

PathwayEmbed: a computational tool to quantify intracellular signaling transduction states from transcriptomic data

Yaqing Huang^{1,2,3}, Sharon Gerecht³, Themis Kyriakides^{1,2}, Micha Sam Brickman Raredon^{2,4,5,*}, 

¹Department of Pathology, Yale University, New Haven, CT 06520, United States

²Vascular Biology and Therapeutics Program, Yale University, New Haven, CT 06520, United States

³Biomedical Engineering, Duke University, Durham, NC 27705, United States

⁴Department of Anesthesiology, Yale School of Medicine, New Haven, CT 06520, United States

⁵Program in Translational Biomedicine, Biological and Biomedical Sciences, Yale University, New Haven, CT 06520, United States

*Corresponding author: Vascular Biology and Therapeutics Program, Yale University, New Haven, CT 06519, United States.

E-mail: michasam.raredon@yale.edu

Associate Editor: Jonathan Wren

Abstract

Motivation: Intracellular signaling pathways regulate essential cellular functions and orchestrate complex biological processes, yet their dynamic activity remains challenging to quantify with precision. Advances in single-cell omics enable pathway activity inference at the transcriptional level; however, existing computational tools often overlook mechanistic features of signaling networks, failing to formally treat the expected directionality of transcriptional change due to signal transduction. To address this technological gap, we have engineered PathwayEmbed, an R-based computational framework for estimating intracellular signal transduction states from single-cell transcriptomic datasets.

Results: PathwayEmbed integrates KEGG pathway information with perturbation-derived RNA sequencing data to assign directional coefficients that capture gene-specific transcriptional responses to pathway activation, repression, and/or signal transduction. These coefficients, in combination with the input data, are used to compute hypothetical ON/OFF range for each pathway. Each cell is then mapped to a specific location between these ON/OFF states, and activity scores are then computed based on the distances to these reference states, providing a continuous and interpretable measure of signaling activity at single-cell resolution. This framework enables robust visualization and quantitative comparison of pathway activity across cell populations. Applied to spatial transcriptomic data, PathwayEmbed captures spatial variation in signaling transduction states and allows comparisons at both temporal and spatial scale. The framework takes tabular data as input and is broadly compatible with established single-cell analysis workflows, supports user-defined pathway ground-truths, and offers a flexible, mechanistically informed approach for quantifying and comparing intracellular signaling activity in a wide variety of contexts.

Availability: PathwayEmbed is an open-source R software under academic free license, and it is available at <https://github.com/raredonlab/PathwayEmbed>. Use-case vignettes are available at <https://raredonlab.github.io/PathwayEmbed/>.

1 Introduction

Intracellular signaling pathways are fundamental to cellular decision-making, coordinating responses to environmental cues and maintaining homeostasis across diverse biological contexts. Dysregulation of these pathways contributes to the pathogenesis of numerous diseases, including cancer, fibrosis, and autoimmune disorders. Accurate quantification of dynamic signaling states is therefore essential for elucidating disease mechanisms, improving diagnosis, and guiding the development of targeted therapies.

Advances in high-throughput sequencing technologies have enabled transcriptional-level inference of intracellular pathway activity (Zhang *et al.* 2020). While gene expression profiles offer a rich resource for this purpose, interpreting accurate pathway activities from such large-scale data remains challenging. Commonly used approaches, for example, Seurat's AddModuleScore (Tirosh *et al.* 2016), calculate a module score using the mean expression of pre-defined gene sets subtracted by control gene sets. AUCell (Aibar *et al.* 2017) and GSVA (Hanzelmann *et al.* 2013) use a gene-ranking approach to determine the correlation between pathways and cells

Received: 12 November 2025. Revised: 8 April 2026. Accepted: 11 May 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.

based on the area under ranked genes. Each of the above methods are straightforward to implement but treat pathways as undirected collections of genes. In other words, genes in given pathways are regarded as upregulated in all conditions. This overlooks critical biological features, including the activating or inhibitory roles of individual molecules, feedback regulation, and differential weighting based on mechanistic relevance. To be specific, transcriptional changes in response to external stimuli might not always be the same at the feature level, depending on the role of corresponding proteins involved. Simple aggregation of gene expression without encoding this directionality may therefore fail to reflect the true net transduction state of signaling pathways, particularly in heterogeneous or perturbed systems. Several existing tools incorporate aspects of regulatory logic or directionality and demonstrate improvement in interpretation of multi-omic data. For example, Signaling Pathway Impact Analysis (Tarca *et al.* 2009) incorporates KEGG interaction topology to identify perturbed pathways but is restricted to population-level analysis and does not produce per-cell activity scores. PROGENy (Schubert *et al.* 2018) and VISION (DeTomaso *et al.* 2019) are applicable to single-cell data and both support signed gene signatures. PROGENy (Schubert *et al.* 2018) derives continuous, signed gene weights by fitting a linear model across a large compendium of bulk RNA-seq perturbation experiments, then applies these weights as a fixed matrix multiplication to score each cell. VISION (DeTomaso *et al.* 2019) computes per-cell weighted expression sums from signed signatures and uses local autocorrelation statistics on the single-cell similarity manifold to identify biologically coherent patterns of variation. However, both approaches generate summary scores with no inherent bounds that based on a predetermined external reference rather than the signaling landscape of the dataset being analyzed.

Inspired by these advances, we developed PathwayEmbed, an R-based computational framework to quantify intracellular signaling transduction activity from transcriptomic data. The key distinction from existing tools is our use of a geometric, reference-state embedding approach: instead of aggregating expression values or computing scores relative to a fixed external population, PathwayEmbed derives its reference anchors from the observed expression extremes of pathway genes within the input data itself, adapted to the biological context at hand. PathwayEmbed accepts raw, normalized, or scaled expression data, offering flexibility for different users and processing workflows. Benchmarking analyses demonstrate both reliability and scalability with increasing dataset size. The resulting pathway scores can be readily visualized with ggplot2 (Wickham 2016) or integrated into established single-cell analysis platforms, including Seurat (Hao *et al.* 2024), to facilitate customized downstream analyses.

2 Approach

PathwayEmbed is an R package for inferring intracellular signaling transduction states from transcriptomic data using a reference-based embedding framework. The framework operates on one pathway at a time, using as input a single-cell expression matrix (cells $\times$ genes) and a focused pathway model (pathway genes $\times$ coefficients), to generate pathway activation scores for each cell. Each gene is assigned a directional coefficient representing its expected transcriptional response during pathway activation.

Directionality is informed by a data-driven perturbation-derived transcriptional data, and tools are provided to allow users to leverage priors in their determination of expected coefficients. In the current implementation, coefficients are discretized to +1 or -1 to encode upregulation or downregulation, respectively; however, the framework is flexible and allows user-defined coefficients to reflect alternative pathway models or context-specific regulatory weights if this is desired. The resulting pathway coefficient table can be viewed using ListPathway() and directly loaded via the LoadPathway() functions. Given an input expression matrix, PathwayEmbed first extracts pathway-associated genes and optionally performs row-wise scaling using DataPreProcess(). Rather than applying weighted aggregation, the method constructs pathway-specific reference states using PathwayMaxMin(). Two high-dimensional reference points, pathway.on and pathway.off, are generated to define the full extent of pathway-transduction space. Briefly, the ON reference takes the expression value that maximizes the signal in the direction indicated by its coefficient, that is, the maximum observed expression for positively signed genes (+1) and the minimum observed expression for negatively signed genes (-1). The OFF state is constructed with the opposite logic. These ON/OFF values define the observed possible expression range for pathway-relevant genes within the input dataset and can be subsequently used for normalization. Each cell is then positioning within this pathway-defined space by computing its distance to the ON and OFF reference states using a user-defined metric (Manhattan distance by default, or alternatively Euclidean distance). Pathway activity is quantified as the normalized relative distance between these two reference states [$\text{score} = \text{distance_to_OFF} / (\text{distance_to_ON} + \text{distance_to_OFF})$], yielding a continuous score bounded between inactive (OFF-like) and active (ON-like) states. Importantly, this formulation defines pathway activity as a relative position within a data-adaptive transduction space, rather than an absolute value derived from predefined signatures. Finally, resulting scores can be merged with cell metadata to enable comparisons for various scientific purposes (Fig. 1). PathwayEmbed is implemented in R (≥ 3.5) and depends on dplyr (Wickham *et al.* 2025), tidyverse (Wickham *et al.* 2019), matrixStats (Bengtsson 2017), ggplot2 (Wickham 2016) and effsize (Torchiano 2020). Detailed function descriptions, biological assumptions, and recommended best practices are provided in the Supplementary Material. The package, along with example data and usage documentation, is available at <https://raredonlab.github.io/PathwayEmbed/>.

3 Application

3.1 Advantages of PathwayEmbed

PathwayEmbed provides a novel framework for quantifying and visualizing pathway activation states at single-cell resolution by integrating curated pathway molecule coefficients with gene expression data. Unlike many existing pathway analysis tools that focus solely on gene set enrichment or overall gene relevance, PathwayEmbed incorporates the directional influence (activating/inhibitory, produced/degraded) of each pathway component according to information from prior perturbation data. This enables a more mechanistic and accurate interpretation of pathway dynamics, which we have validated using known perturbation datasets,

PathwayEmbed

Reference-based pathway activity scoring for single-cell data

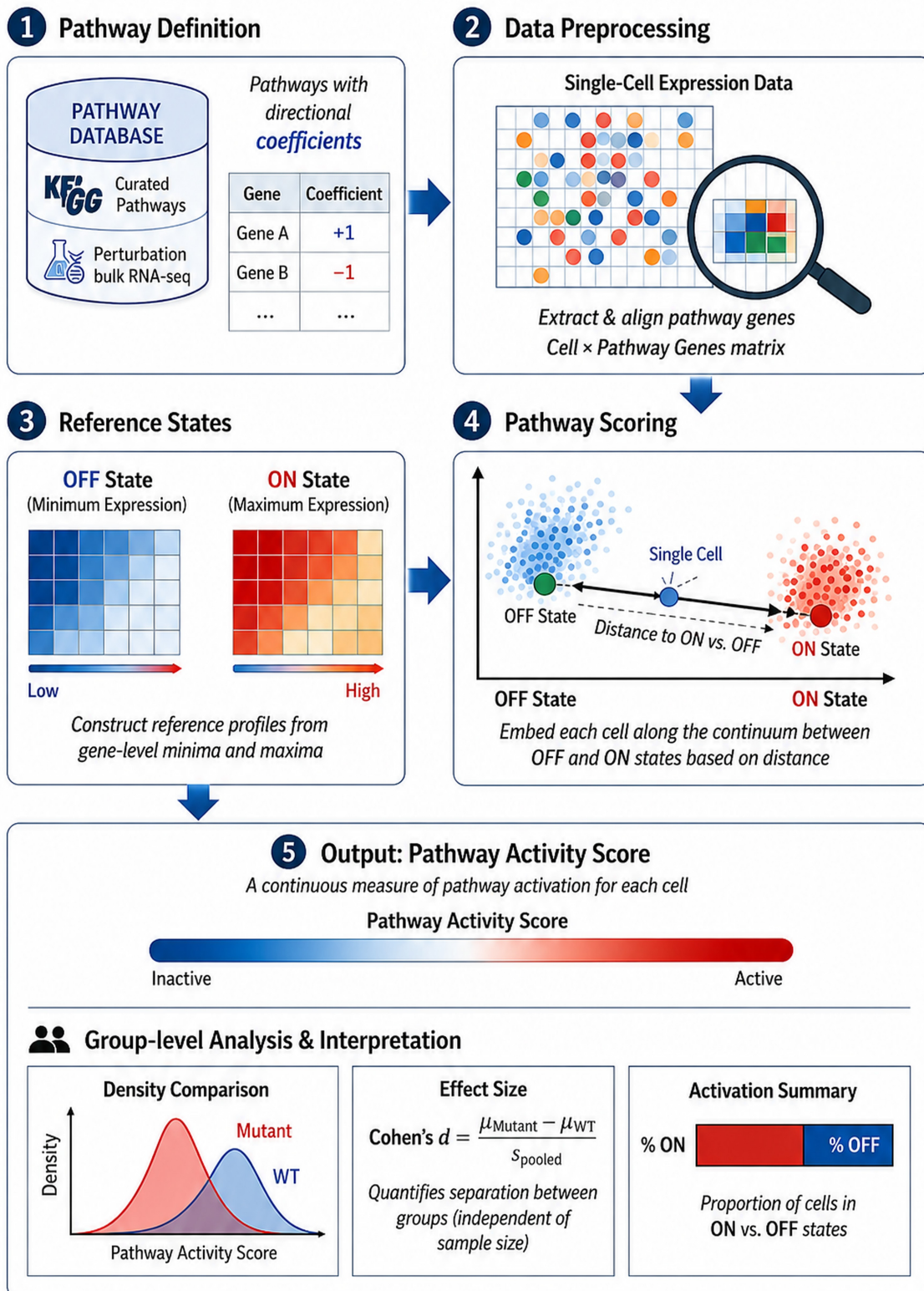


Figure 1 Overview of PathwayEmbed.

such as *Ctnnb1* enhancer and Notch knock out single-cell RNA-seq experiments (Hua *et al.* 2024) (Supplementary Figs. 5 and 6). Pathway coefficients are fully customizable (an example is provided in the vignettes detailing appropriate methods), allowing researchers to adapt pathway definitions to specific biological contexts or experimental conditions, which is frequently necessary during applied biological analysis. Moreover, PathwayEmbed supports input of either raw or normalized counts or Seurat object, providing flexibility for diverse preprocessing workflows (vignette demonstration in Toy Set: https://raredonlab.github.io/PathwayEmbed/articles/examples_updated.html).

By normalizing molecular distances relative to hypothetical “On” and “Off” states, the resulting scores are easily interpretable compared to other decomposition approaches (Fan *et al.* 2016, Tomfohr *et al.* 2005). This design provides several advantages. By encoding directionality through reference-state construction rather than direct weighting, PathwayEmbed preserves interpretability while capturing gene-specific regulatory behavior. The use of dataset-derived extrema enables adaptation to diverse biological contexts, including temporal and spatial transcriptomic data. However, as pathway activity is defined relative to dataset-specific reference states, it is important to note that scores are inherently context-dependent. As such, comparisons across independent datasets may require additional normalization or calibration. Overall, the pairwise distance method approach simplifies interpretation, producing continuous activation scores that can be visualized using density plots or integrated with cell metadata for downstream comparative analyses.

3.2 Compute net transduction states of individual cells

To illustrate its utility, we first generated a synthetic toy dataset containing two populations of cells expressing Wnt pathway molecules and other random genes (Supplementary Figs. 2–4). Using normalized values, we calculated the net Wnt transduction score for individual cells and visualized the density distribution with the embedded plotting function. The shift in Wnt transduction states between predefined wild-type (WT) and mutant groups remained highly consistent across different computational batch sizes. Effect size analysis using Cohen’s *d* revealed a large separation between the two populations distributions, with minimal correlation between Cohen’s *d* values and the number of run iterations, indicating the robustness of the approach for detecting differences in signaling transduction states. We further applied PathwayEmbed to public scRNA-seq datasets from perturbation experiments (Remark *et al.* 2023, Hua *et al.* 2024), and the results aligned with the expected biological interventions, including gene knockout and aging (Supplementary Figs. 5–7). Detailed descriptions of each experiment are provided in the Supplementary Data.

3.3 Map spatial distribution of signaling activation

We further applied PathwayEmbed to mouse embryonic spatial transcriptomics data (Chen *et al.* 2022) to examine Wnt, Notch,

TGF- β , Hippo, and HIF-1 α signaling transduction states. By integrating pathway activation scores with spatial coordinates and evaluating the resulting patterns with Moran’s *I*, we observed clear spatial segmentation of signaling states across embryonic regions during development (Supplementary Figs. 8 and 9). Density plots and spatial feature plots enabled direct comparison of relative transduction states across developmental time points (Supplementary Figs. 8 and 9). A detailed vignette demonstrating this workflow is available online at: https://raredonlab.github.io/PathwayEmbed/articles/spatial_pathway.html.

4 Conclusion

PathwayEmbed provides a flexible and mechanistically informed framework for quantifying pathway activity at single-cell resolution, offering insights that complement and extend beyond conventional gene set enrichment methods. Its core innovation is the geometric positioning of each cell between data-derived ON and OFF reference states constructed from a perturbation-informed pathway gene set, producing a continuous and interpretable transduction score without reliance on fixed external reference panels. This design captures pathway-specific directional structure that adapts to the biological context of each dataset. Continued refinement of molecular coefficients, coupled with integration of experimental validation, will further improve accuracy and broaden applicability, enabling more precise and biologically grounded analyses of intracellular signaling in single-cell omics.

Acknowledgement

The authors used artificial intelligence (AI) tools, including ChatGPT (OpenAI), for grammar and language editing and for assistance with figure revision.

Author contributions

Yaqing Huang (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal]), Sharon Gerech (Funding acquisition [equal], Supervision [equal], Writing—review & editing [equal]), Themis Kyriakides (Funding acquisition [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal]), and Micha Sam Brickman Raredon (Conceptualization [equal], Data curation [equal], Funding acquisition [equal], Methodology [equal], Project administration [equal], Visualization [equal], Writing—review & editing [equal]).

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

M.S.B.R. performs data-science consultation for, and holds stock in, Humacyte, Inc. Humacyte does not influence the conduct, description, or interpretation of any research lead by M.S.B.R. All other authors declare no competing financial interests.

Funding

This work was supported by T32GM086287 from the National Institute of General Medical Sciences (NIGMS), laboratory startup funds from the Yale School of Medicine and the Yale Department of Anesthesiology, and American Heart Association (AHA 26POST1553068, <https://doi.org/10.58275/AHA.26POST1553068.pc.gr.240533>). The opinions expressed are those of the authors and do not necessarily represent the thoughts or opinions of NIGMS, NIH, or the United States government.

Data availability

All data needed to evaluate the conclusions in the paper are included in the manuscript and package and can be accessed publicly.

References

- Aibar S, González-Blas CB, Moerman T *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6.
- Bengtsson H. 2017. matrixStats: functions that apply to rows and columns of matrices (and to Vectors). R package version 0.52.2. <https://github.com/HenrikBengtsson/matrixStats>.
- Chen A, Liao S, Cheng M *et al*. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;**185**:1777–92.e21.
- DeTomaso D, Jones MG, Subramaniam M *et al*. Functional interpretation of single cell similarity maps. *Nat Commun* 2019;**10**:4376.
- Fan J, Salathia N, Liu R *et al*. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 2016;**13**:241–4.
- Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;**14**:7.
- Hao Y, Stuart T, Kowalski MH *et al*. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304.
- Hua X, Zhao C, Tian J *et al*. A Ctnnb1 enhancer transcriptionally regulates wnt signaling dosage to balance homeostasis and tumorigenesis of intestinal epithelia. *Elife* 2024;**13**:RP98238.
- Remark LH, Leclerc K, Ramsukh M *et al*. Loss of notch signaling in skeletal stem cells enhances bone formation with aging. *Bone Res* 2023;**11**:50.
- Schubert M, Klingner B, Klünemann M *et al*. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun* 2018;**9**:20.
- Tarca AL, Draghici S, Khatri P *et al*. A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**:75–82.
- 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.
- Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 2005;**6**:225.
- Torchiano M. 2020. effsize: efficient effect size computation.
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016.
- Wickham H, Averick M, Bryan J *et al*. Welcome to the tidyverse. *JOSS* 2019;**4**:1686.
- Wickham H, Henry L, Müller K *et al*. 2025. dplyr: A grammar of data manipulation. In. R package version 1.1.4, <https://dplyr.tidyverse.org>.
- Zhang Y, Ma Y, Huang Y *et al*. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput Struct Biotechnol J* 2020;**18**:2953–61.