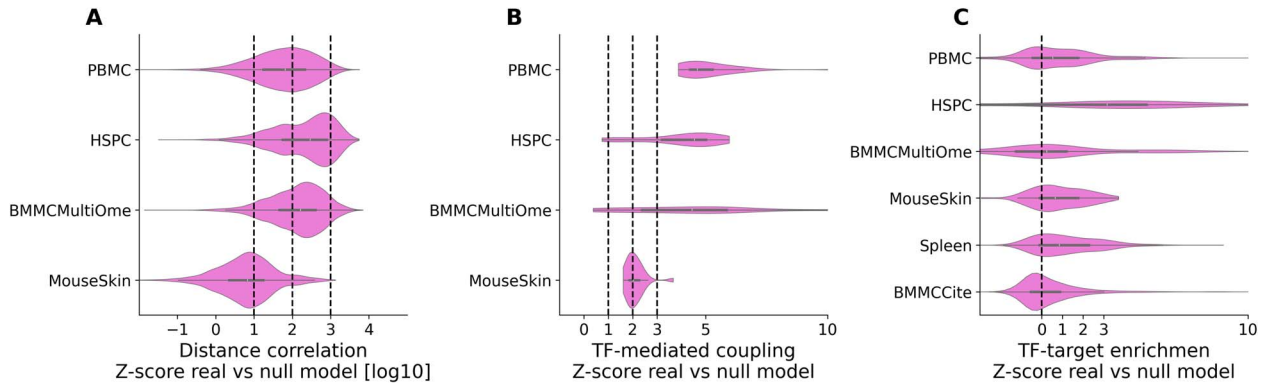

Figure 5 Distance correlation **A**, TF-coupling **B**, and TF-target enrichment **C** for the peak/ADT–mRNA–lncRNA experiment, reported as Z-score between real and permutation-based null model.

topic genes and each TF's target gene set. We compared this aggregate overlap score against a null distribution obtained from randomly sampled and bias-corrected gene sets as described above (see Methods). All data sets show mildly good levels of TF-enrichment [Fig. 5C], indicating that bionSBM tends to identify mRNA topics structured by known transcriptional regulatory programs. We further investigated the contribution of each modality to the results in Fig. 5 by running the same analysis using combinations of the modalities, in line with the approach used to evaluate cluster NMI/NMI∗ and topic specificity and distinctiveness. The results show an interplay between chromatin and transcriptome across all the combinations of modalities. However, the lncRNA and the GEX, along with the peaks, tend to show the lowest distribution of dCor, TF-coupling, and TF-enrichment, supporting the argument that it is important to separate different gene types within RNA data to extract the maximum structure from the data [Supplementary Fig. S5 A, B, C]. These results are also consistent with the fold-change metric between the observed values and null models, without normalizing by variance, which shows the magnitude of the observed effect independently of the variance of the null model [Supplementary Fig. S5 D, E, F].

To exemplify the TF–target gene relationship between mRNA and peak topics, we assigned a topic to each cell type using the following criterion: for mRNA and peak modalities, we computed $P_c(topic \mid cell\,type)$ and assigned to each cell type the topic with the highest distinctiveness (Methods). We then examined the top-ranked genes by $P(gene \mid topic)$ in the selected mRNA topic and compared them with TF binding motifs enriched in the corresponding peak topic for the same cell type obtained with HOMER [37]. By linking chromatin topics to transcriptome topics, we could extract meaningful case studies that substantiate the interpretability of bionSBM.

This approach recovered well-established regulatory programs across data sets. In PBMCs, the peak-topic with the highest

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

distinctiveness for the B memory cell population was enriched for the PAX5 motif [Fig. 6A], an essential gene required for B cell commitment from lymphoid progenitors [38]. Coherently, the same topic was also the highest enriched in B naive cells [Supplementary Fig. S6 A], and the same gene was found to be in the most enriched mRNA-topic in both cell states [Fig. 6A, Supplementary Fig. S6 A]. Notably, the topics were highly enriched for both mRNA and ATAC features only in B cell populations [Supplementary Fig. S7 A], highlighting the specificity of our approach. In the HSPC data set, granulocyte-monocyte progenitors (GMPs) showed the strongest association with a peak-topic motif for IRF8 [Fig. 6B], a TF known to regulate GMP differentiation [39]. The corresponding highest-associated mRNA-topic for the same cell type contained the same TF [Fig. 6B], again connecting the two feature spaces mechanistically. Lastly, in the BMMCMultiOme data set, the proerythroblasts, normoblasts, and erythroblasts had a high probability of a peak-topic showing the KLF1 motif [Fig. 6C, Supplementary Fig. S6 B, Supplementary Fig. S6 C], a master regulator of erythropoiesis [40], and the same gene belonged to the most relevant mRNA-topic for the three above-mentioned cell types, which are the primary cell states evolving during erythropoiesis [Fig. 6C, Supplementary Fig. S6 B, Supplementary Fig. S6 C]. Again, the selected mRNA- and peak-topics were not only the ones with the highest distinctiveness among all other topics, but they were also highly enriched only for the cell types under examination [Supplementary Fig. S7], giving evidence that the explanations are specific. We next evaluated the association of the selected mRNA and peak topics with the bionSBM clusters using $P_c(\text{topic} \mid \text{cluster})$. The cluster-enrichment step is label-independent but is applied to topics previously selected using the annotated cell types [Fig. 6]. Specifically, for each selected topic, we identified the bionSBM clusters showing the strongest enrichment without using cell-type labels. We then examined the cell-type composition of these clusters as a *post hoc* validation step. For the selected topics, the most strongly enriched clusters were predominantly composed of the expected cell types, supporting the biological consistency between the inferred cell partitions and the regulatory programs captured by the selected topics [Supplementary Fig. S8].

Overall, the discovered peak and mRNA assignments consistently recovered canonical lineage-defining TFs, supporting the bionSBM integrative topic model's ability to capture biologically meaningful, cell-type-specific regulatory programs across modalities and to link the transcriptome to the epigenome.

Discussion

In this study, we present bionSBM, a graph-based SBM to integrate single-cell multi-omic data [Fig. 1]. This approach has both theoretical and practical advantages over the state-of-the-art topic models for the single-cell field. First, our method can be applied to any multi-modal data, since it handles an arbitrary number of partitions (i.e. modalities). Second, the method automatically detects the optimal number of topics and clusters without requiring these values to be specified as hyperparameters. Moreover, bionSBM treats modalities symmetrically *a priori* and does not introduce modality-specific weighting parameters. Finally, because topic modelling algorithms are Bayesian, they require specification of a prior distribution over topics. Rather than imposing a structured distribution that constrains cells and features to follow a single mixture model, we employ a flat (non-informative) prior.

Integrating multiple omics layers is key to uncovering new biological insights. By modelling separate topic compositions for each network layer, bionSBM captures coherent relationships across modalities, thereby inferring modality-specific topic structures within a unified probabilistic framework and enabling independent yet coherent biological interpretations across omics layers.

bionSBM also improves performance in cell-type retrieval, outperforming state-of-the-art approaches and more accurately recovering known biological annotations [Fig. 2]. These performance gains come along with inherent interpretability. The model generates specific [Fig. 3], distinct [Fig. 4], and stable [Supplementary Fig. S9, Supplementary Fig. S10] topic representations. These representations support cross-modal regulatory analyses [Fig. 5, Supplementary Fig. S5] and provide biologically meaningful explanations of annotated cell groups and clusters [Fig. 6, Supplementary Fig. S6, Supplementary Fig. S7], Supplementary Fig. S8. This balance between accuracy and interpretability is essential, as predictive improvements alone are unlikely to drive meaningful biological discovery.

We released a Python implementation of the tool, integrated into the single-cell universe, leveraging the convenient *AnnData* and *Muon* packages to guarantee reproducibility and broad usability of the method. Here, we have applied bionSBM to single-cell multi-omic measurements with scRNA-seq, paired with either scATAC-seq or surface protein measurements; however, the framework can be used for any new omics measurements that can be summarized as a count matrix, making the approach truly generalizable.

Methods

Single-cell data processing

The raw reads of the PBMC [32] data set were processed using *Cell Ranger ARC 2.0.2* [41], aligning them to the complete human genome (T2T) [42]. The gene expression count matrices for all other data sets (BMMCMultiOme [33], BMMCCite [35], HSPC [34], MouseSkin [4], Spleen [28]) were obtained from the literature, along with the cell type annotations. From the total GEX count matrices, which contain all the families of transcripts that scRNA-seq can profile, we extracted the mRNA and lncRNA counts using the genome annotations from the appropriate GTF file for the data set (see Supplementary Table S1 for the files). All matrices underwent the same preprocessing according to standard single-cell best practices [6] using *Scanpy* [43]: quality control, doublet identification [44], filtering, library-size normalization, and log transformation. The PBMC peak count matrix was built with Episcanpy [45] after MACS2 [46] peak calling. All the other peak count matrices were obtained from the literature. Similar to scRNA-seq data, preprocessing for scATAC-seq followed best practices, including quality control, filtering, and library size normalization. In both scRNA-seq and scATAC-seq matrices, we kept only the highly variable features. The ADT (surface protein) count matrices were library-size normalized only because the number of features is minimal, so no additional filters are needed.

Stochastic block models

The statistical inference procedure and the definitions of topics and probability distributions are based on the hSBM algorithm. The input matrices are transformed into a weighted multipartite network. Each node represents a feature or a cell (i.e. mRNA, lncRNA, peak ADT). To build the network, we use the log-counts of the data with standard

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

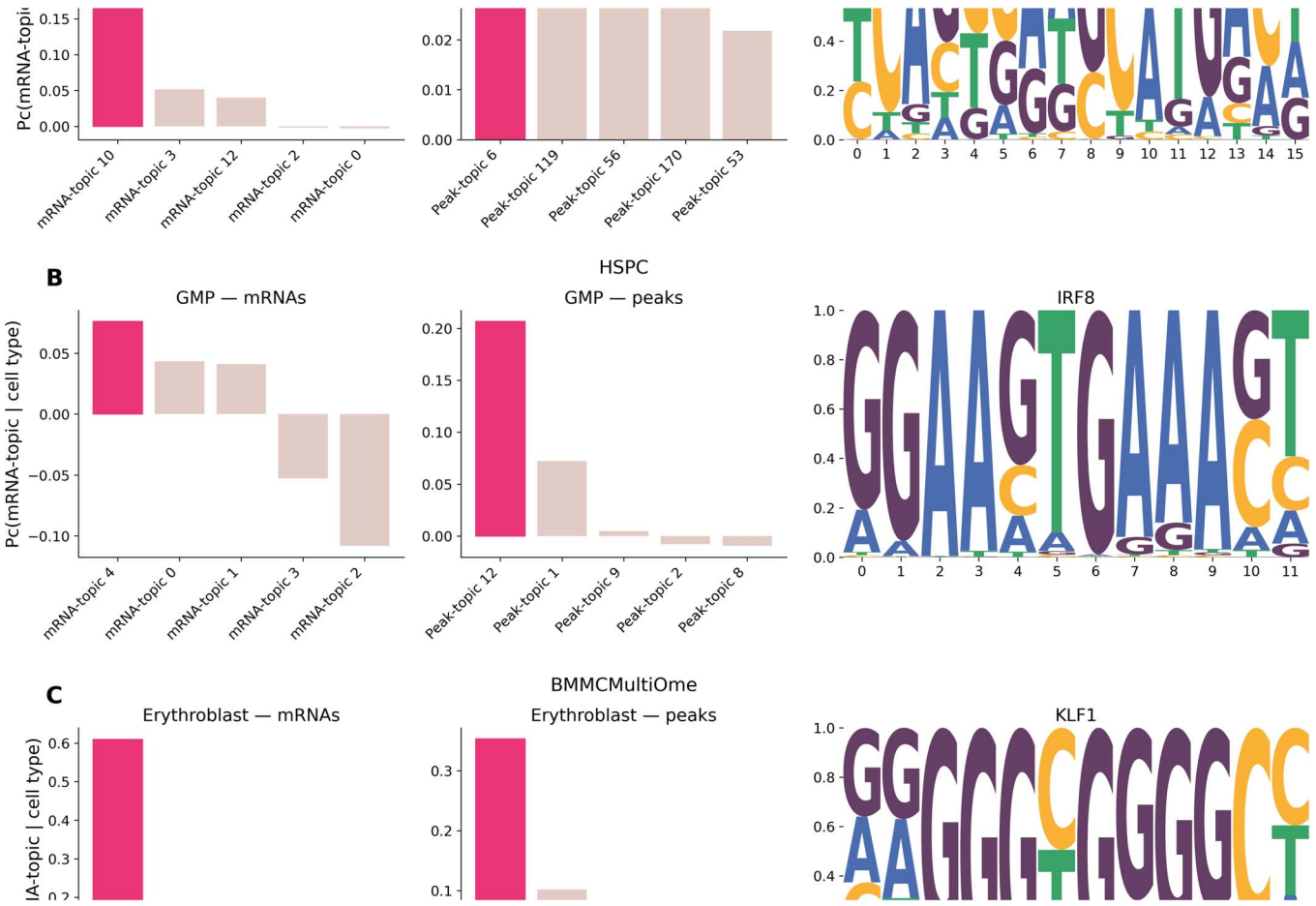

Figure 6 We present a row for each dataset. **Left**: enrichment profile of the mRNA-topic with the highest distinctiveness for the given cell type. **Middle**: enrichment profile of the peak-topic with the highest distinctiveness for the given cell type. **Right**: sequence logo of a significantly enriched TF motif identified in peaks belonging to the selected peak-topic (HOMER analysis).

preprocessing, as discussed in the previous paragraph. Each cell is connected to a feature, in one of the network partitions, if that feature is represented by a non-zero entry in the count matrix; the edge is weighted by the count itself. In the comparison of bionSBM to other algorithms, focusing on the differences, advantages, and drawbacks among algorithms, we used this single strategy to build the bionSBM network, rather than optimizing preprocessing for bionSBM. The algorithm defines $P(model \mid data)$, where the data are the graph and the model is the particular partition constructed during inference. The inference process begins with nodes (i.e. genes or cells) assigned to random blocks (either clusters or topics). The process consists of a Markov Chain Monte Carlo (MCMC) simulation to adjust node assignments to blocks, thereby minimizing the DL [47], defined in natural units as the number of bits required to encode the model and the data. Moves in the MCMC are accepted with probability [21]:

$$P(r \to s|t) = \frac{e_{ts} + \epsilon}{e_t + \epsilon B}, \tag{1}$$

which represents the probability of moving a node from block (cluster or topic) r to block s if a random neighbour of the considered node belongs to block t; given that B is the total number of blocks, e_{ts} is the number of edges from group t to group s and e_t is the total number of edges within group t. Such probability is proportional to the ratio between the number of edges that connect group t and block s and the number of edges that go to block t to group s and e_t is the total number of edges within group t. Such probability is proportional to the ratio between the number of edges that connect group t and block s and the number of edges that go to block s, plus a term that is necessary in order to have a uniform probability if there are no connections between the blocks considered. For instance, if all the edges from t go to group s, e_{ts} equals e_t and the probability becomes uniform: $1/B$. The algorithm supports multiple random initializations, i.e. different starting points for estimating the posterior probability.

It is essential to highlight that, given the probability of accepting a move, the advantage of using multibranch SBM on multi-omics problems is clear. When moving a cell from block r to block s, the algorithm samples a neighbour of the cell and it labels its group as t. At this point, the model only considers the links between the cell and the nodes in the group, so the normalizations of other branches are irrelevant. The final output of the model consists of a hierarchy of topics and clusters. For each level in the hierarchy, the model returns the grouping of features into topics, the grouping of cells into clusters, and the probability that each cell belongs to each topic [18, 19, 22]. When the data set annotation is not available, we suggest that the user excludes partitions where the number of clusters or topics is either one or is disproportionately high with respect to the expected

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

heterogeneity of the data, and investigates the heterogeneity of the data using all remaining levels to perform a multi-scale analysis.

A first implementation of multipartite hSBM, the so-called nSBM, is available at https://github.com/BioPhys-Turin/nsbm [22] based on the graph-tool library [48].

Experiment running and metrics

bionSBM has been run 25 times for each combination of feature spaces, with the initial seed changed each time and the number of initializations set to 7 as suggested in [23]. The ShareTopic, Mowgli, COBOLT, MultiVI, and totalVI experiments have also been run 25 times. MultiVI and totalVI were run using the *scvi-tools* framework [49]. All the algorithms have been run with default parameters, with the number of topics (or factors, or latent dimensions) equal to the number of cell types annotated in each data set. In Mowgli, we set the regularization strength to 0.05 across all feature spaces to treat them equally; 0.05 comes from the average suggested weight in their code between RNA and ATAC data. For each experiment outcome and each level of the bionSBM hierarchy, we have measured the NMI/NMI∗ between the algorithms' clustering and the original cell-type annotation.

The NMI/NMI∗ is a version of the normalized mutual information (NMI) [50] that accounts for randomness, with NMI∗ quantifying the agreement between the ground-truth label and a random clustering of cells. In this way, we account for partitions with a large number of clusters that would otherwise yield a falsely high NMI. On a log scale, the lower limit is 0, indicating that the clustering is entirely dominated by randomness; i.e. the NMI equals NMI∗. For the comparator methods, the number of topics/factors/latent dimensions was set to the number of annotated cell types for benchmarking. bionSBM inference itself does not use the annotations. After inference, for each run and experiment, we used the annotations only to identify the hierarchy level with the highest NMI/NMI∗. The resulting scores were then averaged across runs for each method–feature-space combination to produce Fig. 2.

Topics specificity, distinctiveness, and stability

To compute the specificity, distinctiveness, and stability we selected, for each experiment and for each run, the level with the closest number of clusters to the ground truth annotation. After setting the level, we assigned to each group of cells the topic with the highest probability. Let t and i be a topic and a cell, respectively, we used a modified version of the $P(t|i)$ defined as $P_c(t|i)$ (where c stands for centred) to exclude topics that are important for all the cells, which would be the equivalent of the topic of articles and prepositions in NLP, as suggested in [18]. The centred probability is defined as

$$P_c(t \mid i) = P(t \mid i) - \frac{1}{|\mathcal{C}|}\sum_{j\in\mathcal{C}} P(t \mid j), \tag{2}$$

where $\mathcal{C}$ denotes the set of all cells. Let $\mathcal{C}_\tau$ be the set of cells belonging to a particular group τ, we can derive the $P_c(t \mid \tau)$ as

$$P_c(t \mid \tau) = \frac{1}{|\mathcal{C}_\tau|}\sum_{i\in\mathcal{C}_\tau} P_c(t \mid i) \tag{3}$$

We assigned a topic to each cell group according to $t^*_\tau = \arg\max_t P_c(t \mid \tau)$. We then counted the number of distinct topics assigned to the cell groups and divided that number by the total number of cell groups: let N_{ct} be the total number of cell types and let N_{assigned} be the number of distinct dominant topics assigned across the N_{ct} cell types, the specificity is then defined as

$$\text{Specificity} = \frac{N_{\text{assigned}}}{N_{\text{ct}}} \tag{4}$$

We select the hierarchy level in bionSBM with the closest number of topics to Mowgli and ShareTopic. We average the values obtained from the tool runs to produce the bar plot in Fig. 3.

Once we marked the most important topic for each group as described above, we quantified the distinctiveness of these assigned topics, which is defined as

$$\text{Distinctiveness}(t^*_\tau) = \frac{1}{|\mathcal{C}_\tau|}\sum_{i\in\mathcal{C}_\tau} P_c(t^*_\tau \mid i) - \frac{1}{|\mathcal{C}\setminus\mathcal{C}_\tau|}\sum_{i\in\mathcal{C}\setminus\mathcal{C}_\tau} P_c(t^*_\tau \mid i), \tag{5}$$

where t^*_τ is the topic assigned to group $\mathcal{C}_\tau$, and $P_c(t|i)$ is defined as in 2. In other words, it is defined as the difference between the average probability that the topic belongs to the cells within the group and the average probability that it belongs to cells outside the group. For distinctiveness, we average the values across multiple model runs to produce Fig. 4.

To quantify the stability of topic–feature associations across independent runs of the same experiment, we measured the similarity of the top three most important topics identified for each cell type. For a given run r, let $P_{c,r}(t \mid i)$ denote the centred posterior probability of topic t for cell i, as obtained from the topic–document matrix as shown in Eq 2.

For each cell type τ, a cell-type-specific topic score was computed by averaging centred probabilities across cells of that type:

$$P_{c,r}(t \mid \tau) = \frac{1}{|\mathcal{C}_\tau|}\sum_{i\in\mathcal{C}_\tau} P_{c,r}(t \mid i), \tag{6}$$

where $\mathcal{C}_\tau \subset \mathcal{C}$ is the subset of cells annotated as cell type τ. For each run and cell type, the set $\mathcal{T}_r^\tau$ of the top $N_{\text{top}} = 3$ topics with the highest values of $P_{c,r}(t \mid \tau)$ was selected. Each topic $t \in \mathcal{T}_r^\tau$ is associated with a topic–feature probability vector $\boldsymbol{\phi}_{r,t}$.

To compare two runs r and r', we computed pairwise similarities between all topic vectors in $\mathcal{T}_r^\tau \times \mathcal{T}_{r'}^\tau$ using the cosine similarity

$$S_{\cos}(p, q) = \frac{p \cdot q}{\|p\|\,\|q\|}, \tag{7}$$

where p and q are normalized topics–features distributions.

For each cell type and run pair, the similarity was defined as the maximum similarity across possible combinations given the three top topics for run pairs:

$$S^\tau_{\text{stab}}(r, r') = \max_{\substack{t\in\mathcal{T}_r^\tau \\ u\in\mathcal{T}_{r'}^\tau}} S_{\cos}\left(\boldsymbol{\phi}_{r,t}, \boldsymbol{\phi}_{r',u}\right) \tag{8}$$

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

Choosing the three top topics makes the metric robust to slight variations in the top rankings, which arise from data noise and the approximation introduced by the sampling procedure. The final stability score for an experiment was obtained by averaging $S^{\tau}_{\text{stab}}(r, r')$ across all cell types and all unordered pairs of runs.

Topics analysis and interpretation

For biological interpretation, we used the reference cell-type annotations to identify topics with the highest mean loading in each annotated population. We then assessed the motif and regulon content of these topics. This analysis evaluates concordance between inferred topics and established cell-type regulatory programs; it is not intended as a label-independent discovery analysis. For the TF-based analyses, random mRNA gene sets were matched to the observed topic gene sets in size, mean expression, detection coverage, and genomic gene length. Mean expression and coverage were computed from the expression matrices, with coverage defined as the fraction of cells with non-zero expression. Gene lengths were obtained from the appropriate GTF files (Supplementary Table S1). To account for these covariates, we divided genes into up to five quantile bins for each variable (expression, coverage, and length) and grouped them by their combinations of bins. Random gene sets were then sampled without replacement, preserving the number of observed topic genes within each stratum. Observed topic genes were excluded from the sampling pool whenever sufficient alternative genes were available. No random peak gene sets were constructed.

To quantify statistical dependence between RNA and chromatin latent representations, we consider two matrices $X \in \mathbb{R}^{N\times K_x}$ and $Y \in \mathbb{R}^{N\times K_y}$, where rows correspond to cells and columns to topics and each topic is mean-centred across cells. For each pair of topics (t, u), we compute the dCor, which captures arbitrary statistical dependence. Let $x_t \in \mathbb{R}^N$ and $y_u \in \mathbb{R}^N$ denote the tth and uth topic vectors across cells. We first construct pairwise distance matrices:

$$D^x_{ij} = |x^{(i)}_t - x^{(j)}_t| \quad \text{and} \quad D^y_{ij} = |y^{(i)}_u - y^{(j)}_u| \tag{9}$$

for all cell pairs (i, j). These matrices are then double-centred to remove mean effects:

$$A_{ij} = D^x_{ij} - \bar{D}^x_{i\cdot} - \bar{D}^x_{\cdot j} + \bar{D}^x_{\cdot\cdot} \tag{10}$$

$$B_{ij} = D^y_{ij} - \bar{D}^y_{i\cdot} - \bar{D}^y_{\cdot j} + \bar{D}^y_{\cdot\cdot}, \tag{11}$$

where $\bar{D}_{i\cdot}$, $\bar{D}_{\cdot j}$, and $\bar{D}_{\cdot\cdot}$ denote row, column, and global means, respectively. Distance covariance and variance are defined as

$$\text{dCov}(x_t, y_u) = \mathbb{E}[A_{ij}B_{ij}] \tag{12}$$

$$\text{dVar}(x_t) = \mathbb{E}[A^2_{ij}] \quad \text{and} \quad \text{dVar}(y_u) = \mathbb{E}[B^2_{ij}] \tag{13}$$

and dCor is given by

$$\text{dCor}(x_t, y_u) = \frac{\text{dCov}(x_t, y_u)}{\sqrt{\text{dVar}(x_t)\text{dVar}(y_u)}} \tag{14}$$

This procedure yields a dependence matrix $\mathbf{C} \in \mathbb{R}^{K_x\times K_y}$ with entries $C_{tu} = \text{dCor}(x_t, y_u)$. For computational efficiency, dCor is evaluated on a random subset of 3000 cells. We then constructed a null model by independently permuting each topic across cells:

$$x_t \to \pi^x_{t,m}(x_t), \quad y_u \to \pi^y_{u,m}(y_u), \tag{15}$$

which preserves marginal distributions. The same procedure is applied to the permuted data over 20 replicates, yielding a set of null modelling correlation matrices $\{\widehat{\mathbf{C}}^{(m)}\}^{20}_{m=1}$, which averages as μ_{tu} with SD σ_{tu}. We then defined a Z-score

$$z^{\text{dCor}}_{tu} = \frac{C_{tu} - \mu_{tu}}{\sigma_{tu}}, \tag{16}$$

which provides a normalized measure of how strongly RNA and chromatin topic structures are statistically coupled beyond chance. The same procedure is repeated for all 25 bionSBM runs. The resulting data are plotted in Fig. 5A.

To quantify the regulatory consistency between chromatin accessibility and gene expression topics inferred by bionSBM, we evaluated the enrichment of TF regulons within gene sets. Let $\mathbf{Z} \in \mathbb{R}^{|\mathcal{G}|\times K}$ denote the gene–topic matrix for GEX, where $\mathcal{G}$ is the set of genes and K the number of topics. For each topic t, we define the corresponding gene set $\mathcal{G}_t = \{g \in \mathcal{G} \mid Z_{gt} > 0\}$. For each ATAC topic a, we extract a set of TFs $\mathcal{F}_a$ based on significantly enriched motifs (Benjamini–Hochberg FDR [51] < 0.05 from HOMER analysis). Using a prior TF–target network (CollecTRI [52, 53]), where each TF f is associated with a regulon $R_f \subseteq \mathcal{G}$, we compute a coupling score:

$$S_{\text{coup}}(\mathcal{G}_t, \mathcal{F}_a) = \sum_{f\in\mathcal{F}_a} \frac{|\mathcal{G}_t \cap R_f|}{|R_f|} \tag{17}$$

This score measures the fraction of each TF regulon captured by the gene set, normalized by regulon size. The null model was computed over $N^{\text{coup}}_{\text{perm}} = 1000$ permutations. From this null distribution, we estimate the mean μ^{coup}_{ta} and standard deviation $\sigma^{\text{coup}}_{ta}$, and compute a Z-score:

$$z^{\text{coup}}_{ta} = \frac{S_{\text{coup}}(\mathcal{G}_t, \mathcal{F}_a) - \mu^{\text{coup}}_{ta}}{\sigma^{\text{coup}}_{ta}} \tag{18}$$

This procedure is repeated for all ATAC–GEX topic pairs across $N_{\text{runs}} = 25$ independent runs, yielding a matrix of regulatory coupling scores. For both observed and randomized data, coupling scores were calculated for all peak–mRNA topic pairs, ranked, and the maximum score retained. The observed maximum was then compared with the distribution of maximum scores obtained from the randomized data. The resulting data are plotted in Fig. 5B.

To assess whether the gene sets associated with each mRNA-topic capture biologically meaningful regulatory structure, we performed a TF enrichment analysis using curated regulons. For each topic t, where $\mathcal{G}$ denotes the set of all genes, we defined a global overlap score

$$S_{\text{enr}}(\mathcal{G}_t) = \sum_f |\mathcal{G}_t \cap R_f|, \tag{19}$$

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

where R_f is the target set of TF f. To evaluate the significance of this score, we constructed a null distribution by sampling random gene sets $\mathcal{G}_t^{(m)}$ of the same size, coverage, expression, and gene length as $\mathcal{G}_t$ from $\mathcal{G}$ and computing $S_{\text{enr}}(\mathcal{G}_t^{(m)})$ over $N_{\text{perm}}^{\text{enr}} = 1000$ permutations. The mean and standard deviation of the null distribution were estimated as

$$\mu_{t,\text{null}}^{\text{enr}} = \frac{1}{N_{\text{perm}}^{\text{enr}}} \sum_{m=1}^{N_{\text{perm}}^{\text{enr}}} S_{\text{enr}}(\mathcal{G}_t^{(m)}) \tag{20}$$

and

$$\sigma_{t,\text{null}}^{\text{enr}} = \sqrt{\frac{1}{N_{\text{perm}}^{\text{enr}} - 1} \sum_{m=1}^{N_{\text{perm}}^{\text{enr}}} \left[S_{\text{enr}}(\mathcal{G}_t^{(m)}) - \mu_{t,\text{null}}^{\text{enr}} \right]^2}. \tag{21}$$

We then defined the enrichment Z-score as

$$z_t^{\text{enr}} = \frac{S_{\text{enr}}(\mathcal{G}_t) - \mu_{t,\text{null}}^{\text{enr}}}{\sigma_{t,\text{null}}^{\text{enr}}}, \tag{22}$$

which quantifies the deviation of the observed TF–target overlap from the permutation-based null expectation in units of the null standard deviation. The resulting Z-scores are plotted in Fig. 5C.

For all three metrics described in this chapter, we also computed the ratio between the real and null model means, namely

$$R_M = \frac{\mu_{\text{real}}^M}{\mu_{\text{null}}^M}, \qquad M \in \{\text{dCor}, \text{coup}, \text{enr}\} \tag{23}$$

to produce Supplementary Fig. S5 D–F.

Key Points

- We introduce a probabilistic framework based on multipartite graphs for the joint analysis of single-cell multi-omic data.
- The model enables unified representation of heterogeneous modalities, identifying shared latent biological topics.
- It captures regulatory structure by linking genes, peaks, and surface proteins within a single generative framework.
- The approach shows strong interpretability and competitive performance compared with existing integration methods.
- The framework provides a strategy to uncover coordinated molecular programs in complex datasets.

Acknowledgements

We thank the BMC Bioinformatics Core Facility for providing access to their HPC cluster. Figure 1 was created with Biorender.

Author contributions

Gabriele Malagoli (Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing—original draft, Writing—review & editing), Filippo Valle (Conceptualization, Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing), Andreina Tirabassi (Conceptualization, Writing—original draft, Writing—review & editing), Annalisa Marsico (Conceptualization, Writing—original draft, Writing—review & editing), Loredana Martignetti (Conceptualization, Investigation, Formal analysis, Writing—original draft, Writing—review & editing), Michele Caselle (Conceptualization, Investigation, Formal analysis, Methodology, Supervision, Writing—original draft, Writing—review & editing), and Maria Colomé-Tatché (Conceptualization, Investigation, Formal analysis, Methodology, Supervision, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Competing interests

The authors declare no competing interests.

Funding

G.M. is supported by the Helmholtz International Lab Causal Cell Dynamics (InterLabs-0029)Grant support from the Initiative and Networking Fund of the Hermann von Helmholtz-Association Deutscher Forschungszentren e.V.-. G.M. is also supported by the Helmholtz Association under the joint research school Munich School for Data Science - MUDS and by the German Research Foundation project ID 213249687-SFB 1064. MCT was supported by a ATR2024-154840 grant financed by MICIU/AEI/10.13039/501100011033.

Data availability

Code and tutorial for bionSBM are available at https://github.com/gmalagol10/bionsbm.git. The code to reproduce all the results is available at https://github.com/gmalagol10/bionsbm/tree/main/reproducibility.

References

1. Ashley EA. Towards precision medicine. *Nat Rev Genet* 2016; **17**:507–22. https://doi.org/10.1038/nrg.2016.86
2. Van de Sande B, Lee JS, Mutasa-Gottgens E *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov* 2023; **22**:496–520. https://doi.org/10.1038/s41573-023-00688-4
3. Tirosh I, Izar B, Prakadan SM *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016; **352**:189–96. https://doi.org/10.1126/science.aad0501
4. Ma S, Zhang B, LaFave LM *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020; **183**:1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056
5. Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017; **14**:865–8. https://doi.org/10.1038/nmeth.4380
6. Heumos L, Schaar AC, Lance C *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023; **24**:550–72. https://doi.org/10.1038/s41576-023-00586-w

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

7. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986; **323**:533–6. https://doi.org/10.1038/323533a0
8. Athaya T, Ripan RC, Li X *et al.* Multimodal deep learning approaches for single-cell multi-omics data integration. *Brief Bioinform* 2023; **24**:bbad313. https://doi.org/10.1093/bib/bbad313
9. Yan X, Chen J, Zheng R *et al.* ACE: a versatile contrastive learning framework for single-cell mosaic integration. *Genomics Proteomics Bioinformatics* 2025; **23**:qzaf062. https://doi.org/10.1093/gpbjnl/qzaf062
10. Yan X, Fang Z, Ang KS *et al.* Mosaic integration of spatial multi-omics with SpaMosaic. *Nat Genet* 2026; **58**:1126–37. https://doi.org/10.1038/s41588-026-02573-3
11. Wenming W, Zhang W, Ma X. Network-based integrative analysis of single-cell transcriptomic and epigenomic data for cell types. *Brief Bioinform* 2022; **23**:bbab546. https://doi.org/10.1093/bib/bbab546
12. Ma X, Sun P, Zhang Z-Y. An integrative framework for protein interaction network and methylation data to discover epigenetic modules. *IEEE/ACM Trans Comput Biol Bioinform* 2019; **16**:1855–66. https://doi.org/10.1109/TCBB.2018.2831666
13. Ma X, Liu Z, Zhang Z *et al.* Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data. *BMC Bioinformatics* 2017; **18**:72. https://doi.org/10.1186/s12859-017-1490-6
14. Lazzardi S, Valle F, Mazzolini A *et al.* Emergent statistical laws in single-cell transcriptomic data. *Phys Rev E* 2023; **107**:044403. https://doi.org/10.1103/PhysRevE.107.044403
15. David MB, Ng AY, Jordan MI. Latent Dirichl*et al*location. *J Mach Learn Res* 2003; **3**:993–1022. https://doi.org/10.5555/944919.944937
16. Lee DD, Sebastian Seung H. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999; **401**:788–91. https://doi.org/10.1038/44565
17. Paul WH, Laskey KB, Leinhardt S. Stochastic blockmodels: first steps. *Soc Networks* 1983; **5**:109–37. https://doi.org/10.1016/0378-8733(83)90021-7
18. Valle F, Osella M, Caselle M. A topic modeling analysis of TCGA breast and lung cancer transcriptomic data. *Cancers* 2020; **12**. https://doi.org/10.3390/cancers12123799
19. Gerlach M, Peixoto TP, Altmann EG. A network approach to topic models. *Sci Adv* 2018; **4**:eaaq1360. https://doi.org/10.1126/sciadv.aaq1360
20. Hyland C, Tao Y, Azizi L *et al.* Multilayer networks for text analysis with multiple data types. *EPJ Data Sci* 2021; **10**. https://doi.org/10.1140/epjds/s13688-021-00288-5
21. Peixoto TP. Nonparametric Bayesian inference of the microcanonical stochastic block model. *Phys Rev E* 2017; **95**:012317. https://doi.org/10.1103/PhysRevE.95.012317
22. Valle F, Osella M, Caselle M. Multiomics topic modeling for breast cancer classification. *Cancers* 2022; **14**. https://doi.org/10.3390/cancers14051150
23. Malagoli G, Valle F, Barillot E *et al.* Identification of interpretable clusters and associated signatures in breast cancer single-cell data: a topic modeling approach. *Cancers* 2024; **16**. https://doi.org/10.3390/cancers16071350
24. El Kazwini N, Sanguinetti G. SHARE-topic: Bayesian interpretable modeling of single-cell multi-omic data. *Genome Biol* 2024; **25**:55. https://doi.org/10.1186/s13059-024-03180-3
25. Huizing G-J, Deutschmann IM, Peyré G *et al.* Paired single-cell multi-omics data integration with Mowgli. *Nat Commun* 2023; **14**:7711. https://doi.org/10.1038/s41467-023-43019-2
26. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol* 2021; **22**:351. https://doi.org/10.1186/s13059-021-02556-z
27. Ashuach T, Gabitto MI, Koodli RV *et al.* MultiVi: deep generative model for the integration of multimodal data. *Nat Methods* 2023; **20**:1222–31. https://doi.org/10.1038/s41592-023-01909-9
28. Gayoso A, Steier Z, Lopez R *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* 2021; **18**:272–82. https://doi.org/10.1038/s41592-020-01050-x
29. Virshup I, Bredikhin D, Heumos L *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
30. Virshup I, Rybakov S, Theis FJ *et al.* AnnData: access and store annotated data matrices. *J Open Source Software* 2024; **9**:4371. https://doi.org/10.21105/joss.04371
31. Bredikhin D, Kats I, Stegle O. MUON: multimodal omics analysis framework. *Genome Biol* 2022; **23**:42. https://doi.org/10.1186/s13059-021-02577-8
32. 10x genomics. *PBMC from a Healthy Donor, No Cell Sorting (10k), Single Cell Gene Expression Dataset*. https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-no-cell-sorting-10-k-1standard-2-0-0. 2021. Accessed 14 Nov 2022.
33. Stuart T, Butler A, Hoffman PJ *et al.* Comprehensive integration of single-cell data. *Cell* 2019; **177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031
34. Li C, Virgilio MC, Collins KL *et al.* Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nat Biotechnol* 2022; **41**:387–98. https://doi.org/10.1038/s41587-022-01476-y
35. Luecken M, Burkhardt D, Cannoodt R *et al.* A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: **1** Vanschoren J, Yeung S (ed), 2021. url: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/158f3069a435b314a80bdcb024f8e422-Paperround2.pdf. Accessed 20 Feb 2022.
36. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat* 2007; **35**:2769–94. https://doi.org/10.1214/009053607000000505
37. Heinz S, Benner C, Spann N *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010; **38**:576–89. https://doi.org/10.1016/j.molcel.2010.05.004
38. Cobaleda C, Schebesta A, Delogu A *et al.* Pax5: the guardian of B cell identity and function. *Nat Immunol* 2007; **8**:463–70. https://doi.org/10.1038/ni1454
39. Netherby CS, Messmer MN, Burkard-Mandel L *et al.* The granulocyte progenitor stage is a key target of IRF8-mediated regulation of myeloid-derived suppressor cell production. *J Immunol* 2017; **198**:4129–39. https://doi.org/10.4049/jimmunol.1601722
40. Gnanapragasam MN, Jiang P, Dhara A *et al.* KLF1 coordinates specialized transcriptional networks required to maintain the integrity of terminal erythropoiesis. *J Cell Sci* 2025; **138**. https://doi.org/10.1242/jcs.264036
41. 10x genomics. *Cell Ranger ARC*. Version 2.0.2. Software for analysis of Chromium Single Cell Multiome ATAC + Gene Expression data. 2022. Accessed 14 Nov 2022.
42. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022; **376**:44–53. https://doi.org/10.1126/science.abj6987

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

43. Alexander Wolf F, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018; **19**:15–5. https://doi.org/10.1186/s13059-017-1382-0
44. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* 2019; **8**:281–291.e9. https://doi.org/10.1016/j.cels.2018.11.005
45. Danese A, Richter ML, Chaichoompu K *et al.* EpiScanpy: integrated single-cell epigenomic analysis. *Nat Commun* 2021; **12**:1–8. https://doi.org/10.1038/s41467-021-25131-3
46. Zhang Y, Liu T, Meyer CA *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; **9**:R137–9. https://doi.org/10.1186/gb-2008-9-9-r137
47. Rissanen J. Modeling by shortest data description. *Automatica* 1978; **14**:465–71. https://doi.org/10.1016/0005-1098(78)90005-5
48. Peixoto T, P. The graph-tool python library. *figshare* 2014. https://doi.org/10.6084/m9.figshare.1164194 url: http://figshare.com/articles/graph_tool/1164194 (10 September 2014, date last accessed)
49. Gayoso A, Lopez R, Xing G *et al.* A python library for probabilistic analysis of single-cell omics data. *Nat Biotechnol* 2022; **40**:163–6. https://doi.org/10.1038/s41587-021-01206-w
50. Rosenberg A, Hirschberg J. V-measure: a conditional entropy-based external cluster evaluation measure. In: Eisner J (ed). *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics; 2007. pp. 410–20, 2007 url: https://aclanthology.org/D07-1043/.
51. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995; **57**:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
52. Badia-i-Mompel P, Santiago JV, Braunger J *et al.* DecoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinf Adv* 2022; **2**:vbac016. https://doi.org/10.1093/bioadv/vbac016
53. Müller-Dott S, Tsirvouli E, Vazquez M *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res* 2023; **51**:10934–49. https://doi.org/10.1093/nar/gkad841

Downloaded from academic.oup.com/bib/article/27/6/bbag608/9062982 by GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek user on 23 November 2026

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Briefings in Bioinformatics, 2026, **27**, bbag608
https://doi.org/10.1093/bib/bbag608
Problem Solving Protocol