

Systems biology

Sparse dimensionality reduction for analyzing single-cell-resolved interactions

Niklas Brunn^{1,2,*} , Maren Hackenberg^{1,2} , Camila L. Fullio^{3,4} , Tanja Vogel³ , Harald Binder^{1,2,5} 

¹Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau 79104, Germany

²Freiburg Center for Data Analysis, Modeling and AI (FDMAI), University of Freiburg, Freiburg im Breisgau 79104, Germany

³Institute of Anatomy and Cell Biology, Department Molecular Embryology, Faculty of Medicine, University of Freiburg, Freiburg im Breisgau 79104, Germany

⁴Faculty of Biology, University of Freiburg, Freiburg im Breisgau 79104, Germany

⁵Centre for Integrative Biological Signaling Studies (CIBSS), University of Freiburg, Freiburg im Breisgau 79104, Germany

*Corresponding author. Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Straße 26, Freiburg im Breisgau 79104, Germany. E-mail: niklas.brunn@uniklinik-freiburg.de.

Associate Editor: Yoshihiro Yamanishi

Abstract

Summary: Several approaches have been proposed to reconstruct interactions between groups of cells or individual cells from single-cell transcriptomics data, leveraging prior information about known ligand–receptor interactions. To enhance downstream analyses, we present an end-to-end dimensionality reduction workflow, specifically tailored for single-cell cell–cell interaction data. In particular, we demonstrate that sparse dimensionality reduction can pinpoint specific ligand–receptor interactions in relation to clusters of cell pairs. For sparse dimensionality reduction, we focus on the Boosting Autoencoder approach. Overall, we provide a comprehensive workflow, including result visualization, that simplifies the analysis of interaction patterns in cell pairs. This is supported by a Jupyter notebook that can readily be adapted to different datasets.

Availability and implementation: <https://github.com/NiklasBrunn/Sparse-dimension-reduction>.

1 Introduction

Recent enhancements of techniques for measuring gene expression levels of cells along with their spatial localization in the tissue have led to a rapidly evolving research field for computational tools that enable a detailed study of cell–cell communication (Armingol *et al.* 2024). Pioneering approaches integrate prior knowledge about known ligand–receptor interactions into a computational framework for reconstructing directed cell–cell interactions (CCIs) from transcriptomics data by leveraging the aggregated ligand expressions in sender cells and the aggregated cognate receptor expressions in receiver cells, of a specific type respectively (Cabello-Aguilar *et al.* 2020, Efremova *et al.* 2020, Jin *et al.* 2021).

However, treating cells of the same type as a single entity may overlook potential within-cell type variability of interaction patterns (Armingol *et al.* 2024). More recent approaches thus compute CCI scores at the single-cell level (Li and Yang 2022, Bafna *et al.* 2023, Raredon *et al.* 2023, Wilk *et al.* 2024). For example, NICHES (<https://github.com/msraredon/NICHES>) (Raredon *et al.* 2023) and Scriabin (Wilk *et al.* 2024) construct cell–cell interaction matrices (CCIMs) based on known ligand–receptor interactions using single-cell gene expression data. Such CCIMs consist of the interaction scores for each active ligand–receptor interaction (features) in pairs of single cells (observations).

Importantly, standard scRNA-seq analysis relies on clustering individual cells first, followed by identification of marker genes by determining differential gene expression. In contrast, interaction-pattern analysis on a CCIM instead applies dimensionality reduction to cell–cell pairs, clusters these pairs based on interaction score profiles, and uses post-hoc statistical testing to identify ligand–receptor pairs characterizing each interaction cluster (Raredon *et al.* 2023).

Here, we demonstrate a workflow that uses sparse dimensionality reduction to identify characteristic ligand–receptor interactions for clusters of cell pairs in an end-to-end manner, i.e. with ligand–receptor pairs identification already integrated into dimensionality reduction. While sparse dimensionality reduction can potentially be obtained by several approaches (Boileau *et al.* 2020, Song *et al.* 2021), we focus on the Boosting Autoencoder (<https://github.com/NiklasBrunn/BoostingAutoencoder>) approach (BAE) (Hackenberg *et al.* 2025), which we have recently introduced in the context of analyzing single-cell gene expression data. The BAE provides an interpretable representation of cell pairs in a low-dimensional latent space, where each latent dimension is linked to a sparse characterizing set of ligand–receptor interactions by model-selected weights.

However, different signs in the learned weights can make the interpretation challenging. We demonstrate how specifically the BAE approach can be easily adapted for improved

interpretation by integrating a soft clustering component into the neural network architecture.

Central to our improved approach is that the learned representation of cell pairs, cluster memberships, and selected interactions for the different clusters can be used to generate easy-to-interpret 2D visualizations using UMAP (McInnes *et al.* 2018). Additionally, ranked lists of selected interactions per cluster, along with their importance scores can be extracted from the model output.

We provide Jupyter notebooks in our GitHub repository to analyze interaction patterns in single-cell pairs, which can be easily adapted for different datasets and to explore the functionality of the adapted BAE approach.

Overall, our work advances CCI analyses with the following key contributions. (i) We present a novel application of a sparse dimensionality reduction approach to single-cell-resolved CCI data analysis, combining structured cell pair representation learning with ligand–receptor interaction selection. (ii) Interpretability is enhanced by integrating a soft-clustering component into the original BAE architecture, which allows to pinpoint specific ligand–receptor interactions in relation to clusters of cell pairs. (iii) We provide a user-friendly notebook that guides users through the entire CCI analysis pipeline, from raw gene expression data to interaction pattern results visualization. (iv) We introduce minor modifications to the original BAE optimization framework, improving both computational efficiency and model flexibility (Supplementary Material 1.1, available as [supplementary data](#) at *Bioinformatics Advances* online).

2 The boosting autoencoder

The Boosting Autoencoder (BAE) is a deep learning approach for sparse and interpretable representation learning, originally designed for the analysis of single-cell RNA sequencing data (Hackenberg *et al.* 2025), which we adapt here for the analysis of sparse CCI data (see [Supplementary Material 2](#), available as [supplementary data](#) at *Bioinformatics Advances* online for a discussion on data sparsity).

A BAE uses an autoencoder architecture comprising two concatenated neural networks, the encoder and the decoder. In the present setting of CCIs, the encoder defines a parametric mapping from the feature space of ligand–receptor interactions to a low-dimensional latent space and the decoder vice versa. The objective of the optimization is to minimize a reconstruction loss measuring the deviation of the model reconstruction to the input, under the constraint of reducing the dimensionality during the forward pass.

To enable a direct connection of interactions to latent dimensions, the BAE encoder consists of a single-layer linear neural network that is parameterized solely by a weight matrix. During training, the BAE iteratively updates the encoder weight matrix based on componentwise boosting (Tutz and Binder 2006, 2007), a stepwise feature selection approach. Specifically, the negative gradients of the autoencoder reconstruction loss with respect to the latent representation are used to construct responses, to which linear models of the input interactions are fitted in a stepwise manner. In each iteration, only the weight corresponding to the specific interaction that improves most the current linear model fit is updated. For each dimension of the latent representation, the coefficients of this linear model represent the corresponding row of the encoder weight matrix. As the encoder weights

corresponding to the model coefficients are initialized from zero, model training results in a sparse weight matrix. The componentwise boosting algorithm is embedded into a gradient-based optimization framework for finding jointly optimal encoder weights and decoder parameters. It offers the advantage of incorporating user-defined constraints into the feature selection step, e.g. to obtain disentangled latent dimensions (Hackenberg *et al.* 2025).

In our adaptation of the BAE for analyzing single-cell-resolved interaction patterns in CCIMs, we use the BAE with the disentanglement constraint to learn low-dimensional representations of cell pairs in largely uncorrelated latent dimensions, each characterized by a specific small set of ligand–receptor interactions. During parameter optimization, the BAE iteratively links interactions to latent dimensions, where the corresponding encoder weights can have different signs. Consequently, each latent dimension can capture two distinct groups of cell pairs, represented by opposite signs.

To further enhance the interpretation of learned interaction patterns in latent dimensions, we introduce the split-softmax transformation after the encoder. Specifically, this transformation allows to split two different groups of cell pairs potentially represented in the same latent dimension, while keeping track of the selected characterizing interactions for each group. The split-softmax transformation first splits the activation in latent dimensions per cell pair into their positive and negative versions, followed by a softmax transformation, to model associations of cell pairs to different latent dimensions. The split-softmax transformation allows for an equal contribution of positive and negative signals in the latent representation, whereas direct application of the softmax function would map negative signals toward zero, thus overlooking potentially meaningful signals. Each output dimension of the split-softmax transformation represents a cluster to which observations can be assigned with a certain probability. The number of clusters is thereby controlled by the user-defined number of latent dimensions for the dimensionality reduction. Supporting details about the adapted BAE model, its optimization, and guidance on controlling the sparsity level are provided in [Supplementary Material 1](#), available as [supplementary data](#) at *Bioinformatics Advances* online.

In contrast to frequently used step-by-step approaches, where dimensionality reduction is followed by clustering and subsequent differential interaction analysis, our BAE approach optimizes smoothly all model components in an end-to-end manner. An overview of the model architecture is provided in [Fig. 1a](#).

An analysis demonstrating the scalability of the BAE with respect to increasing observation and feature sizes is presented in [Supplementary Material 4](#), available as [supplementary data](#) at *Bioinformatics Advances* online.

3 Analysis of single-cell interaction patterns

3.1 Analysis workflow

In our proposed workflow for analyzing single-cell-resolved interaction patterns, the input to the BAE is a CCIM. Details on preprocessing and CCIM construction are provided in [Supplementary Material 3](#), available as [supplementary data](#) at *Bioinformatics Advances* online. A CCIM can be constructed using tools including NICHES (Raredon *et al.* 2023)

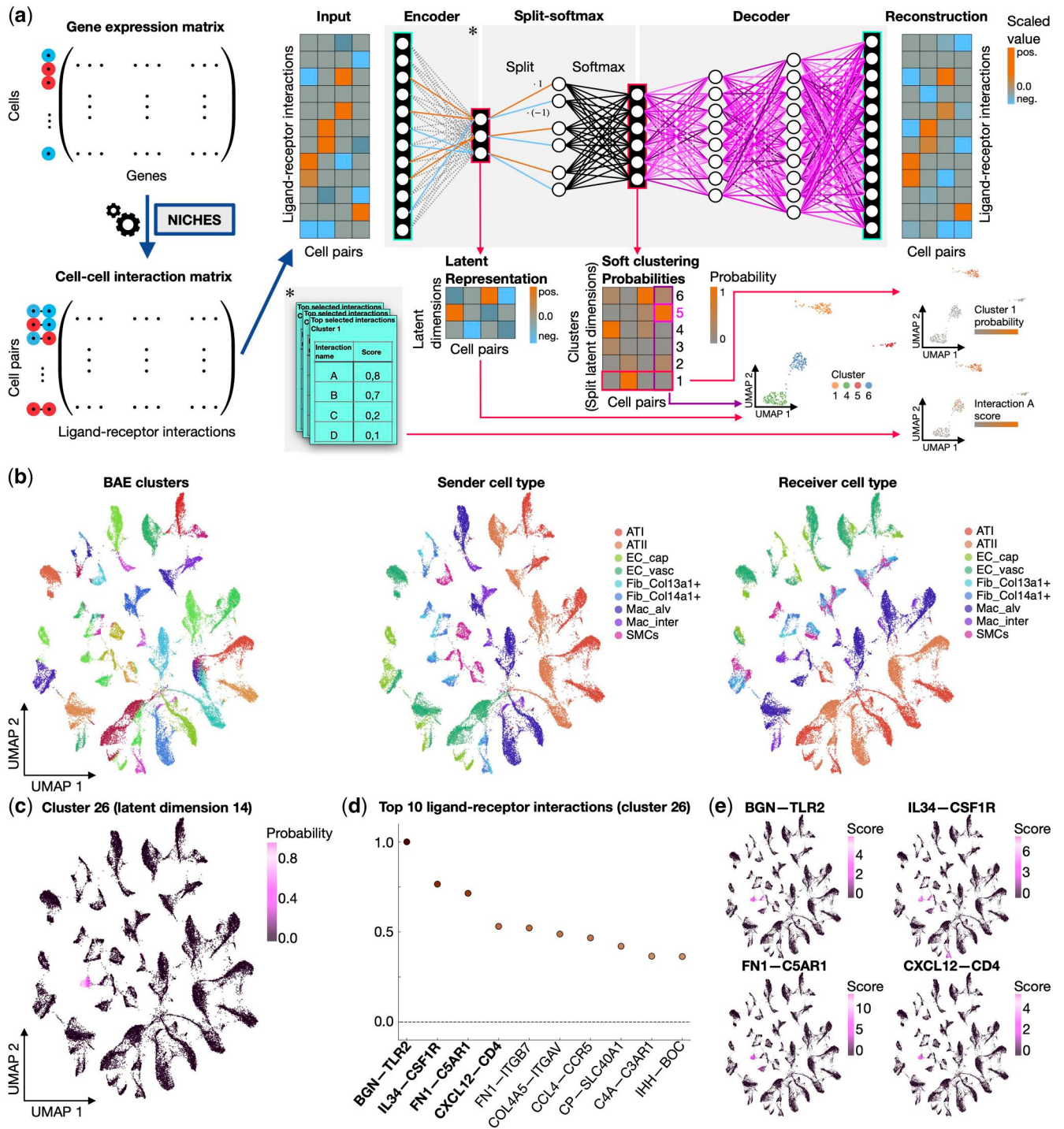


Figure 1. Schematic overview and application of the BAE. (a) (Left) The input data for training a BAE to learn single-cell interaction patterns is a CCIM constructed, e.g. by NICHES with standardized features. (Right) The BAE learns a structured and sparse connection of ligand–receptor interactions to latent dimensions (Encoder). The split-softmax transformation computes the cluster membership probabilities of cell pairs based on their representation. Each cell pair gets assigned to the cluster with the highest probability. The decoder network maps the split-softmax representation back into the ligand–receptor interaction space for reconstructing the interaction scores of cell pairs. 2D UMAP embeddings of the BAE latent representation can be used to visually inspect the results. Plots can be, e.g. colored by the cluster probabilities of cell pairs or scores of BAE-selected ligand–receptor interactions. (b–e) Result visualization of the BAE application to single-cell cell–cell interaction data computed by NICHES from rat data. The number of latent dimensions was set to $d = 30$, resulting in 60 clusters to which the model can assign cells. (b) (From left to right) 2D UMAP embedding of the latent representation of cell pairs colored by cluster membership. UMAP embedding of cell pairs colored by sender cell type. UMAP embedding of cell pairs colored by receiver cell type. (c) UMAP embedding colored exemplarily by the split-softmax representation of cell pairs in dimension 14 corresponding to cluster 26. (d) Scatter plot of the top selected ligand–receptor interactions for cluster 26. Dots represent min-max normalized positive encoder weights for the latent dimension 14 of the BAE in descending order (from left to right). (e) Top left/top right/bottom left/bottom right: UMAP embedding of cell pairs colored by interaction scores for the top 1/2/3/4 ligand–receptor interactions selected by the BAE for cluster 26.

using either gene expression data from scRNA-seq or from a spatial transcriptomics experiment (Fig. 1a).

After training a BAE on the CCIM, it provides multiple outputs for exploring interaction patterns. First, a 2D UMAP representation of the data can be computed based on the learned latent representation, to visually explore similarities in the interaction patterns of cell pairs. Second, cell pairs are assigned to clusters with learned probabilities by the soft clustering component. Third, sparse lists of ranked ligand–receptor interactions help to pinpoint interactions that characterize clusters.

The BAE results can be visually inspected by coloring the 2D UMAP representation with the information from the different result modalities. Complementary metadata, such as cell type information, can thereby enhance the interpretation of the results.

Further, our implementation of the workflow can handle spatial transcriptomics data and contains visualization functions to enhance the intuitive assessment of the learned latent patterns, making it easy for researchers to interpret and explore the data. Additionally, we provide Jupyter notebooks that allow users to either reproduce our analysis or to adapt it to other datasets, as well as to explore the model functionality using simulated data.

3.2 Visualization of results

As an exemplary application, we considered a NICHES CCIM of scRNA-seq data derived from rat lung tissue (Raredon et al. 2019).

We colored cell pairs in the 2D UMAP representation based on their cluster identity with the highest probability, the previously defined sender cell types, and the receiver cell types respectively (Fig. 1b). Comparing the three plots reveals that cell pairs with similar sender and receiver types have similar interaction profiles.

The individual clusters can be further examined for their specific characteristic interaction profiles (Fig. 1c–e). For example, inspection of the cluster membership probabilities of cell pairs in cluster 26 shows a separation from other cell pairs based on the latent patterns (Fig. 1c). The top characterizing interactions, which were selected by the BAE for cluster 26, are shown in Fig. 1d. These ligand–receptor interactions can be used to color the cell pair representation in the UMAP plot to see how well the interaction score patterns match the cluster identity of the cell pairs (Fig. 1e).

Selected ligand–receptor interactions can guide follow-up experiments to confirm cluster-specific signaling. For example, in cluster 26 the highlighted cell pairs identify fibroblasts as sender cells and interstitial macrophages as receiver cells (Fig. 1b and c). In the rat lung, the top two BAE selected ligand–receptor pairs (BGN–TLR2 and IL34–CSF1R) are indirectly supported by ligand/receptor expression in rat lungs and related systems, together with cross-species functional evidence, although neither interaction has yet been demonstrated directly in rats (Westergren-Thorsson et al. 1993, Koslowski et al. 2001, Schaefer et al. 2005, Hwang et al. 2012, Wang et al. 2012, Boström and Lundberg 2013, Dai et al. 2015, Meziani et al. 2018, Irvine et al. 2020, Lelios et al. 2020). More details are provided in [Supplementary Material 5](#), available as [supplementary data](#) at *Bioinformatics Advances* online.

A further demonstration of the analysis workflow on spatial transcriptomics data is provided in [Supplementary](#)

[Material 6](#), available as [supplementary data](#) at *Bioinformatics Advances* online.

4 Conclusion

We present a comprehensive workflow for sparse dimensionality reduction of single-cell-resolved interaction data, facilitated by the Boosting Autoencoder (BAE). To enhance the interpretability of the resulting low-dimensional embedding, we integrated a soft clustering component into the end-to-end approach. This integration enables the model to pinpoint specific interactions in relation to clusters of cell pairs, facilitating deep insights into the underlying biological processes. With our workflow implementation, researchers can perform the analysis in a customizable Jupyter notebook with their own data and visualize the results.

Author contributions

All authors contributed to the conception and design of the work. N.B. implemented the approach, and performed the computational analysis. H.B., M.H., and N.B. wrote the manuscript. C.F. and T.V. commented on the manuscript. All authors approved the final version of the manuscript.

Supplementary data

[Supplementary data](#) are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [322977937 (GRK 2344) to C.F., H.B., M.H., N.B., and T.V., 499552394 (SFB 1597) to H.B. and M.H.].

Data availability

The data underlying this article are available at <https://zenodo.org/records/6846618> and can be processed for the analysis following the instructions and scripts in our GitHub repository <https://github.com/NiklasBrunn/Sparse-dimension-reduction>.

References

- Armingol E, Baghdassarian HM, Lewis NE. The diversification of methods for studying cell–cell interactions and communication. *Nat Rev Genet* 2024;25:381–400.
- Bafna M, Li H, Zhang X. Clarify: cell–cell interaction and gene regulatory network refinement from spatially resolved transcriptomics. *Bioinformatics* 2023;39:i484–93.
- Boileau P, Hejazi NS, Dudoit S. Exploring high-dimensional biological data with sparse contrastive principal component analysis. *Bioinformatics* 2020;36:3422–30.
- Boström EA, Lundberg P. The newly discovered cytokine il-34 is expressed in gingival fibroblasts, shows enhanced expression by pro-inflammatory cytokines, and stimulates osteoclast differentiation. *PLoS One* 2013;8:e81665.

- Cabello-Aguilar S, Alame M, Kon-Sun-Tack F *et al.* Singlecellsignalr: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res* 2020;**48**:e55.
- Dai H, Pan L, Lin F *et al.* Mechanical ventilation modulates toll-like receptors 2, 4, and 9 on alveolar macrophages in a ventilator-induced lung injury model. *J Thorac Dis* 2015;**7**:616.
- Efremova M, Vento-Tormo M, Teichmann SA *et al.* Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;**15**:1484–506.
- Hackenberg M, Brunn N, Vogel T *et al.* Infusing structural assumptions into dimensionality reduction for single-cell rna sequencing data to identify small gene sets. *Commun Biol* 2025;**8**:414.
- Hwang S-J, Choi B, Kang S-S *et al.* Interleukin-34 produced by human fibroblast-like synovial cells in rheumatoid arthritis supports osteoclastogenesis. *Arthritis Res Ther* 2012;**14**:R14.
- Irvine KM, Caruso M, Cestari MF *et al.* Analysis of the impact of csf-1 administration in adult rats using a novel csf1r-mapple reporter gene. *J Leukoc Biol* 2020;**107**:221–35.
- Jin S, Guerrero-Juarez CF, Zhang L *et al.* Inference and analysis of cell–cell communication using cellchat. *Nat Commun* 2021;**12**:1088.
- Koslowski R, Pfeil U, Fehrenbach H *et al.* Changes in xylosyltransferase activity and in proteoglycan deposition in bleomycin-induced lung injury in rat. *Eur Respir J* 2001;**18**:347–56.
- Lelios I, Cansever D, Utz SG *et al.* Emerging roles of il-34 in health and disease. *J Exp Med* 2020;**217**:e20190290.
- Li R, Yang X. De novo reconstruction of cell interaction landscapes from single-cell spatial transcriptome data with deeplinc. *Genome Biol* 2022;**23**:124.
- McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426, 2018, preprint: not peer reviewed.
- Meziani L, Mondini M, Petit B *et al.* Csf1r inhibition prevents radiation pulmonary fibrosis by depletion of interstitial macrophages. *Eur Respir J* 2018;**51**:1702120.
- Raredon MSB, Adams TS, Suhail Y *et al.* Single-cell connectomic analysis of adult mammalian lungs. *Sci Adv* 2019;**5**:eaaw3851.
- Raredon MSB, Yang J, Kothapalli N *et al.* Comprehensive visualization of cell–cell interactions in single-cell and spatial transcriptomics with niches. *Bioinformatics* 2023;**39**:btac775.
- Schaefer L, Babelova A, Kiss E *et al.* The matrix component biglycan is proinflammatory and signals through toll-like receptors 4 and 2 in macrophages. *J Clin Invest* 2005;**115**:2223–33.
- Song D, Li K, Hemminger Z *et al.* scPNMF: sparse gene encoding of single cells to facilitate gene selection for targeted gene profiling. *Bioinformatics* 2021;**37**:i358–66.
- Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics* 2006;**62**:961–71.
- Tutz G, Binder H. Boosting ridge regression. *Comput Stat Data Anal* 2007;**51**:6044–59.
- Wang Y, Szretter KJ, Vermi W *et al.* Il-34 is a tissue-restricted ligand of csf1r required for the development of Langerhans cells and microglia. *Nat Immunol* 2012;**13**:753–60.
- Westergren-Thorsson G, Hernnäs J, Särnstrand B *et al.* Altered expression of small proteoglycans, collagen, and transforming growth factor-beta 1 in developing bleomycin-induced pulmonary fibrosis in rats. *J Clin Invest* 1993;**92**:632–7.
- Wilk AJ, Shalek AK, Holmes S *et al.* Comparative analysis of cell–cell communication at single-cell resolution. *Nat Biotechnol* 2024;**42**:470–83.