ACPL: a root-free polar ordering for cyclic single-cell trajectories

Shreyo Ghosh

School of Engineering, Rutgers University

ghoshshreyo2007@gmail.com

August 28, 2026

Abstract. Trajectory inference on cyclic biological processes has a specific geometric failure mode. Graph constructions that minimise Euclidean cost route across the interior of a closed loop rather than around it, placing a backward edge at the G2M to G1 closure, the most consequential transition in the cell cycle. We describe a construction that removes this possibility by ordering cells along the angular coordinate of a planar embedding before any graph is built, and admitting edges only in the direction of increasing smoothed arc length. The resulting directed graph is acyclic by construction, is invariant to the choice of starting cell, and is built in $O(N \log N)$ time. Angular ordering of cell-cycle data is established practice and is not claimed here; the contribution is the directed-graph construction and its evaluation. Across four public datasets, scored by an origin-free rank statistic necessary because ACPL takes no starting cluster while its baselines are given one, it recovers cyclic order more accurately than Monocle-style minimum spanning trees and Slingshot run on their native inputs, consistently across ten embedding seeds, with six of sixteen paired bootstrap comparisons excluding zero after Holm correction for multiplicity and no comparison favouring a baseline. It is exceeded by tricycle, a reference-based method, on two of three datasets tested, and recovers no significant signal on a yeast series where tricycle cannot run at all. The construction cannot determine the rotational direction of the recovered cycle, so independence from a starting cluster is a trade rather than an advantage. This article supersedes the earlier ACPL preprint, engrXiv 10.31224/7087, whose evaluation contained errors; they are itemised in Section 6.

Keywords: cell cycle; trajectory inference; polar coordinates; directed acyclic graph; single-cell RNA sequencing.

1 Introduction

Single-cell RNA sequencing captures a static snapshot of cellular states from which trajectory inference reconstructs an ordering. For cyclic processes the reconstruction has a geometric

character that linear and branching methods do not address.

Projection of cell-cycle expression data into two dimensions by a topology-preserving algorithm such as UMAP (McInnes et al., 2018) produces a closed loop that faithfully reflects the cyclic topology of the underlying process. That fidelity creates the failure. Graph constructions based on Euclidean proximity, including the minimum spanning trees underlying Monocle (Trapnell et al., 2014; Qiu et al., 2017), minimise total spanning cost, and on a circular manifold the cheapest connection between the end and the start of the cycle is the chord through the interior rather than the arc around the rim. Cells completing mitosis sit close in the projection to cells re-entering G1 because the cycle closes geometrically, despite being maximally distant in time. A greedy spanning tree connects them, and the backward edge lands at the loop closure. The defect cannot be repaired by assigning direction afterwards, because the edge is already in the graph; it has to be prevented when the graph is built.

The literature has noted the difficulty without solving it in these terms. Haghverdi et al. (2016) observed it for diffusion pseudotime, and Wolf et al. (2019) motivated PAGA partly by the same concern, though PAGA operates at cluster resolution and gives up single-cell ordering. RNA velocity (La Manno et al., 2018; Bergen et al., 2020) supplies direction from splicing kinetics but requires spliced and unspliced counts unavailable in legacy SMART-seq2 data. Slingshot (Street et al., 2018) fits principal curves through clusters and offers no formal acyclicity guarantee.

1.1 What is and is not new here

Ordering cells by angular position around an origin in a two-dimensional embedding is established practice in cell-cycle inference and is not a contribution of this work. Zheng et al. (2022) infer cell-cycle position directly as a polar angle and note that this is usually done by a variant of the polar angle; Liang et al. (2020) fit a sinusoidal autoencoder to the same circular structure; Liu et al. (2017) recover cell-cycle order by consensus travelling-salesman ordering; Liu et al. (2022) and Hsiao et al. (2020) address the same problem by other means.

What is offered here is narrower. ACPL, for Arc plus LOESS Polar-Linearisation, is a construction that turns an angular ordering into a directed graph that is acyclic by construction, is invariant to the choice of starting cell, and is built in $O(N \log N)$ time; an evaluation of that construction against both general-purpose trajectory methods and a reference-based cell-cycle specialist, with baselines run on their native inputs; and an explicit account of what the construction cannot do.

2 Methods

2.1 Polar transform

Given a two-dimensional embedding $X = \{x_1, \ldots, x_N\}$, the hub c^* is chosen to minimise the variance of radial distances,

$$c^* = \arg\min_c \mathrm{Var}(\{\|x_i - c\|_2\}), \tag{1}$$

optimised by Nelder-Mead from the coordinate median. Polar coordinates follow as $\theta_i = \mathrm{atan2}(y_i - c_y^*,\, x_i - c_x^*)$ and $r_i = \|(x_i - c_x^*,\, y_i - c_y^*)\|_2$. The angle is unwrapped to a continuous increasing variable to accommodate multi-lap data.

The radial coordinate is discarded. On a closed loop cells at every phase lie at approximately equal distance from the hub, so r carries no temporal signal; applying the same smoothing pipeline to r instead of θ scores below the Cartesian baseline, confirming this empirically.

2.2 Arc length and the necessity of smoothing

Cumulative arc length is computed along the θ-ordered traversal and smoothed by LOESS to give the time proxy s^*. The smoothing is not cosmetic. Raw cumulative arc length $s_i = \sum_{k<i} \|x_{k+1} - x_k\|_2$ is strictly increasing in traversal index by positive definiteness of the norm, so a monotonicity filter applied to it admits every edge in traversal order and returns perfect structural accuracy as an arithmetic identity rather than as geometric inference. Smoothing disrupts strict monotonicity at curvature reversals and makes the filter selective.

This has a practical consequence worth stating plainly: any implementation of this construction that reports perfect accuracy is measuring a tautology, not performance.

2.3 Graph construction and its properties

Edges are admitted only in the direction of increasing s^*. The three results below are direct consequences of totally ordering cells by a scalar and are stated as properties rather than theorems.

Property 1 (Acyclicity)**.** *The directed graph admitting edge $(i \to j)$ only when $s_j^* > s_i^*$ contains no directed cycle, since a cycle $i_1 \to \cdots \to i_k \to i_1$ would imply $s_{i_1}^* > s_{i_1}^*$.*

Property 2 (Root invariance)**.** *The global ordering is determined by s^* alone and does not depend on a designated starting cell.*

Property 3 (Cost)**.** *Sorting and smoothing are $O(N \log N)$ and the K-nearest valid neighbour step after sorting is $O(NK)$, giving $O(N \log N)$ for fixed K.*

Table 1. τ_{cyc} against baselines on native inputs, across ten embedding seeds. Baselines computed from principal components are seed-independent.

Dataset	Slingshot (PCA)	MST (PCA)	ACPL, mean (SD)	seeds ahead
Nestorowa HSC	0.4096	0.2473	0.4301 (0.0092)	10/10
Buettner mESC	0.3940	0.1618	0.4516 (0.0139)	10/10
Leng hESC	0.6963	0.4684	0.7721 (0.0318)	10/10
Richard T cells	0.1046	0.1041	0.1579 (0.0076)	10/10

3 Evaluation protocol

Four datasets are used, all from the Bioconductor `scRNAseq` package: Nestorowa HSC (Nestorowa et al., 2016), $N = 1920$, mouse, Seurat phase scoring; Buettner mESC (Buettner et al., 2015), $N = 288$, mouse, FACS labels; Leng hESC (Leng et al., 2015), $N = 247$, human, FACS labels; Richard T cells (Richard et al., 2018), $N = 572$, mouse, Seurat phase scoring. Gene identifiers are Ensembl in current releases and are mapped to symbols before phase scoring.

Three features of the protocol require comment.

Chance levels are computed and reported. Agreement statistics on ordinal phase labels have chance levels set by label imbalance. For a tie-counting windowed concordance the chance level is $(1 + \sum_k p_k^2)/2$, which is 66.67, 66.77, 67.59 and 69.31 percent on these four datasets and is never 50 percent.

Orientation is handled identically for every method. Resolving handedness by taking whichever direction scores higher against the ground-truth labels uses the answer key to orient the answer and inflates the null.

The primary statistic is origin-free. We report τ_{cyc}, the maximum Kendall τ_b against ordinal phase rank over a grid of origin rotations and both traversal directions, with the identical maximisation applied under permutation so that the selection is priced into the null. This is necessary because the construction takes no starting cluster while its baselines are given one: the same perfect cyclic ordering scores $\tau = +0.82$ or -0.27 depending only on where it is taken to begin, so a fixed linear scale from G1 through S to G2M rewards methods that are told where to start.

Baselines are run on their native inputs, ten principal components, in addition to the shared two-dimensional embedding.

Table 2. Against tricycle (Zheng et al., 2022). Paired bootstrap, 400 resamples.

Dataset	ACPL	tricycle	Difference, 95% CI
Nestorowa HSC	0.4499	0.6773	-0.2275 $[-0.2577, -0.1892]$
Buettner mESC	0.4311	0.5256	-0.0945 $[-0.1892, -0.0241]$
Richard T cells	0.1704	0.1784	-0.0080 $[-0.0945, +0.0629]$

4 Results

4.1 Against general-purpose trajectory methods

Forty of forty seed-by-dataset comparisons favour ACPL over Slingshot in native form, and forty of forty over MST in native form. Paired bootstrap comparisons on a fixed embedding number sixteen across the four datasets. Seven exclude zero unadjusted and six survive Holm correction over the family; all six favour ACPL and no comparison favours a baseline under either accounting.

Against MST in native form the margin is large and survives adjustment on three datasets: Nestorowa $+0.203$, Buettner $+0.269$ and Leng $+0.313$, each at $p_{\mathrm{Holm}} = 0.016$. Against Slingshot in native form, the stronger of the two baselines, it survives on Leng ($+0.085$, $p_{\mathrm{Holm}} = 0.022$) but not on Nestorowa, where the unadjusted interval $[+0.012, +0.068]$ becomes $p_{\mathrm{Holm}} = 0.070$ and a Bonferroni simultaneous interval of $[-0.005, +0.085]$. That comparison is the narrowest in the table, mean $+0.0205$ across ten embedding seeds with standard deviation 0.0092 and a least favourable seed of $+0.0028$. It is reported here as suggestive, not as established.

One observation constrains the interpretation. On Nestorowa, Slingshot scores 0.3452 on the two-dimensional embedding and 0.4096 on principal components, so the embedding degrades it, while ACPL attains 0.4499 on that same embedding. The advantage is therefore not attributable to the embedding being generally favourable. It is specific to a construction organised around the loop closure that the embedding produces, which a principal curve does not exploit.

4.2 Against a reference-based method

tricycle exceeds ACPL on two of three datasets, significantly in both, and ties on the third. The Nestorowa margin of 0.227 is more than five times ACPL's own advantage over native Slingshot on the same data. tricycle projects onto a reference learned from external mouse neurosphere data and therefore brings information ACPL does not have. That explains the result without diminishing it.

4.3 A negative result on yeast

Because tricycle depends on a mammalian reference it cannot be applied outside the organisms that reference covers; on *Saccharomyces cerevisiae* it finds no projection genes and fails outright. This suggests a domain ACPL might occupy. It does not.

On the Spellman alpha-factor synchronised series (Spellman et al., 1998), restricted to the first full cycle so that the planar construction is within scope ($n = 10$ timepoints), circular correlation against experimental timepoint is $+0.088$ ($p = 0.81$), against $+0.245$ for MST on the embedding and $+0.180$ for MST on principal components. On the full 18-point series spanning about 1.8 cycles the construction scores -0.115 ($p = 0.64$) while MST on principal components reaches $+0.545$ ($p = 0.02$).

Two qualifications apply and neither rescues the result. The sample is small, and $p = 0.81$ at $n = 10$ is not evidence of absence. And the full series is outside scope, since a planar construction places every cell on one circle and cannot separate successive laps. The burden was on ACPL to demonstrate a capability where its competitor is structurally excluded, and it did not.

5 Discussion

5.1 What these results support

On mammalian single-cell cell-cycle data ACPL recovers cyclic order more accurately than Monocle-style minimum spanning trees and Slingshot run on their native inputs, consistently across four datasets and ten embedding seeds, with no comparison favouring a baseline. It requires no starting cluster, admits only forward edges by construction, and runs in $O(N \log N)$.

It is exceeded by a reference-based specialist on two of three datasets tested, does not generalise to yeast, and cannot separate laps in multi-cycle data. Root-freedom differentiates it from Slingshot but not from tricycle, which also requires no starting cluster.

5.2 Root-freedom is a trade

On Nestorowa and Buettner the inferred cycle runs opposite to G1 through S to G2M. This is the correct cycle traversed the wrong way, established by the three forward gaps between phase means summing to 4π rather than 2π, and nothing internal to ACPL detects it. Independence from a starting cluster is bought at the cost of being unable to orient the result, so an external anchor is mandatory rather than a supplementary convenience. Schwabe et al. (2020) resolve the same ambiguity using RNA velocity, which is the right kind of evidence: external, and about the arrow of time rather than about the labels being scored.

5.3 Limitations

ACPL operates on a two-dimensional embedding, and rigorous recovery theory for circular ordering attaches to spectral embeddings rather than to UMAP (Recanati et al., 2018). Four datasets is a small evaluation, and on two of them the phase labels are themselves inferred from the same expression data by Seurat scoring, so the ground truth is not independent of the input.

Sixteen comparisons invite a multiplicity objection. It is answered rather than deferred: Holm correction over the family, which is valid under the arbitrary dependence induced by sharing bootstrap resamples within a dataset, retains six of the seven, and Bonferroni simultaneous intervals retain the same six (`analysis/09_multiplicity.R`). The comparison that does not survive is ACPL against native Slingshot on Nestorowa, which is also the narrowest margin in the table. That adjustment cost the single result this work would otherwise have leaned on hardest, and the claim has been narrowed to match rather than restated at a weaker threshold.

6 Relationship to the earlier preprint

This article supersedes engrXiv 10.31224/7087 and should be cited in its place. The earlier version's evaluation contained the following errors.

Its stated random baseline of approximately 50 percent was wrong by 16.7 to 19.3 percentage points, for the reason given in Section 3. Orientation was resolved asymmetrically, with baselines receiving a maximisation over two orientations that the proposed construction did not, in two of the benchmark scripts. Two datasets carried incorrect species designations: Leng is human embryonic stem cells rather than mouse, and Richard is mouse rather than human. Section 4.9 of that version analysed data attributed to GEO accession GSE64016 and described as HeLa; that accession is the Oscope series of Leng et al. (2015), comprising 213 unsorted H1 hESC and 247 H1-Fucci sorted cells, contains no HeLa cells, and has no subset of the stated size. That section is withdrawn and the data underlying it cannot presently be identified.

Four of the five dataset loaders in the accompanying repository no longer execute against current Bioconductor releases, following a change in the gene-identifier convention returned by `scRNAseq`.

Correcting the first two errors reverses the sign of three of four reported comparisons, one of them from -6.7 to $+0.8$ percentage points, and removes that version's central empirical claim that a minimum spanning tree outperformed the method on three of five datasets. Also withdrawn are the claim to introduce a polar coordinate prior, addressed in Section 1.1; any scaling advantage over Slingshot, which that version's own supplementary material contradicted; and the designation of the three formal results as theorems.

The general methodological issues these errors illustrate are developed separately.

Data and code availability

All datasets are public. Nestorowa, Buettner, Leng and Richard are available through the Bioconductor `scRNAseq` package; the yeast series is from Spellman et al. (1998) through `yeastCC`.

Code is at https://github.com/shreyosecret/ACPL. The repository now contains the corrected dataset loaders, the metric implementations, and the evaluation scripts implementing the protocol of Section 3, in `R/` and `analysis/` respectively. The code accompanying the earlier version of this work is retained unmodified under `deprecated/` so that the changes described in Section 6 can be inspected directly; it is not maintained, four of its five dataset loaders no longer execute against current Bioconductor releases, and it should not be used to reproduce anything reported here.

Running `run_all.R` from the repository root reproduces the analyses in order and regenerates session information. Package versions move underneath these datasets, as Section 6 documents, so the recorded session information should be treated as part of the result rather than as an appendix to it.

Acknowledgements

The author thanks Samiran Ghosh for informal statistical consultation on the bootstrap and permutation designs. He is not an author and bears no responsibility for the conclusions. AI-assisted tools were used for LaTeX typesetting, prose refinement and code review in accordance with ISCB policy; all analyses, derivations, interpretations and conclusions are the author's own.

References

Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38:1408–1414, 2020.

Florian Buettner, Kedar N. Natarajan, F. Paolo Casale, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10): 845–848, 2016.

Chiaowen Joyce Hsiao, PoYuan Tung, John D. Blischak, et al. Characterizing and inferring

quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Research*, 30(4): 611–621, 2020.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, et al. RNA velocity of single cells. *Nature*, 560: 494–498, 2018.

Ning Leng, Li-Fang Chu, Chris Barry, et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nature Methods*, 12(10):947–950, 2015.

Shaoheng Liang, Fang Wang, Jincheng Han, and Ken Chen. Latent periodic process inference from single-cell RNA-seq data. *Nature Communications*, 11:1441, 2020.

Jiajia Liu, Mengyuan Yang, Weiling Zhao, and Xiaobo Zhou. CCPE: cell cycle pseudotime estimation for single cell RNA-seq data. *Nucleic Acids Research*, 50(2):704–716, 2022.

Zehua Liu, Huazhe Lou, Kaikun Xie, et al. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nature Communications*, 8:22, 2017.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

Sonia Nestorowa, Fiona K. Hamey, Blanca Pijuan-Sala, et al. A single-cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. *Blood*, 128(8):e20–e31, 2016.

Xiaojie Qiu, Qi Mao, Ying Tang, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, 2017.

Antoine Recanati, Thomas Kerdreux, and Alexandre d'Aspremont. Reconstructing latent orderings by spectral clustering. In *arXiv:1807.07122*, 2018.

Arianne C. Richard, Aaron T. L. Lun, Winnie W. Y. Lau, et al. T cell cytolytic capacity is independent of initial stimulation strength. *Nature Immunology*, 19(8):849–858, 2018.

Daniel Schwabe, Sara Formichetti, Jan Philipp Junker, Martin Falcke, and Nikolaus Rajewsky. The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology*, 16(11):e9946, 2020.

Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

Kelly Street, Davide Risso, Russell B. Fletcher, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19:477, 2018.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, et al. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20:59, 2019.

Shijie C. Zheng, Genevieve Stein-O'Brien, Jonathan J. Augustin, et al. Universal prediction of cell-cycle position using transfer learning. *Genome Biology*, 23:41, 2022.