CoPhaser: generic modeling of biological cycles in scRNA-seq with context-dependent periodic manifolds

Yves Paychère[1], Andrea Salati[1], Cédric Gobet[1], Felix Naef[1*]

[1*]Institute of Bioengineering, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, CH-1015, Switzerland.

*Corresponding author(s). E-mail(s): felix.naef@epfl.ch;
Contributing authors: yves.paychere@epfl.ch; andrea.salati@epfl.ch; cedric.gobet@epfl.ch;

Abstract

Biological cycles such as the cell cycle and circadian rhythms are fundamental cell-autonomous processes that are often obscured in single-cell RNA sequencing (scRNA-seq) data by other sources of variation, including cell-type identity, metabolic or disease-associated states. Disentangling these continuous periodic trajectories from other forms of cellular variability remains a major challenge in single-cell analysis. Here, we introduce CoPhaser, an algorithm designed to learn context-dependent periodic manifolds that decompose scRNA-seq count data into independent periodic and non-periodic sources of variation, while maintaining interpretability of manifold coordinates across diverse biological contexts. CoPhaser employs a biologically informed variational autoencoder with a structured latent space that explicitly separates cycle phases from cellular context, while controlling their mutual information. By modeling gene expression as context-modulated harmonic functions, the model captures flexible but biologically motivated deformations of periodic manifolds. We demonstrate that CoPhaser recovers biologically accurate continuous cell-cycle phases across diverse RNA sequencing technologies, including highly heterogeneous situations such as development and cancer, without requiring prior knowledge of gene expression programs or cell-cycle states. In two cancer applications, CoPhaser reveals subtype-specific proliferation dynamics, identifying quiescent primitive states in relapsed pediatric acute myeloid leukemia and distinguishes proliferation-driven gene overexpression from constitutive overexpression to uncover potential robust therapeutic targets in triple-negative breast cancer. CoPhaser readily generalizes to other periodic systems, enabling the reconstruction of circadian clocks in the mouse aorta and the identification of cell-type and subtype-specific circadian differences. Furthermore, it maps continuous endometrial remodeling across the human menstrual cycle, uncovering distinct transcriptional dynamics between women with endometriosis and controls. Together, CoPhaser provides a versatile and interpretable framework for dissecting the interplay between cellular identity and biological cycles in single-cell data.

Keywords: Periodic Manifolds, scRNA-seq, Autoencoder, Biological cycles, Cell Cycle, Circadian Cycle, Menstrual Cycle

1 Introduction

The advent of single-cell genomics has accelerated the possibility of reconstructing molecularly defined biological state spaces at a fine cellular resolution [1]. An emerging property is that cells evolve dynamically along structured subspaces, commonly called manifolds, which have low dimensions compared to the high-dimensional molecular space [2]. Learning and representing such manifolds so that they are interpretable and biologically insightful is a major challenge, and has been successfully approached using variational autoencoders (VAEs) [3], notably to predict single-cell responses to perturbations[4, 5].

To be biologically relevant, these low-dimensional spaces should respect the topological structure of the underlying processes. Indeed, these processes live in subspaces with distinct shapes, with building blocks such as lines and trees, relevant in the context of cell differentiation or development, and also cycles which are common in molecular biology[6, 7]. Here, we view cell states as the Cartesian product of coordinates from multiple underlying processes, such as cell identity and the cell cycle. We will focus on modeling biological cycles while accounting for the heterogeneity brought by these other biological degrees of freedom.

Biological periodicity is widespread, with many important biological and cellular processes having a temporally periodic structure. Among the prime examples of biological cycles in mammals, both the cell division and circadian cycles are cell-autonomous periodic processes driven by core gene regulatory networks, including the periodic regulation of mRNA abundance through transcription and mRNA stability [8, 9]. These two cycles have periods in the range of one day, whereas other cycles can be significantly longer, such as the human menstrual cycle. Depending on the context, these cycles may tick only in subpopulations of cells (*e.g.* cycling tumor cells), be paused and eventually restart, or be permanently arrested.

Of particular relevance, these cycles manifest as gene expression profiles that are periodic with respect to a latent state variable, hereafter called the phase. These phases should be biologically interpretable, linking to molecularly defined events such as the G1/S transition in the cell cycle progression, or the peak activity of the BMAL1/CLOCK transcription factors during the circadian cycle. A major challenge is to learn the phase coordinates of individual cells from (sparse) scRNA-seq counts, especially in the presence of heterogeneity and confounding sources of variability. Indeed, cell-autonomous cycles are embedded within cells existing in diverse biological contexts, which we define as different cell types, metabolic conditions, or disease progression states, or in fact any discrete and continuous cell states leading to cell state variability that cannot be explained by the cyclic process itself. Thus, while many biological cycles are often highly conserved owing to their fundamental biological functions within cells, tissues and organisms, contexts can nevertheless influence their properties, for example by shifting the means and amplitudes of the gene expression profiles (Fig. 1A). Such modified gene expression profiles complicate the assignment and interpretation of cellular phases, since then there is no fixed reference.

The cell cycle is ubiquitous in development and regeneration, while it is also frequently misregulated in disease. Numerous methods have sought to estimate cell cycle states from scRNA-seq (reviewed in [10]). Such methods typically seek to either regress out cell-cycle related variability for downstream analyses [7, 11], or study gene expression dynamics [12]. Common models for cell cycle phase inference consist of autoencoders (AEs), of which Cyclum [13] and DeepCycle [14] are representatives. These AEs infer a latent phase describing cell-cycle progression and enables high-resolution reconstruction of transcriptional trajectories in deeply sequenced, actively proliferating,

and homogeneous systems. While successful in homogeneous populations of cells, these methods do not consider that the cell-cycle can be modified by the cellular context, limiting their applicability in heterogeneous situations such as development [15] or disease [16, 17].

While the abundance and detectability of cell cycle transcripts in mammalian cells is favorable for the phase inference task, other cycles are more challenging as the availability of sufficiently highly expressed periodic transcripts can be more scarce. This is notably the case for the circadian cycle, where phase assignment algorithms have been most successful in bulk [18, 19]. Nevertheless, the assignment of circadian phases in single cells is possible when the sequencing coverage is high [20, 21]. Notably, Tempo [20] implements a probabilistic model for circadian phase inference that uses variational inference. By modeling rhythmic gene expression using shared oscillatory waveforms across cells, Tempo can allocate phase coordinates in relatively homogeneous circadian systems. However, Tempo does not explicitly model cell-type or state-dependent modulation of circadian gene expression. However, how to approach circadian heterogeneity and contextual dependence at the single-cell level is still open, and can be approached in multiple ways, including the recent VAE SNOW [22]. SNOW decomposes scRNA-seq time-series data generically into time-dependent and time-independent components, facilitating joint batch correction and temporal analysis. SNOW handles arbitrary time-series designs and does not impose an explicit periodic topology on the latent space, and requires observed experimental time making it well suited to synchronized or externally driven time courses. Similarly, CellUntangler [23] is a VAE designed to disentangle different biological processes, including cycles. However, it does not register cells on a strict 1D cyclic manifold, and its decoder architecture is not designed to yield a phase definition that is comparable across different contexts.

Here, we formalize the idea that context-dependent periodic manifolds can be learned in an unsupervised manner, enabling the definition of phase coordinates that remain interpretable and comparable across cellular contexts. We develop a generic model, *CoPhaser* (for context-aware phaser), applicable to biological cycles and data modalities, without requiring external labels or prior knowledge of periodic gene profiles, while remaining flexible enough to support transfer learning from independently learned or calibrated cycles. With its flexible but explicit parameterization of the expected expression levels in function of the phase and context latent coordinates, CoPhaser consists of a biologically informed variational autoencoder with a structured latent space that explicitly separates cycle phases from cellular context, while controlling their mutual information. By modeling gene expression as context-modulated harmonic functions, CoPhaser captures flexible deformations of biological cycles.

We demonstrate that CoPhaser recovers biologically accurate continuous cell-cycle phases across diverse RNA sequencing technologies, including highly heterogeneous situations such as development and cancer, without requiring prior knowledge of gene expression programs or cell-cycle states. We applied CoPhaser to two cancer datasets, revealing subtype-specific proliferation dynamics, identifying quiescent primitive states in relapsed pediatric acute myeloid leukemia, and distinguishing proliferation-driven gene overexpression from constitutive overexpression to uncover robust therapeutic targets in triple-negative breast cancer. Furthermore, CoPhaser generalizes to other periodic systems, enabling the reconstruction of circadian clocks in the mouse aorta and the identification of cell-type– and subtype-specific circadian differences. Finally, we employed CoPhaser to map continuous endometrial remodeling across the human menstrual cycle, uncovering distinct transcriptional dynamics between women with endometriosis and controls.

2 Results

2.1 CoPhaser enables context-dependent learning of periodic manifolds

While many biological cycles are highly conserved, they are rarely fully invariant. Instead, their gene expression trajectories measured in single-cell RNA-sequencing (scRNA-seq) manifest as periodic manifolds that can be influenced by the underlying cellular context (Fig. 1A). To register those cycles across contexts, we allowed the manifolds to be shifted and scaled according to a context variable. To disentangle these periodic dynamics from cell-specific context, we developed CoPhaser, a variational autoencoder (VAE) with a structured latent space composed of an f-space encoding a periodic phase and a cell-context z-space (Fig. 1B). The model uses factorized variational posteriors consisting of a Power Spherical distribution [24] for the phase and a multivariate Gaussian for the non-periodic latent variable z. The decoders, yielding the expected log-means of the measured scRNA-seq counts for each cell, consist of a Fourier series to capture the periodic structure and standard decoders for the context-dependent mean and amplitude corrections (Fig. 1B and Methods 3.1).

The encoder is informed with two different sets of genes: a set of seed rhythmic genes (X_f) determines the f-space, whereas the z-space is inferred from a broader set that includes highly variable genes (Methods 3.8.1).

To ensure the identifiability of the f and z coordinates and avoid that cyclic variability is attributed to the more general z-space, we trained a Mutual Information Neural Estimator (MINE) [25] to explicitly penalize dependencies between the two latent spaces (Methods 3.3).

Formally, for each cell c and gene g, the log-expected gene counts $\log \mathbb{E}[X_{cg}]$ are modeled as the sum of three contributions: a periodic component $F_g(\varphi_c)$, a baseline expression μ_{cg}, and a context-dependent mean shift $\Delta\mu_g(z_c)$. In addition, we allow for a context-dependent cycling amplitude $\lambda(z_c)$ that scales the periodic component $F_g(\varphi_c)$ (Fig. 1B). This parameterization enables the model to learn flexible, context-dependent deformations of the periodic manifold, while preserving a coherent and shared definition of the phase (φ_c) across cells. The model is trained using a custom loss based on the Evidence Lower Bound (ELBO), with observation noise modeled by a Negative Binomial distribution (Methods 3.4).

2.2 Accurate recovery of phases and context-dependent modulations in simulated data

To validate our approach, we simulated data recapitulating cell-cycle gene expression counts using biologically plausible Fourier coefficients (Methods 3.7). We also introduced a one-dimensional cellular context that linearly modulates both the mean gene expression levels and the cycle amplitude. CoPhaser correctly decomposed these sources of variations: the inferred z-space accurately recovers the contextual gradient (Fig. 1C), while the f-space faithfully encodes the ground-truth (GT) cell-cycle phase (Fig. 1D,E). Consequently, the inferred gene expression profiles (Fig. 1F), cycle amplitudes (Fig. 1G), and mean shifts $\Delta\mu$ closely match the simulated GT, resulting in accurate reconstruction of gene expression even in the presence of strong contextual distortions (Fig. 1I,J).

To assess the roles of our model components, we performed ablation studies. We found that removing the contextual z-space resulted in a model that fails to distinguish between cyclic and contextual variance, producing severe artifacts in the inferred phases (Fig. S1A-C). This limitation is also observed in standard Principal Component Analysis (PCA), which similarly lacks the

capacity to model context-dependent deformations and thus confounds the two sources of variation (Fig. S1D-F). Furthermore, controlling the mutual information (I) between the f- and z-spaces proved to be critical, as removing this constraint led to information leakage and degraded separation of the two latent representations (Fig. S1G–K).

2.3 Context-aware inference of continuous cell-cycle phases across datasets and technologies

We validated CoPhaser on cell-cycle analysis in scRNA-seq datasets of increasing complexity. First, using scEU-seq data from the human RPE1-FUCCI cell line [8], we observed a monotonic relationship between the inferred phase and the FUCCI-derived cell cycle phases (Fig. 2A). Note that this monotonic, but non-identical, relationship between FUCCI- and mRNA- derived phases is expected as the two methods are reporting on biochemical processes occurring on different regulatory layers. In fact, the FUCCI-based phase appears less informative in the interval inferred phase $\in (-0.5\pi, 0)$, resulting in an apparent compression and discontinuity in gene expression profiles at FUCCI phase $\sim -0.5\pi$, whereas CoPhaser infers a smooth progression, indicating that mRNA contains biologically relevant information on cell-cycle progression when the FUCCI signal is not sensitive (Fig. 2B). The latent context z-space captures the experimental conditions in the original study (Fig. S2A), however, in this case, without inducing substantial differences in the predicted gene profiles (Fig. S2B). This is consistent with the high correlation between CoPhaser and PCA-based phase inference on this dataset (Fig. S2C,D). Indeed, PCA-based phases are reliable when the context-induced variability is limited [26]. Moreover, all methods capture the expected monotonic increase in the mean number of unique molecular identifiers (UMIs) per cell over the cycle (Fig. S2D). Finally, the inferred peak phases of cell-cycle genes show a good agreement with independent annotations from Cyclebase [27] (Fig. S2E).

We further evaluated the robustness of the latent representations using two batches of RPE1 scRNA-seq data (10x Genomics) [28]. While these experiments were performed in the same conditions, the z-space was sensitive enough to separate the two batches (Fig. 2C), reflecting that the two batches slightly differed in their proliferation status.

The inferred phases recapitulate canonical cell cycle gene expression patterns (Fig. 2D) with the correct temporal ordering of peak phases (Fig. 2E). Consistent with prior knowledge, the UMI count increases monotonically across the cycle (Fig. 2F) and sharply drops following mitosis. Furthermore, histone transcripts, which were not part of the rhythmic input gene set, consistently and sharply peak during the inferred S-phase alongside known markers, providing independent validation of the phase inference (Fig. 2G).

A potential limitation of our approach is the requirement for an input list of cycling genes. However, because Fourier coefficients are inferred for all context input genes, we hypothesized that this list could be refined iteratively for cycles that induce strong gene expression signatures. We tested this strategy on the RPE1 dataset by initializing the model with a minimal set of just two cycling genes, *TOP2A* and *PCNA*, while using the 2000 most variable genes as context. This minimal setup already yielded a good approximation of the cell cycle, with a circular correlation of 0.47 relative to the reference (Fig. S3A, B). We note however that despite the inferred gene peaks following the correct temporal ordering (Fig. S3C), the inferred amplitude of *PCNA* became exaggerated (Fig. S3B), indicating overfitting, particularly in the low-expression interval $(-\pi, 0)$.

This overfitting manifested in artifacts in both the reconstructed UMI counts and histone expression profiles (Fig. S3D,E).

To test how performance improves with more cycling genes, we defined a gene selection score combining mean expression and inferred amplitude to identify robust markers (Methods 3.8.1). We selected the twenty most informative genes and retrained the model. Using this refined gene set, the inferred phases closely match those obtained with the full canonical 98-gene list, without visible artifacts and faithfully recapitulating the expected biological cell cycle gene expression dynamics (Fig. S3F-J).

Finally, we evaluated our method on a complex and highly challenging dataset of mouse embryonic cells profiled by VASA-seq [15], which was provided with independent S-phase annotations. Despite the rich cellular heterogeneity and numerous proliferative cell types, the inferred f-space cleanly separates S-phase cells from other populations (Fig. 2H). As expected, the dominant source of variation in the z-space corresponded to cell identity (Fig. 2I), closely resembling the original embedding [15]. Crucially, owing to the explicit separation of latent spaces, the z-space is largely devoid of cell cycle–related variation (Fig. S4A). Histone expression increases and peaks during S-phase, followed by a decline as G2 genes peak, in agreement with established cell cycle biology (Fig. S4B, C). However, significant mean shifts are visible between cell types, with genes peaking in G1 or at the G1 to S transition being more affected than the others (Fig. S4D). As extreme examples, *Cdca7*, which peaks at the S-phase transition, displays high mean shifts between cell types, while *Kif11*, peaking in G2, displays none.

To systematically assess performance, we computed the Jensen–Shannon (JS) divergence between the inferred phase distributions of annotated S-phase cells and non–S-phase cells. A perfect separation yields a score of 1, whereas random assignments yield 0. Here, CoPhaser outperformed all alternatives, including PCA projections, whether mean-centered per cell type or globally (Fig. 2J). The assignments from Seurat perform poorly, with a separation close to random (Fig. S4G), and the leading PCs yield only moderate separation (Fig. S4H). In contrast, our approach achieves a clear separation with limited overlap confined to cell cycle transitions (Fig. S4I), underscoring the advantages of a structured latent space and explicit modeling of context-dependent cycle deformations.

2.4 Analysis of subtype-specific proliferation rates and phase-dependent versus constitutive gene dysregulation in breast cancer

Having validated our model on controlled and heterogeneous non-tumoral datasets, we next investigated its performance in highly variable tumoral contexts. Tumors often contain a large fraction of quiescent cells arrested in G0/G1, which can obscure cell cycle inference. To address this, we augmented our model with a Bernoulli VAE [29] that generically classifies cells as cycling or non-cycling, thereby reducing the influence of quiescent cells on phase inference (Methods 3.2). Specifically, the Bernoulli encoder takes as inputs all cyclic genes and the latent z, and the learned indicator functions then multiply the Fourier series so that cells classified as non-cycling are not influenced by the periodic gene expression.

We first analyzed a breast cancer dataset comprising 69 surgical tissue specimens from 55 patients [30], focusing exclusively on tumoral epithelial cells. We used this dataset, for which the original authors provided a cycling annotation, to validate our cycling classifier. Cells identified as non-cycling were effectively restricted to the G0/G1 interval (Fig. S5A). Our classification

was less strict than the original annotation (Fig. S5B, C); this difference likely stems from the original annotation that relies on *Mki67* mRNA expression, which is very low during G1, potentially leading to false negatives in the annotation from the original authors. Feature attribution via Integrated Gradients [31] between cells in G1 and cells classified as non-cycling, displayed shared negative markers between subtypes consisting mostly of cell cycle regulators, as well as fewer positive markers, among which epithelial markers such as keratins (Fig. S5D–F).

Analyzing the predicted cycling logits for cells inferred to be in G1 or G0, we observed a striking divergence between breast cancer subtypes, with triple-negative (TN) tumors assigned with a markedly higher probability of cycling (Fig. S5). To further characterize the signal driving this classification, we compared the gene expression signatures of G0/G1 cells classified as cycling and non-cycling. This analysis revealed an upregulation of cell cycle–associated genes, most notably the proliferation marker *BIRC5* (Fig. S5H, I), suggesting that the classifier identifies a G0-like state.

Having validated our classifier, we investigated the predictions of the other components of our model. In this dataset, the local structure of the z-space is dominated by inter-patient variability (Fig. S6A), while the global organization preserves tumor subtypes (Fig. 3A). As observed previously, the inferred peaking times of cell cycle genes follow the expected order (Fig. 3B), and *H4C3* peaks with S-phase genes, exhibiting similar profiles across subtypes (Fig. 3C). Gene expression profiles are well aligned between subtypes (Fig. S6B), and the mean UMI count increases monotonically along the inferred cell cycle (Fig. S6C).

By quantifying the fraction of proliferating tumoral cells, we identified significant differences between subtypes, with triple-negative breast cancer (TN) exhibiting the highest proliferation rates (Fig. 3D). Analysis of inferred mean shifts further revealed that TN cells display elevated expression of *CDC20* (Fig. 3E), consistent with prior reports [32]. However, since *CDC20* is not constitutively overexpressed but rather upregulated specifically during the S/G2/M phase, the predicted expression dynamics suggest limited efficacy of therapies targeting *CDC20* in slow-dividing populations such as persistent cancer cells [16]. In contrast, CoPhaser identifies genes such as *KRT14* as being constitutively overexpressed in TN cells (Fig. 3F), consistent with their role as known markers of TN breast cancer [33].

An advantage of our model is that the inferred mean shifts are decoupled from differences in proliferation rates. For example, pseudobulk expression of *UBE2C* suggests higher expression in TN tumors (Fig. S6D). However, inspection of its inferred expression profile along the cell cycle (Fig. S6E) reveals dynamics similar to other subtypes, indicating that the apparent overexpression arises primarily from the increased fraction of cycling cells rather than a true context-specific upregulation.

2.5 CoPhaser reveals persistence-driven relapse in pediatric AML

We next analyzed a longitudinal pediatric acute myeloid leukemia (pAML) dataset comprising scRNA-seq data from 28 patients sampled at diagnosis, remission, and relapse [34]. In this dataset, the contextual z-space was primarily structured by tumor subtype and patient ID (Fig. 3G, Fig. S6F), with non-tumoral cells clustering together (Fig. S6G). This organization highlights the dominant variance induced by tumor driver mutations, which exceeds that associated with the diverse cell types present in AML (Fig. S6H). Despite this heterogeneity, the inferred cell cycle is robust: the peaking times of canonical genes follow the expected temporal progression, with a

notably prolonged G1 phase (Fig. 3H), and the histone gene *H4C3* consistently peaking alongside S-phase genes across all subtypes (Fig. 3I). Moreover, gene expression profiles are well aligned across subtypes, differing mainly by context-dependent shifts in the (log) means (Fig. S6I), and the mean number of UMIs per cell increases monotonically over the inferred cell cycle (Fig. S6J).

We next quantified the fraction of actively proliferating malignant cells, defined as cells with an inferred phase in $(\pi/4, \pi)$ and predicted to cycle, corresponding to non-G0/G1 cells, for each malignant cell type. Pre-B cells, late erythrocytes, common lymphoid progenitors (CLP), granulocyte–monocyte progenitors (GMP), and progenitor cells exhibited the highest proliferative fractions (Fig. 3J). Notably, the fraction of cycling cells within each cell type did not change significantly between diagnosis and relapse across patients. In contrast, examining the composition of proliferating malignant cells revealed a marked relative decrease in GMP cells and a concomitant increase in progenitor cells at relapse (Fig. 3K). Analysis of non-cycling (G0/G1) malignant/tumor cells further confirmed a global shift toward more primitive cell states following treatment (Fig. 3L), consistent with the original study [34]. Importantly, this shift appears to be driven by differential persistence rather than elevated proliferation rates, as these primitive populations do not cycle more frequently (Fig. 3D). Hematopoietic stem cells (HSCs) exemplify this behavior: although they rarely proliferate, they constitute a substantial fraction of G0/G1 cells, reflecting their long lifetime.

2.6 CoPhaser accurately recovers context-modulated circadian oscillations in the mouse aorta

We next evaluated the ability of our model to capture circadian rhythms in gene expression. To this end, we analyzed a high-coverage scRNA-seq dataset from mouse aorta sampled every 6 h over a 24 h period [20], namely at external Zeitgeber (ZT) times ZT0, 6, 12, 18. The performance of phase assignment in the circadian field is often assessed with the median absolute deviation (MAD; Methods 3.10.1) between predicted phase and external ZT. Here, the inferred phases closely tracked the ZT, with a MAD of 1.67 h across all cell types (Fig. 4A), with the lowest MAD for the smooth muscle cells (SMCs) and the highest for endothelial cells (Fig. 4B).

The Fourier coefficients inferred by the model recapitulate the canonical circadian clock structure (Fig. 4C) [19], and the inferred phases reveal rhythmic gene expression profiles across cell types (Fig. 4D). Notably, the overall MAD is lower than that obtained with *Tempo* [20], with particularly pronounced improvements for fibroblasts and macrophages (Table 1). These cell types are represented by relatively few cells but exhibit strong clock gene expression (Fig. S7A). While minimizing MAD is not an explicit modeling objective, given that we expect genuine biological desynchrony in circadian phases across individual cells, achieving a lower MAD without using external time during training suggests that CoPhaser more accurately captures the underlying clock structure. We attribute this improvement to the shared optimization across cell types, which stabilizes parameter estimation, especially for sparsely represented populations.

In parallel, the z-space faithfully captures cellular identity, separating the major vascular and immune cell types (Fig. 4E). Analyzing the inferred context-dependent shifts of the means reveals a cell-type-specific regulation of clock-related genes. In particular, *Npas2* and *Hlf* are predicted to be expressed at very low levels in endothelial cells (Fig. 4F). Inspection of reconstructed expression profiles confirms that both genes are nearly silent throughout the circadian cycle in endothelial cells (Fig. 4G), a pattern that is also evident at the pseudobulk level (Fig. S7B), while both genes

display robust oscillations in other cell types. In contrast, *Nfil3* expression is largely restricted to smooth muscle cells (SMCs) (Fig. 4F,G). Analysis of the inferred amplitude scaling factor (λ) indicates that global circadian rhythmicity is strongest in smooth muscle cells and fibroblasts, while appearing slightly dampened in endothelial cells and macrophages (Fig. S7C).

A key strength of CoPhaser is that it does not require prior clustering by cell type, but instead resolves cell heterogeneity directly from the data. Focusing on SMCs, the model resolves three distinct subpopulations characterized by differential expression of *Rgs5* (Fig. 4H, Fig. S7D). Using established marker genes [35], we identify the *Rgs5*-high and -intermediate clusters as originating from the descending thoracic aorta, whereas the *Rgs5*-low cluster corresponds to cells from the aortic arch (Fig. S7E). Beyond this anatomical stratification, the model predicts subtle, yet consistent differences in circadian gene regulation between these subpopulations, notably for *Nfil3*, *Per1* and *Nampt* (Fig. S7F). These differences are apparent both in the reconstructed single-cell expression profiles (Fig. 4I) and at the pseudobulk level (Fig. S7G), highlighting our model's ability to resolve fine-grained context-dependent circadian regulation within closely related cell populations.

Table 1 Median Average Deviation (MAD) in hours in function of the cell types. Mean over 5 independent runs, with the standard deviation indicated in parentheses. Our model was trained both with the same quality check filters as Tempo, and with an additional minimal number of 5000 UMIs per cell.

Method	Gene set	SMCs	Fibroblasts	Endothelial cells	Macrophages
Tempo	All	1.88 (0.03)	2.53 (0.36)	3.39 (0.4)	2.99 (0.01)
Our Method	Core clock	1.82 (0.03)	2.30 (0.04)	3.48 (0.29)	2.50 (0.06)
Our Method (high-quality cells)	Core clock	**1.73 (0.10)**	**2.01 (0.04)**	**2.84 (0.07)**	**2.29 (0.14)**
Random	-	5.91 (0.02)	5.79 (0.07)	5.17 (0.28)	5.34 (0.06)

2.7 CoPhaser reveals menstrual cycle gene expression dynamics and endometriosis-associated alterations

Finally, we applied CoPhaser to resolve the continuous dynamics of the human menstrual cycle. Throughout this cycle, endometrial stromal cells from the functionalis layer undergo extensive transcriptional remodeling under the influence of ovarian hormones [36], transitioning from proliferative endometrial stromal (eStromal) cells to decidualized stromal (dStromal) cells characteristic of the secretory phase. To infer a minimal cycling gene list, we used the I between the menstrual cycle phase and the gene expression (Method 3.8) Consistent with this biology, the phases inferred by CoPhaser closely match the annotated menstrual stages (Fig. 5A), with cells from the same subject assigned to highly similar phases (circular standard deviation 0.1π; Fig. S8A).

The different stromal cell states, corresponding to distinct stages of the menstrual cycle, follow the expected temporal progression (Fig. 5B) and are well integrated in the z-space. This indicates that the dominant source of gene expression variation among stromal cells can be summarized by a single one-dimensional phase, with the notable exception of an additional variance component associated with the cell cycle, specifically observed in cells annotated as eStromal proliferating (Fig. 5C). The curated set of rhythmic input genes (Methods 3.8) spans all phases of the menstrual cycle; however, CoPhaser additionally identifies many rhythmic genes among the

broader set of context genes (Fig. S8B). The reconstructed expression profiles reveal diverse temporal dynamics, ranging from sharp transitions—such as *CXCL14*, which is abruptly induced in mid–secretory phase, likely corresponding to the opening of the window of implantation [37], to smoother transitions, such as that observed for *MMP11* (Fig. 5E).

The original study described a distinct population of eStromal cells marked by high expression of matrix metalloproteinases (MMPs). CoPhaser predicts that this population is largely restricted to the menstrual phase, with minimal overlap with other eStromal cells (Fig. 5A). This prediction is further supported by the observation that few subjects harbor both populations simultaneously (Fig. 5F). We then examined differences between patients with endometriosis and control subjects. Because each individual spans only a narrow range of phases (Fig. S8B), and only a limited number of subjects cover each phase of the cycle (Fig. S8C), statistical power to detect deregulated genes across many subjects is limited. Among genes highlighted originally, *GREB1* and *DKK1* show modest differences limited to the late secretory phase. In contrast, analysis of inferred mean shifts revealed that *PAEP* is consistently more highly expressed in subjects with endometriosis compared to controls (Fig. 5G).

Finally, we leveraged the versatility of CoPhaser to analyze cell cycle dynamics occurring within the endometrium across the menstrual cycle. The model recovers the expected peaking phases and expression profiles of canonical cell cycle genes (Fig. S8D, E). The cell types identified as proliferative are consistent with the original annotations, with "Cycling" cells and "Cycling eStromal" cells exhibiting the highest proliferative activity (Fig. S8F). Notably, a subtype of endometrial perivascular cells (ePVs), annotated as "group 1b" based on *STC2* expression, also appears highly proliferative. As expected, the majority of proliferation occurs during the proliferative phase of the menstrual cycle (Fig. S8G). Across these proliferative cell types, we did not detect major differences in cell cycle gene expression between endometriosis patients and controls (Fig. S8H).

3 Methods

3.1 Model architecture

The model comprises two parallel encoders and two parallel decoders, designed to separate periodic dynamics from context-dependent variability. Two distinct gene sets are used as encoder inputs: a set of seed rhythmic genes (X_f) informing the rhythmic encoder, and a broader set of highly variable genes (X_z), which includes the rhythmic genes, informing the context encoder.

Rhythmic encoder.

The rhythmic encoder processes the seed rhythmic genes X_f. Raw counts are log-normalized as

$$\log\left(1 + \frac{X_{cg}}{N_c} \cdot 10^4\right) - b,$$

where X_{cg} denotes the count of gene g in cell c, N_c is the library size of cell c, and b is a baseline correction factor. The normalized expression is then projected through a three-layer fully connected neural network (Linear–Linear–Linear) into a two-dimensional latent space. Although functionally equivalent in expressivity to a single linear layer, this architecture has been shown to improve generalization when trained by gradient descent [38].

The resulting two-dimensional vector is normalized to unit length and used as the mean direction parameter $\mu_\varphi(X_f)$ of a Power Spherical posterior

$$q(\varphi \mid x) = \text{PowerSpherical}(\mu_\varphi(X_f), \kappa(X_f)),$$

defined on the unit circle [24]. The concentration parameter $\kappa(X_f)$ is inferred via a separate neural network and informs on the uncertainty of the inferred phase.

Context encoder.

The context encoder captures non-periodic sources of variation, such as cell type or technical effects, and closely follows a standard VAE architecture used in scVI [3]. Specifically, this encoder maps the variable genes X_z to a latent variable $z \in \mathbb{R}^{d_z}$ using a multivariate normal posterior

$$q(z \mid x) = \mathcal{N}(\mu_z(x), \text{diag}(\sigma_z^2(x))),$$

where $\mu_z(x)$ and $\sigma_z(x)$ are produced by a two-layer fully connected neural network with GELU activations [39]. The dimensionality d_z is a tunable hyperparameter; we typically use $d_z = 10$ for complex biological contexts and $d_z = 2$ for simpler settings.

Decoders and generative model.

Three decoders are trained jointly from the latent spaces to reconstruct gene expression: (i) a rhythmic Fourier decoder capturing periodic expression patterns (F_g), (ii) a context-dependent mean-shift decoder ($\Delta\mu$), and (iii) a context-dependent amplitude decoder (λ). The rhythmic component is decoded for all genes used by the model, not only the seed rhythmic genes, and is represented using a Fourier series with typically three harmonics ($N_{\text{harm}} = 3$) unless specified otherwise. The context-dependent decoders mirror the architecture of the context encoder.

We adopt a parametrized approach so that the expected gene expression is modeled as

$$\mathbb{E}[X_{cg}] = N_c \cdot \exp(\mu_g + \Delta\mu_{cg} + b_c \cdot \lambda_c \cdot F_g(\varphi_c)), \tag{1}$$

where N_c is the library size, μ_g is the gene-specific log-mean expression across the dataset, $\Delta\mu_{cg}$ is a context-dependent mean shift, and λ_c is a context-dependent amplitude scaling factor derived from z_c. The rhythmic component is given by

$$F_g(\varphi) = \sum_{k=1}^{3} [a_{gk} \cos(k\varphi) + b_{gk} \sin(k\varphi)],$$

capturing the periodic expression pattern of gene g.

Context-dependent learning of corrections to the gene expression mean.

To explicitly correct for context-dependent deformations of the periodic manifold, and for situations where the coverage of phases does not cover the cycle uniformly, we additionally decode a context-dependent shift in the mean $\Delta\mu_{cg}^{\text{enc}}$ for the rhythmic encoder after an initial number of training epochs. This allows the model to adjust baseline expression levels and amplitudes prior to phase

inference. The normalized counts used as input to the rhythmic encoder are then given by

$$\text{NormalizedCounts}_{cg} = \frac{\log\left(1 + \frac{X_{cg}}{N_c \cdot \exp(\Delta\mu_{cg}^{\text{enc}})} \cdot 10^4\right)}{\lambda_c} - b, \tag{2}$$

where $\Delta\mu_{cg}^{\text{enc}}$ and λ_c are both derived from the context latent variable z_c.

3.2 Cycling classifier

For datasets containing substantial populations of non-cycling cells, such as adult tissues with G0-arrested cells, we include a cycling classifier to infer the probability that a cell is actively cycling. This classifier is implemented as a Relaxed Bernoulli (Gumbel–Softmax) encoder [40, 41], and is jointly trained with the rest of the model. Thus, while we illustrate the idea with G0 cells, this is generic and the nature of the G0-like state will depend on the cycle being considered. This encoder takes as input the gene expression X and the context latent variable z, and outputs a probability b_c of being in the cycling state.

During training, samples are drawn from the Relaxed Bernoulli distribution to enable differentiable optimization. The temperature parameter is annealed exponentially over training to approximate a discrete Bernoulli distribution, allowing the model to make increasingly sharp cycling versus G0-like assignments.

When the prior probability of cycling is set to a value less than one, the expected expression is modeled as

$$\mathbb{E}[X_{cg}] = N_c \cdot \exp(\mu_g + \Delta\mu_{cg} + b_c \cdot \lambda_c \cdot F_g(\varphi_c) + (1 - b_c) \cdot F_g(\varphi_{\text{G1}}))\,, \tag{3}$$

where $b_c \in \{0, 1\}$ denotes the cycling status of cell c, and φ_{G1} is a fixed reference phase corresponding to the G1 state when modeling the cell cycle (hence the notation), accounting for the transcriptional similarity between G0 and G1 cells. For other cycles, this reference can be chosen to represent arrested cycling, as for example, in a knock-out model of the circadian clock.

Although not strictly required, for datasets with a high fraction of non-cycling cells, we initialized the rhythmic decoder using Fourier coefficients inferred from the RPE1 dataset [28], and fixed $\varphi_{\text{G1}} = -1.5$ radians.

3.3 Mutual information neural estimation (MINE) network

To encourage the rhythmic phase φ and the contextual latent variable z to capture distinct sources of variation, we employ a mutual information neural estimator (MINE) [25]. The MINE network estimates the mutual information $I(\varphi; z)$ by discriminating between samples drawn from the joint distribution $P(\varphi, z)$ and samples drawn from the product of marginals $P(\varphi)P(z)$, obtained by randomly permuting one variable within each minibatch.

The estimated mutual information is incorporated into the training objective as a penalty term, thereby discouraging statistical dependence between the phase and context representations. This regularization promotes disentangled latent spaces in which z captures non-periodic contextual variation while φ encodes cyclic dynamics.

In practice, the model is often able to capture contextual information more easily than a periodic structure. To prevent the context encoder from absorbing the cycle signal, we explicitly control

gradient flow during MINE optimization. Specifically, when computing the mutual information loss, the projected periodic variable in f-space is detached from the computational graph, such that gradients from the MINE loss update the context encoder but not the rhythmic encoder. This asymmetric gradient application encourages z to relinquish cycle-related information while preserving the ability of the rhythmic encoder to capture the periodic manifold.

3.4 Loss function

The model is trained by maximizing the evidence lower bound (ELBO), augmented with additional regularization terms to promote biologically meaningful and stable solutions. The objective to be minimized is then

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = & -\mathbb{E}_q[\log p(x \mid z, \varphi, b)] \\ & + \beta_z \, D_{\text{KL}}(q(z \mid x) \,\|\, p(z)) \\ & + \beta_\varphi \, D_{\text{KL}}(q(\varphi \mid x) \,\|\, p(\varphi)) \\ & + \beta_b \, D_{\text{KL}}(q(b \mid x) \,\|\, p(b)) \,.\end{aligned} \tag{4}$$

The reconstruction term $\log p(x \mid \cdot)$ is modeled using a negative binomial likelihood for count data. The Kullback–Leibler (KL) divergence terms regularize the approximate posteriors toward their respective priors: $p(z)$ is a standard multivariate normal distribution $\mathcal{N}(0, I)$, $p(\varphi)$ is a uniform distribution on the circle, and $p(b)$ is a Bernoulli prior representing cycling status. The weighting coefficients β_z, β_φ, and β_b implement a β-VAE framework [42]. In practice, β_φ was set to 0.1 to prevent premature collapse of the phase posterior to a uniform distribution, which would hinder learning of cyclic structure, while β_b was set to 10 to encourage the exclusion of non-cycling (G0-like) cells, which are transcriptionally similar to G1-like cells.

To further disentangle periodic dynamics from contextual variation, we include a mutual information penalty between the phase variable φ and the contextual latent variable z:

$$\mathcal{L}_{\text{MINE}} = \lambda_{\text{MINE}} \, I(\varphi; z), \tag{5}$$

where $I(\varphi; z)$ is estimated using a mutual information neural estimator (MINE), and λ_{MINE} controls the strength of the penalty. This term encourages φ to capture periodic structure while z encodes non-periodic contextual effects. Complete independence is not enforced, as some biological contexts (for example, cell type) are intrinsically associated with cycling propensity.

In the absence of additional constraints, the model can converge to degenerate solutions in which inferred phases collapse to a narrow range. To prevent this behavior, we introduce a circular entropy regularization term that encourages a uniform phase distribution early during the training. The phase distribution is estimated using a differentiable circular histogram with N_{bins} bins. Let c_k denote the center of bin k and $w = 2\pi/N_{\text{bins}}$ the bin width. For a batch of phases $\{\varphi_i\}_{i=1}^{B}$, the contribution of phase φ_i to bin k is computed using a Gaussian kernel on the circular distance:

$$K_{ik} = \exp\left(-\frac{1}{2}\left(\frac{d(\varphi_i, c_k)}{w}\right)^2\right), \tag{6}$$

where $d(\varphi_i, c_k)$ denotes the shortest angular distance between φ_i and c_k. The normalized probability mass of bin k is

$$p_k = \frac{\sum_{i=1}^{B} K_{ik}}{\sum_{j=1}^{N_{\text{bins}}} \sum_{i=1}^{B} K_{ij}}. \tag{7}$$

The circular entropy is then defined as $H(\varphi) = -\sum_{k=1}^{N_{\text{bins}}} p_k \log p_k$, yielding the loss term

$$\mathcal{L}_{\text{entropy}} = -\lambda_{\text{entropy}} H(\varphi), \tag{8}$$

where λ_{entropy} controls the strength of the regularization. Importantly, to avoid over-constraining the phase distribution, this term is exponentially annealed during training (Methods 3.5). Its purpose is to route the optimization towards biologically relevant solutions early during the training.

Finally, to stabilize the geometry of the periodic latent space, we constrain the two-dimensional output u of the rhythmic encoder (prior to normalization onto the unit circle) to lie close to the unit circle. This is achieved by penalizing both the mean squared deviation of the radius from unity and its variance:

$$\mathcal{L}_{\text{radial}} = \lambda_{\text{radial}} \left(\mathbb{E}\left[(\|u\| - 1)^2\right] + \text{Var}(\|u\|) \right). \tag{9}$$

3.5 Training procedure

The model is trained for 200 epochs using the Adam optimizer with a learning rate of 0.01. To promote stable convergence, we adopt a staged training strategy. As described above, the weight of the circular entropy loss decays exponentially according to

$$\exp\left(-\frac{\max(\text{epoch}, 30) - 30}{100} \right) \times \text{entropy_weight_factor}, \tag{10}$$

where `entropy_weight_factor` is a tunable hyperparameter.

To further mitigate training instabilities, corrections for the mean and amplitude in the rhythmic encoder ($\Delta\mu_{cg}^{\text{enc}}$) are enabled only after epoch 50, and the rhythmic amplitude parameter (λ) is optimized only from epoch 100 onward. For the cycling classifier, the temperature of the Relaxed Bernoulli distribution is annealed from a high initial value to a low final value over the course of training.

The mutual information neural estimator (MINE) is trained in parallel with the variational autoencoder (VAE), also using the Adam optimizer, to continuously estimate and minimize the mutual information between the dynamically evolving z- and f-spaces.

Although the model can be trained from random initialization, and was trained as such for cell-cycle datasets with low fraction of G0 cells and for menstrual cycle data, we found that non-random initialization of the Fourier coefficients was beneficial for more challenging settings. These include contexts with a low fraction of cycling cells and circadian datasets, in which only a small number of lowly expressed genes exhibit oscillatory behavior. In these cases, prior initialization was not strictly required but substantially reduced the likelihood of convergence to biologically implausible solutions, such as aberrant inferred peaking phases or flat histone expression profiles.

Initialization of Fourier coefficients when specified were taken from the RPE1 dataset from [28] for cell-cycle analyses, and from [19] for circadian analyses, using only the genes shown in Fig. 1B. The Fourier coefficients were held fixed during the initial training epochs and subsequently unfrozen to allow adaptation to the target dataset.

3.6 Phase definition

Because phase inference is invariant to global phase rotations and phase orientation, an external convention is required to fix both the origin and the direction of the inferred phase.

For the cell cycle, the model automatically determines the phase orientation by enforcing the expected ordering of peak expression of canonical cell-cycle genes, such that genes peaking in G1 precede those peaking in S phase, and so on. We fixed the origin in this work such that $-\pi$ and π correspond to the M phase. This alignment was achieved using the mean UMI count per cell, which exhibits a characteristic decrease during mitosis.

For the circadian cycle, the indeterminacy was resolved using the external sampling time, and for the menstrual cycle, we fixed the phase using the annotated menstrual stages.

3.7 Simulations

To simulate biologically plausible gene expression profiles for the cell cycle, we used the Fourier coefficients of the rhythmic gene set fitted on the two batches of RPE1 scRNA-seq from [28]. We then drew a phase and context variable ranging from 0 to 1, from uniform distributions for each of the 10,000 cells. For one third of the rhythmic genes, we drew a maximal mean shift from a uniform distribution ranging from -5 to 5 (log), and shifted the mean gene expression of each cell based on the formula

$$\mu_{gc} = F_g(\varphi_c) + \Delta\mu_{max,g} \cdot z_c \tag{11}$$

with μ_{gc} the expected log fraction of count of gene g for the cell c, $F_g(\varphi_c)$ the Fourier coefficient of the gene g evaluated at the phase φ_c, $\Delta\mu_{max,g}$ the maximal mean shift for the gene g and z_c the context variable for the cell c.

Furthermore, we scaled the Fourier coefficient by an amplitude factor ranging from 0.8 to 1.7 linearly based on the context value. We then simulated 2000 context variables, with randomly drawn mean log fraction of counts, and maximal fold changes randomly drawn between -3 and 3, which modifies the expected log fraction by scaling them linearly between their maximal fold change and their basal expression, according to the context variable. We then transformed these expected log fractions into counts using a negative binomial noise with as mean the expected fractions multiplied by a target of 10,000 counts per cell, and a dispersion parameter of 0.1.

3.8 Gene sets selection

The model requires two gene sets: a rhythmic gene set, known to oscillate with the cycle of interest, and a context gene set, providing a context sufficient to recapitulate the different populations of cells found in the dataset. For the context genes, we always used the 2000 most variable genes in the dataset using the function *highly_variable_genes* from *Scanpy* with the flavor *Seurat* [43]. For the rhythmic gene set we tailored the list for each of the cycles: for the cell cycle, we use a list of 98 cell cycle genes similar to the widely used list from [44], for the circadian cycle, we used a

list of 25 core clock genes created by [20], and for the menstrual cycle, we used a list of 93 genes, obtained by selecting the 3000 most variable genes in the stromal cells, and taking the genes with an I greater than 0.1 between their counts and the menstrual cycle state using the *Sklearn* function *mutual_info_classif* [45]. All the rhythmic gene lists can be found in Supplementary Table 1.

3.8.1 Rhythmic gene selection from fitted context genes

To select the genes added to the rhythmic gene set using the two-gene fit (See Results 2.3), we sorted the context genes according to their mean fraction of counts multiplied by the exponential of the fitted peak-to-peak amplitude.

3.9 CoPhaser inference on multiple datasets

3.9.1 RPE1 cells from Battich *et al.*

This dataset of RPE1 cells was sequenced using single-cell EU-labeled RNA sequencing, which is based on CEL-Seq2 [8]. We used the original mapping and removed cells with less than 5% of unspliced counts, more than 70% of unspliced counts, and less than 2,000 UMIs per cell. 2,305 cells passed the quality check (QC) cut-offs and were kept for the training of our model, which was performed from scratch.

3.9.2 RPE1 cells from Lederer *et al.*

For this dataset [28], we mapped raw sequencing reads to the reference genome using *Cell Ranger* [46] (10x Genomics). Spliced and unspliced transcript counts were then extracted and demultiplexed using the *Velocyto* pipeline [47]. We then removed cells with less than 10% of unspliced counts, more than 42% of unspliced counts, fewer than 4,000, and more than 40,000 UMIs per cell. 15,237 cells passed the QC cut-offs and were kept for the training of the model, which was performed from scratch.

3.9.3 Mouse embryonic cells from Salmen *et al.*

For this dataset [15], raw sequencing reads were mapped to the reference genome using *Cell Ranger* [46] (10x Genomics). Spliced and unspliced transcript counts were then extracted and demultiplexed using the *Velocyto* pipeline [47]. We then removed cells with less than 30% of unspliced counts, more than 70% of unspliced counts, fewer than 1,500, and more than 50,000 UMIs per cell. 38,913 cells passed the quality check (QC) cut-offs and were kept for the training of the different models. We trained our model from scratch. For Seurat's discrete cell cycle phase assignment, we used its original implementation in R, and the list of genes from [44], as advised in their tutorial [48]. We inferred cell cycle phases using Tricycle [49], using its original implementation in R, analyzing all cell types individually, using the original author's annotation. For Cyclum [13], we let the model find the best number of linear dimensions, which it found to be one, and used the hyperparameters that they used for mouse embryonic stem cells. For the PCA-derived phase inference, we first selected the genes in our cell cycle gene set and normalized the gene counts using Scannpy *normalize_total* function, with 10,000 as the target, and took the log of 1 plus the normalized counts. Then we mean-centered each gene's log-normalized counts, either estimating the mean using the full dataset or using the original authors' cell types annotation. We then selected the first two PCs to infer the cell cycle phase. Using the original authors' S-phase annotation, we

also selected the first and third PCs, which appeared to contain more variance associated with the cell cycle. We tested scPrisma [50], but not all cells could fit in the GPU used and using only 20% of the dataset led to inconclusive results. Furthermore, scPrisma fits can be improved using prior knowledge, but Seurat's estimation of the phase does not constitute a good prior in this case. We compared all these methods with a random phase assignment, for which the cell phases are drawn from a uniform distribution.

3.9.4 Breast cancer atlas from Pal *et al.*

We analyzed this breast cancer atlas [51], using the original mapping and cell type annotation. We dropped the sample ER_0001 because the cell barcode from the count matrix do not match the one containing the cell type annotation. We merged the data from the three tumor subtypes and selected the tumoral epithelial cells based on the original annotation. We used the cells that passed the original QC, leading to 85,542 cells. We trained the model starting with the Fourier coefficients obtained from the dataset of RPE1 cells from Lederer *et al.*, frozen for the first 20 epochs.

3.9.5 Dataset of pediatric AML cells from Lambo *et al.*

We analyzed this longitudinal dataset of pediatric AML [34], using the original mapping and cell type annotation. We merged all the patients and timepoints and selected the normal cells that had more than 2,000 UMIs and less than 6,000 UMIs and Malignant cells that had between 6,000 and 20,000 UMIs. To reduce training time, we kept only 50% of these cells, leading to 90,891 cells and a training time of less than 15 minutes. Similarly as before, we started the training of the model with the Fourier coefficients obtained from the dataset of RPE1 cells from Lederer *et al.*, frozen for the first 20 epochs.

3.9.6 Dataset collected around the clock from Auerbach *et al.*

We analyzed the dataset collected around the clock from the mouse aorta [20], because its high quality (median UMI around 10,000 counts) allows circadian phase inference at the single cell level. The cell types annotation and counts were downloaded directly from the original work. Based on the histogram of the number of counts per cell, we added a second filtering on their data by removing the cells with fewer than 5,000 UMIs. For comparison with their obtained median absolute deviation (MAD), we used both the subfiltered dataset and the original one and reported the results on both (Table 1). We used the same clock-related gene sets as the original authors, which can be found in Supplementary Table 1.

3.9.7 Dataset from the human endometrium from Mareckova *et al.*

We analyzed the dataset of single-nucleus RNA-seq from human endometrium by [36]. The authors sequenced cells from snap-frozen samples of superficial endometrial biopsies from 63 donors, with and without endometriosis. We used the mapping and counts provided by the authors available at ArrayExpress under the accession number E-MTAB-14039, and selected the stromal cells for the analysis of the menstrual cycle, and randomly selected one quarter of the dataset, leading to 30,536 cells. The cycling gene list was obtained as explained above (3.8). For the cell cycle, we randomly sampled one-third of the dataset for the cell cycle analysis, leading to 93,674 cells.

3.10 Evaluation metrics

3.10.1 Median absolute deviation (MAD)

To evaluate model performance on the mouse aorta dataset collected around the clock, we followed the original study and quantified accuracy using the median absolute deviation (MAD) between the inferred phase and the external sampling time. Specifically, MAD is defined as

$$\mathrm{MAD} = \mathrm{median}(|d(\varphi_i', t_i)|)\,, \tag{12}$$

where $d(\varphi_i', t_i)$ denotes the circular distance between the inferred phase φ_i' and the external time t_i of cell i.

Because both the origin and direction of the inferred phase are arbitrary, we computed MAD after optimally aligning the inferred phases to external time by selecting the shift and orientation that minimized the deviation, independently for each cell type.

As noted previously, a low MAD is not always expected, as it would imply perfect synchrony of circadian phases across individual cells. Nevertheless, achieving a lower MAD without incorporating external time into model selection or training provides evidence that the inferred phases capture biologically meaningful circadian structure.

3.10.2 Jensen–Shannon divergence

To assess agreement between inferred phases and the binary S-phase annotation provided by the original authors [15], we computed the Jensen–Shannon (JS) divergence between the phase distributions of S-phase and non–S-phase cells. The JS divergence is a symmetric and smoothed variant of the Kullback–Leibler divergence, defined as

$$\mathrm{JSD}(P \,\|\, Q) = \frac{1}{2} D_{\mathrm{KL}}(P \,\|\, M) + \frac{1}{2} D_{\mathrm{KL}}(Q \,\|\, M), \tag{13}$$

where $M = \frac{1}{2}(P + Q)$ is the mixture distribution.

To obtain a proper distance metric, we report the square root of the JS divergence, which satisfies the triangle inequality [52]. Using a base-2 logarithm, the resulting Jensen–Shannon distance is bounded between 0 and 1, with 0 indicating identical distributions and 1 indicating completely non-overlapping distributions.

4 Discussion

Biological cycles such as the cell cycle, circadian rhythms, or hormonal cycles are intrinsic, periodic processes that unfold within heterogeneous cellular environments. In vivo, these cycles are not necessarily synchronized across cells and are frequently modulated by cell identity, differentiation state, or disease context. Consequently, cyclic gene expression programs cannot be represented as fixed trajectories but need to be modeled as periodic functions that can be smoothly deformed by the biological context, a construction that we termed context-dependent periodic manifolds.

Current approaches to cycle inference from scRNA-seq typically assume homogeneous dynamics across cells, which both confound the inference of latent phases and limit comparisons in heterogeneous samples. This motivated the development of CoPhaser, a model that preserves a shared and interpretable phase coordinate while allowing to unravel biologically relevant modulations of cyclic

gene expression in several case studies. CoPhaser decomposes gene expression into two components: one derived from a one-dimensional cyclic latent coordinate, and one from a non-cyclic context capturing variability associated with other cellular properties. A mutual information network controls independence between both latent spaces. Compared with existing methods to perform cyclic inference in single-cell data (see Introduction), CoPhaser stands out as with its explicit parameterization of gene expression variability, which is biologically rooted and facilitates interpretation. For instance, we used integrated gradients to reveal gene programs that modulate circadian gene expression programs in different cellular subtypes in the aorta.

A notable benefit of CoPhaser is its applicability across biological cycles of different molecular nature, with widely differing timescales. The same framework applies to fast processes such as the cell cycle and circadian rhythms, as well as slower hormonal cycles, without requiring cycle-specific assumptions beyond periodicity. A key benefit is that, aside from a set of seed genes associated with the cycle period—which can be partial and refined iteratively, the model does not require prior knowledge of cycle orbits or latent cell phases. In practice, CoPhaser converges to biologically meaningful solutions when initialized fully randomly.

The biological examples presented underscore the potential of CoPhaser to generate interpretable biological insights when applied to complex in vivo datasets. In the mouse embryo, CoPhaser uncovered cell-type–specific differences in cell-cycle dynamics, which were particularly pronounced for genes peaking in the G1 and G1/S phases. In cancer datasets, CoPhaser clarified how subtype-specific transcriptional programs reshape cell-cycle dynamics rather than uniformly promoting proliferation. Notably, it enabled the distinction between constitutively overexpressed genes and those whose overexpression is coupled to specific cell-cycle phases, a distinction that is relevant for identifying therapeutic targets in lowly proliferating, persistent cancer cell populations. In acute myeloid leukemia, CoPhaser further enabled the separation of proliferative activity from disease-associated transcriptional states, revealing that lowly proliferating hematopoietic stem cells (HSCs) can increase their relative abundance within the tumoral compartment upon drug treatment while maintaining low proliferation rates.

In circadian aorta, the model revealed cell-type–specific differences in rhythmic amplitude and baseline expression that are obscured by context-agnostic circadian phase inference methods. This enabled precise circadian phase inference across cell types and the identification of cell-type–specific circadian genes. Moreover, the high-resolution context space allowed the discovery of circadian gene expression differences between closely related cell subtypes.

Finally, in the menstrual cycle, CoPhaser captured gradual and continuous transcriptional remodeling across phases, supporting a view of hormonal cycling as a combination of smooth transitions and key turning points that induce strong transcriptional changes. In this context, CoPhaser uncovered distinct gene dynamics in stromal cells between women with and without endometriosis, and allowed the discovery of a subtype of cycling cells.

Limitations. CoPhaser relies on an initial set of rhythmic genes, although iterative refinement reduces sensitivity to this choice. Detecting weak oscillations remains dependent on sequencing depth, particularly for low-amplitude circadian signals. From a computational perspective, training costs exceed those of much simpler phase-calling methods such as PCA, but remain tractable for typically sized scRNA-seq datasets. Furthermore, the model is sensitive to its hyperparameters, which can be tuned using biological prior knowledge, and can be mitigated using transfer learning.

More broadly, this work shows the advantage of encoding biological topology directly into latent representations. Structured latent spaces that reflect known biological principles provide a path toward more interpretable and robust models. CoPhaser contributes to this emerging class of biologically informed single-cell methods by demonstrating how explicit periodic structure and context disentanglement can be combined within a flexible learnable model. Finally, the open and modular implementation enables reuse and transfer learning across datasets, facilitating application to a broad range of biological systems.

5 Supplementary Material

Acknowledgments. We thank members of the Naef lab for insightful discussions. This work was supported by the Swiss National Science Foundation (individual grant 310030B_201267 to F.N.) and by the Chan Zuckerberg Initiative (CZI) Single-Cell Biology Data Insights program (DI3) grant "Space-time Regulation of Biological Cycles in Cancer" awarded to Felix Naef and Nacho Molina.

References

[1] Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., Luo, Y.: Single-cell RNA sequencing technologies and applications: A brief overview. Clinical and Translational Medicine **12**(3), 694 (2022) https://doi.org/10.1002/ctm2.694 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ctm2.694. Accessed 2025-12-24

[2] Wang, L., Zhang, Q., Qin, Q., Trasanidis, N., Vinyard, M., Chen, H., Pinello, L.: Current progress and potential opportunities to infer single-cell developmental trajectory and cell fate. Current opinion in systems biology **26**, 1–11 (2021) https://doi.org/10.1016/j.coisb.2021.03.006 . Accessed 2025-12-24

[3] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature Methods **15**(12), 1053–1058 (2018) https://doi.org/10.1038/s41592-018-0229-2 . Publisher: Nature Publishing Group. Accessed 2025-12-11

[4] Lotfollahi, M., Wolf, F.A., Theis, F.J.: scGen predicts single-cell perturbation responses. Nature Methods **16**(8), 715–721 (2019) https://doi.org/10.1038/s41592-019-0494-8 . Publisher: Nature Publishing Group. Accessed 2025-12-24

[5] Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I.L., Srivatsan, S.R., Naghipourfar, M., Daza, R.M., Martin, B., Shendure, J., McFaline-Figueroa, J.L., Boyeau, P., Wolf, F.A., Yakubova, N., Günnemann, S., Trapnell, C., Lopez-Paz, D., Theis, F.J.: Predicting cellular responses to complex perturbations in high-throughput screens. Molecular Systems Biology **19**(6), 11517 (2023) https://doi.org/10.15252/msb.202211517

[6] Rand, D.A., Raju, A., Sáez, M., Corson, F., Siggia, E.D.: Geometry of gene regulatory dynamics. Proceedings of the National Academy of Sciences **118**(38), 2109729118 (2021) https://doi.org/10.1073/pnas.2109729118 . ZSCC: NoCitationData[s0]. Accessed 2021-11-04

[7] Piran, Z., Cohen, N., Hoshen, Y., Nitzan, M.: Disentanglement of single-cell data with biolord. Nature Biotechnology **42**(11), 1678–1683 (2024) https://doi.org/10.1038/s41587-023-02079-x

. Accessed 2025-12-19

[8] Battich, N., Beumer, J., Barbanson, B., Krenning, L., Baron, C.S., Tanenbaum, M.E., Clevers, H., Oudenaarden, A.: Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. Science **367**(6482), 1151–1156 (2020) https://doi.org/10.1126/science.aax3072 . Publisher: American Association for the Advancement of Science. Accessed 2024-12-27

[9] Wang, J., Symul, L., Yeung, J., Gobet, C., Sobel, J., Lück, S., Westermark, P.O., Molina, N., Naef, F.: Circadian clock-dependent and -independent posttranscriptional regulation underlies temporal mRNA accumulation in mouse liver. Proceedings of the National Academy of Sciences **115**(8), 1916–1925 (2018) https://doi.org/10.1073/pnas.1715225115 . ZSCC: NoCitationData[s0]. Accessed 2021-08-02

[10] Guo, X., Chen, L.: From G1 to M: a comparative study of methods for identifying cell cycle phases. Briefings in Bioinformatics **25**(2), 517 (2024) https://doi.org/10.1093/bib/bbad517 . Accessed 2025-12-19

[11] Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O.: 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) https://doi.org/10.1038/nbt.3102 . Accessed 2025-12-24

[12] Nariya, M.K., Santiago-Algarra, D., Tassy, O., Cerciat, M., Ye, T., Riba, A., Molina, N.: Single-cell multiomics reveals the oscillatory dynamics of mRNA metabolism and chromatin accessibility during the cell cycle. Cell Reports **44**(8), 116089 (2025) https://doi.org/10.1016/j.celrep.2025.116089 . Accessed 2025-12-19

[13] Liang, S., Wang, F., Han, J., Chen, K.: Latent periodic process inference from single-cell RNA-seq data. Nature Communications **11**(1), 1441 (2020) https://doi.org/10.1038/s41467-020-15295-9 . Publisher: Nature Publishing Group. Accessed 2025-12-19

[14] Riba, A., Oravecz, A., Durik, M., Jiménez, S., Alunni, V., Cerciat, M., Jung, M., Keime, C., Keyes, W.M., Molina, N.: Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. Nature Communications **13**(1), 2865 (2022) https://doi.org/10.1038/s41467-022-30545-8 . Number: 1 Publisher: Nature Publishing Group. Accessed 2023-12-04

[15] Salmen, F., De Jonghe, J., Kaminski, T.S., Alemany, A., Parada, G.E., Verity-Legg, J., Yanagida, A., Kohler, T.N., Battich, N., Brekel, F., Ellermann, A.L., Arias, A.M., Nichols, J., Hemberg, M., Hollfelder, F., Oudenaarden, A.: High-throughput total RNA sequencing in single cells using VASA-seq. Nature Biotechnology **40**(12), 1780–1793 (2022) https://doi.org/10.1038/s41587-022-01361-8 . Publisher: Nature Publishing Group. Accessed 2025-12-15

[16] Shen, S., Vagner, S., Robert, C.: Persistent Cancer Cells: The Deadly Survivors. Cell **183**(4), 860–874 (2020) https://doi.org/10.1016/j.cell.2020.10.027 . Accessed 2025-12-24

[17] Gaglia, G., Kabraji, S., Rammos, D., Dai, Y., Verma, A., Wang, S., Mills, C.E., Chung, M., Bergholz, J.S., Coy, S., Lin, J.-R., Jeselsohn, R., Metzger, O., Winer, E.P., Dillon, D.A., Zhao,

J.J., Sorger, P.K., Santagata, S.: Temporal and spatial topography of cell proliferation in cancer. Nature Cell Biology **24**(3), 316–326 (2022) https://doi.org/10.1038/s41556-022-00860-9 . Publisher: Nature Publishing Group. Accessed 2025-12-24

[18] Anafi, R.C., Francey, L.J., Hogenesch, J.B., Kim, J.: CYCLOPS reveals human transcriptional rhythms in health and disease. Proceedings of the National Academy of Sciences of the United States of America **114**(20), 5312–5317 (2017) https://doi.org/10.1073/pnas.1619320114 . Accessed 2025-12-19

[19] Talamanca, L., Gobet, C., Naef, F.: Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans. Science **379**(6631), 478–483 (2023) https://doi.org/10.1126/science.add0846 . Publisher: American Association for the Advancement of Science. Accessed 2023-12-04

[20] Auerbach, B.J., FitzGerald, G.A., Li, M.: Tempo: an unsupervised Bayesian algorithm for circadian phase inference in single-cell transcriptomics. Nature Communications **13**(1), 6580 (2022) https://doi.org/10.1038/s41467-022-34185-w . Publisher: Nature Publishing Group. Accessed 2025-12-17

[21] Duan, J., Ngo, M.N., Karri, S.S., Tsoi, L.C., Gudjonsson, J.E., Shahbaba, B., Lowengrub, J., Andersen, B.: tauFisher predicts circadian time from a single sample of bulk and single-cell pseudobulk transcriptomic data. Nature Communications **15**(1), 3840 (2024) https://doi.org/10.1038/s41467-024-48041-6 . Publisher: Nature Publishing Group. Accessed 2025-12-19

[22] Xu, B., Braun, R.: Variational inference of single cell time series. Bioinformatics (2024). https://doi.org/10.1101/2024.08.29.610389 . http://biorxiv.org/lookup/doi/10.1101/2024.08.29.610389 Accessed 2025-12-22

[23] Chen, S., Regev, A., Condon, A., Ding, J.: CellUntangler: Separating distinct biological signals in single-cell data with deep generative models. Cell Genomics **0**(0) (2025) https://doi.org/10.1016/j.xgen.2025.101073 . Publisher: Elsevier. Accessed 2025-12-24

[24] Cao, N.D., Aziz, W.: The Power Spherical distribution. arXiv. arXiv:2006.04437 [stat] (2020). https://doi.org/10.48550/arXiv.2006.04437 . http://arxiv.org/abs/2006.04437 Accessed 2025-05-05

[25] Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual Information Neural Estimation. In: Proceedings of the 35th International Conference on Machine Learning, pp. 531–540. PMLR, ??? (2018). ISSN: 2640-3498. https://proceedings.mlr.press/v80/belghazi18a.html Accessed 2024-12-18

[26] Schwabe, D., Formichetti, S., Junker, J.P., Falcke, M., Rajewsky, N.: The transcriptome dynamics of single cells during the cell cycle. Molecular Systems Biology **16**(11), 20209946 (2020) https://doi.org/10.15252/msb.20209946 . Accessed 2025-12-24

[27] Santos, A., Wernersson, R., Jensen, L.J.: Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Research **43**(Database issue), 1140–1144 (2015) https://doi.org/10.1093/nar/gku1092 . Accessed 2025-12-15

[28] Lederer, A.R., Leonardi, M., Talamanca, L., Bobrovskiy, D.M., Herrera, A., Droin, C., Khven, I., Carvalho, H.J.F., Valente, A., Dominguez Mantes, A., Mulet Arabí, P., Pinello, L., Naef, F., La Manno, G.: Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations. Nature Methods **21**(12), 2271–2286 (2024) https://doi.org/10.1038/s41592-024-02471-8 . Publisher: Nature Publishing Group. Accessed 2025-12-15

[29] Loaiza-Ganem, G., Cunningham, J.P.: The continuous Bernoulli: fixing a pervasive error in variational autoencoders

[30] Pal, B., Chen, Y., Vaillant, F., Capaldo, B.D., Joyce, R., Song, X., Bryant, V.L., Penington, J.S., Di Stefano, L., Tubau Ribera, N., Wilcox, S., Mann, G.B., Papenfuss, A.T., Lindeman, G.J., Smyth, G.K., Visvader, J.E.: A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO Journal **40**(11), 107333 (2021) https://doi.org/10.15252/embj.2020107333 . Accessed 2025-12-17

[31] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. arXiv. arXiv:1703.01365 [cs] (2017). https://doi.org/10.48550/arXiv.1703.01365 . http://arxiv.org/abs/1703.01365 Accessed 2025-12-24

[32] Le, X., Chen, Q., Wen, Q., Cao, S., Zhang, L., Hu, L., Hu, G., Li, Q., Chen, Z.: Design, synthesis and optimization of Apcin analogues as Cdc20 inhibitors for triple-negative breast cancer therapy. European Journal of Medicinal Chemistry **289**, 117434 (2025) https://doi.org/10.1016/j.ejmech.2025.117434 . Accessed 2025-12-17

[33] Verma, A., Singh, A., Singh, M.P., Nengroo, M.A., Saini, K.K., Satrusal, S.R., Khan, M.A., Chaturvedi, P., Sinha, A., Meena, S., Singh, A.K., Datta, D.: EZH2-H3K27me3 mediated KRT14 upregulation promotes TNBC peritoneal metastasis. Nature Communications **13**, 7344 (2022) https://doi.org/10.1038/s41467-022-35059-x . Accessed 2025-12-17

[34] Lambo, S., Trinh, D.L., Ries, R.E., Jin, D., Setiadi, A., Ng, M., Leblanc, V.G., Loken, M.R., Brodersen, L.E., Dai, F., Pardo, L.M., Ma, X., Vercauteren, S.M., Meshinchi, S., Marra, M.A.: A longitudinal single-cell atlas of treatment response in pediatric AML. Cancer Cell **41**(12), 2117–213512 (2023) https://doi.org/10.1016/j.ccell.2023.10.008 . Accessed 2025-12-15

[35] Dobnikar, L., Taylor, A.L., Chappell, J., Oldach, P., Harman, J.L., Oerton, E., Dzierzak, E., Bennett, M.R., Spivakov, M., Jørgensen, H.F.: Disease-relevant transcriptional signatures identified in individual smooth muscle cells from healthy mouse vessels. Nature Communications **9**(1), 4567 (2018) https://doi.org/10.1038/s41467-018-06891-x . Publisher: Nature Publishing Group. Accessed 2025-12-18

[36] Marečková, M., Garcia-Alonso, L., Moullet, M., Lorenzi, V., Petryszak, R., Sancho-Serra, C., Oszlanczi, A., Icoresi Mazzeo, C., Wong, F.C.K., Kelava, I., Hoffman, S., Krassowski, M., Garbutt, K., Gaitskell, K., Yancheva, S., Woon, E.V., Male, V., Granne, I., Hellner, K., Mahbubani, K.T., Saeb-Parsy, K., Lotfollahi, M., Prigmore, E., Southcombe, J., Dragovic, R.A., Becker, C.M., Zondervan, K.T., Vento-Tormo, R.: An integrated single-cell reference atlas of the human endometrium. Nature Genetics **56**(9), 1925–1937 (2024) https://doi.org/10.1038/s41588-024-01873-w . Publisher: Nature Publishing Group. Accessed 2025-12-18

[37] Wang, W., Vilella, F., Alama, P., Moreno, I., Mignardi, M., Isakova, A., Pan, W., Simon, C., Quake, S.R.: Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. Nature Medicine **26**(10), 1644–1653 (2020) https://doi.org/10.1038/s41591-020-1040-z . Publisher: Nature Publishing Group. Accessed 2025-12-18

[38] Arora, S., Cohen, N., Hu, W., Luo, Y.: Implicit regularization in deep matrix factorization. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 7413–7424. Curran Associates Inc., Red Hook, NY, USA (2019)

[39] Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (GELUs). arXiv. arXiv:1606.08415 [cs] (2023). https://doi.org/10.48550/arXiv.1606.08415 . http://arxiv.org/abs/1606.08415 Accessed 2025-12-24

[40] Jang, E., Gu, S., Poole, B.: Categorical Reparameterization with Gumbel-Softmax. arXiv. arXiv:1611.01144 [stat] (2017). https://doi.org/10.48550/arXiv.1611.01144 . http://arxiv.org/abs/1611.01144 Accessed 2025-12-24

[41] Maddison, C.J., Mnih, A., Teh, Y.W.: The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. arXiv. arXiv:1611.00712 [cs] (2017). https://doi.org/10.48550/arXiv.1611.00712 . http://arxiv.org/abs/1611.00712 Accessed 2025-12-24

[42] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: -VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK (2017)

[43] scanpy.pp.highly_variable_genes. https://scanpy.readthedocs.io/en/stable/generated/scanpy.pp.highly_variable_genes.html Accessed 2025-12-19

[44] Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Gaillard, A., Kolb, K.E., Villani, A.-C., Johannessen, C.M., Andreev, A.Y., Van Allen, E.M., Bertagnolli, M., Sorger, P.K., Sullivan, R.J., Flaherty, K.T., Frederick, D.T., Jané-Valbuena, J., Yoon, C.H., Rozenblatt-Rosen, O., Shalek, A.K., Regev, A., Garraway, L.A.: Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science **352**(6282), 189–196 (2016) https://doi.org/10.1126/science.aad0501 . Publisher: American Association for the Advancement of Science. Accessed 2025-12-19

[45] mutual_info_classif. https://scikit-learn/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html Accessed 2025-12-19

[46] Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J.,

Bielas, J.H.: Massively parallel digital transcriptional profiling of single cells. Nature Communications **8**(1), 14049 (2017) https://doi.org/10.1038/ncomms14049 . Publisher: Nature Publishing Group. Accessed 2025-12-24

[47] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L.E., Liu, Z., Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., Kharchenko, P.V.: RNA velocity of single cells. Nature **560**(7719), 494–498 (2018) https://doi.org/10.1038/s41586-018-0414-6 . Publisher: Nature Publishing Group. Accessed 2025-12-24

[48] Analysis, visualization, and integration of Visium HD spatial datasets with Seurat. https://satijalab.org/seurat/articles/cell_cycle_vignette.html Accessed 2025-12-19

[49] Zheng, S.C., Stein-O'Brien, G., Augustin, J.J., Slosberg, J., Carosso, G.A., Winer, B., Shin, G., Bjornsson, H.T., Goff, L.A., Hansen, K.D.: Universal prediction of cell-cycle position using transfer learning. Genome Biology **23**(1), 41 (2022) https://doi.org/10.1186/s13059-021-02581-y . Accessed 2025-12-15

[50] Karin, J., Bornfeld, Y., Nitzan, M.: scPrisma infers, filters and enhances topological signals in single-cell data using spectral template matching. Nature Biotechnology **41**(11), 1645–1654 (2023) https://doi.org/10.1038/s41587-023-01663-5 . Publisher: Nature Publishing Group. Accessed 2025-12-19

[51] Pal, B., Chen, Y., Vaillant, F., Capaldo, B.D., Joyce, R., Song, X., Bryant, V.L., Penington, J.S., Di Stefano, L., Tubau Ribera, N., Wilcox, S., Mann, G.B., kConFab, Papenfuss, A.T., Lindeman, G.J., Smyth, G.K., Visvader, J.E.: A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO journal **40**(11), 107333 (2021) https://doi.org/10.15252/embj.2020107333

[52] Endres, D.M., Schindelin, J.E.: A new metric for probability distributions. IEEE Transactions on Information Theory **49**(7), 1858–1860 (2003) https://doi.org/10.1109/TIT.2003.813506 . Accessed 2025-12-24

6 Figures

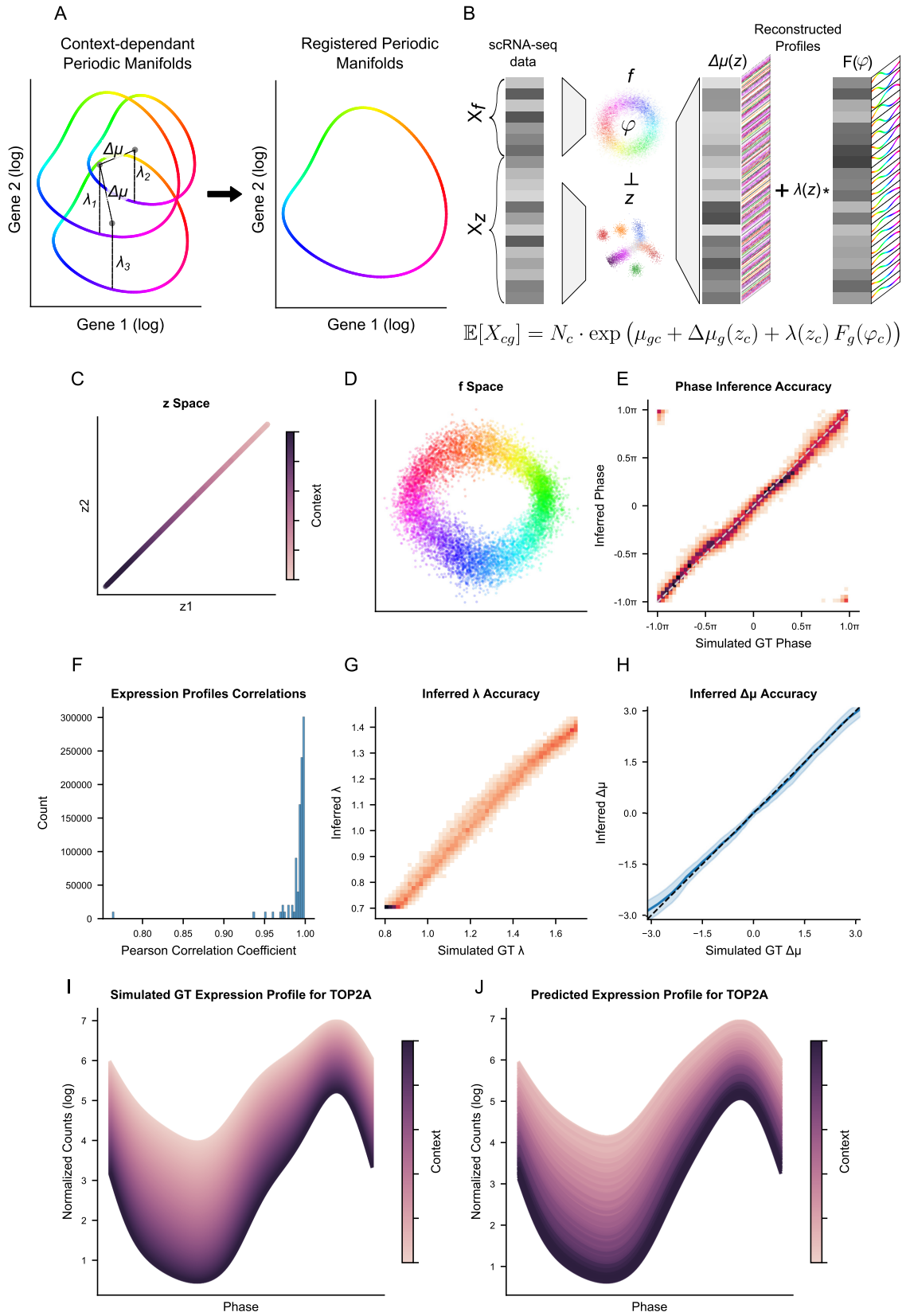
A
Context-dependant Periodic Manifolds
Registered Periodic Manifolds
Gene 2 (log)
Gene 1 (log)
Δμ
λ1
λ2
λ3
B
scRNA-seq data
Reconstructed Profiles
Δμ(z)
F(φ)
f
φ
z
+ λ(z)*
$\mathbb{E}[X_{cg}] = N_c \cdot \exp\left(\mu_{gc} + \Delta\mu_g(z_c) + \lambda(z_c)\, F_g(\varphi_c)\right)$
C
z Space
z1
z2
Context
D
f Space
E
Phase Inference Accuracy
Inferred Phase
Simulated GT Phase
F
Expression Profiles Correlations
Count
Pearson Correlation Coefficient
G
Inferred λ Accuracy
Inferred λ
Simulated GT λ
H
Inferred Δμ Accuracy
Inferred Δμ
Simulated GT Δμ
I
Simulated GT Expression Profile for TOP2A
Normalized Counts (log)
Phase
Context
J
Predicted Expression Profile for TOP2A
Normalized Counts (log)
Phase
Context

Fig. 1 CoPhaser enables context-dependent learning of periodic manifolds using a VAE. A. Schematic overview of the learning task. The model addresses scenarios in which context-dependent periodic manifolds (three are shown) are registered by shifts ($\Delta\mu$) and scaling (λ) (in log). **B.** Schematic of the model architecture. Expression of rhythmic genes (X_f) and context genes (X_z) are projected into a two-dimensional periodic f-space and a multi-dimensional contextual z-space, respectively. Independence between the two latent spaces is enforced by controlling their mutual information (I). The cyclic component of gene expression (F_g) is reconstructed from the cell-specific phase (φ) in f-space using Fourier series. Gene- and cell-specific mean shifts ($\Delta\mu_g(z_c)$) and cell-specific rhythmic amplitudes ($\lambda(z_c)$) are decoded from z-space. Expected gene counts ($\mathbb{E}[X_{cg}]$) are obtained by combining the library size (N_c) with the inferred mean expression fraction (μ_{cg}). **C.** z-space inferred from simulated data containing a one-dimensional contextual variable. **D.** Inferred f-space from simulated data, colored by the simulated ground-truth (GT) phase. **E.** Scatterplot comparing inferred and simulated GT phases, colored by local density (darker indicates high density). **F.** Pearson correlation coefficient between inferred and simulated GT gene expression profiles evaluated at 20 evenly spaced phases in $[-\pi, \pi)$ using inferred and simulated GT Fourier coefficients for a fixed context (Methods 3.7). **G.** Scatterplot comparing inferred and simulated GT rhythmic amplitudes ($\lambda(z_c)$), colored by the local density. **H.** 95 % confidence intervals of inferred mean shifts across the range of simulated GT values across all genes and contexts. **I.** Simulated expression profile of DNA topoisomerase II alpha (*TOP2A*) across the one-dimensional context. Realistic Fourier coefficients for the simulations were obtained from fitting RPE1 scRNA-seq data [28], and $\Delta\mu$ and λ were randomly generated as functions of the contextual variable. **J.** Reconstructed *TOP2A* expression profile using the inferred Fourier coefficients, $\Delta\mu$, and λ.

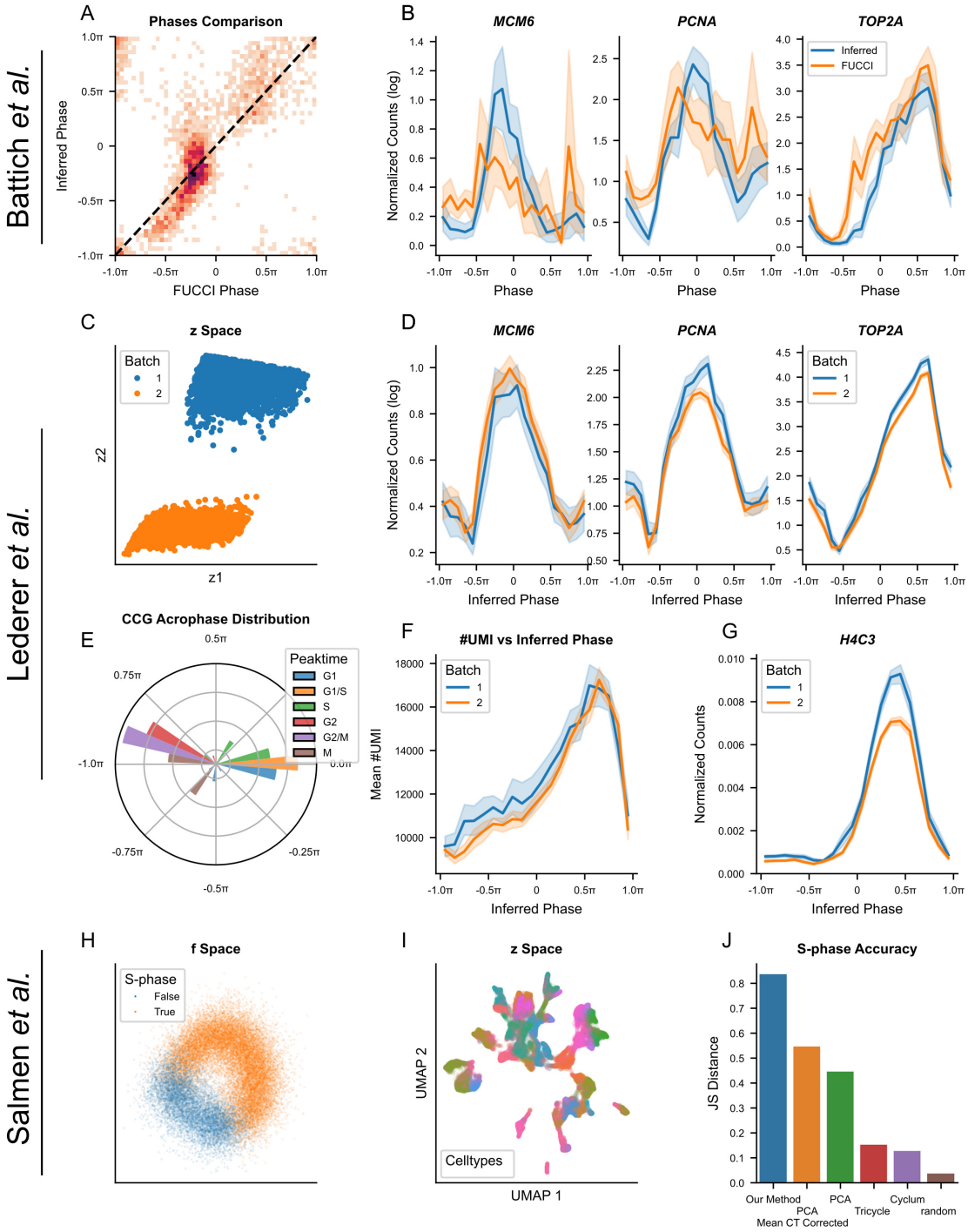

Fig. 2 Context-aware inference of continuous cell-cycle phase across datasets and technologies. A,B. Cell-cycle phase inference on RPE1-FUCCI scRNA-seq data from [8]. **A** Comparison between inferred phases and FUCCI-derived ground-truth phases, color coded by local density. **B** Reconstructed gene expression profiles using inferred phases (blue) and FUCCI-derived phases (orange). Profiles are obtained by binning cells by phase and averaging $\log(1 + \text{counts per } 10{,}000)$ for the indicated gene. **C–G.** Cell-cycle phase inference on two RPE1 scRNA-seq batches (10x Genomics) from [28]. **C,** Inferred z-space colored by batch identity. **D** Reconstructed gene expression profiles colored by batch, obtained by binning cells according to inferred phase and averaging $\log(1+\text{counts per } 10{,}000)$ for the indicated gene. **E.** Distribution of inferred peak phases for cell-cycle genes, colored by their annotated peak phase according to Cyclebase [27]. **F.** Mean number of unique molecular identifiers (UMIs) per cell as a function of inferred phase, colored by batch. **G.** Mean fraction of counts derived from histone genes as a function of inferred phase, colored by batch. **H–J.** Cell-cycle phase inference on mouse embryonic scRNA-seq data (VASA-seq) from [15]. **H.** Projected f-space colored by the original S-phase annotations [15]. **I.** Projected z-space using a Uniform Manifold Approximation and Projection (UMAP) colored by original cell-type annotations [15]. **J.** Jensen–Shannon divergence (base-2 logarithm) between the original S-phase annotation and binned inferred phases, computed for CoPhaser, principal component-based baselines (PC1 and PC3 yielded the best score; mean-centered either by cell type or across all cells), Tricycle [49], Cyclops [18], and a random phase assignment.

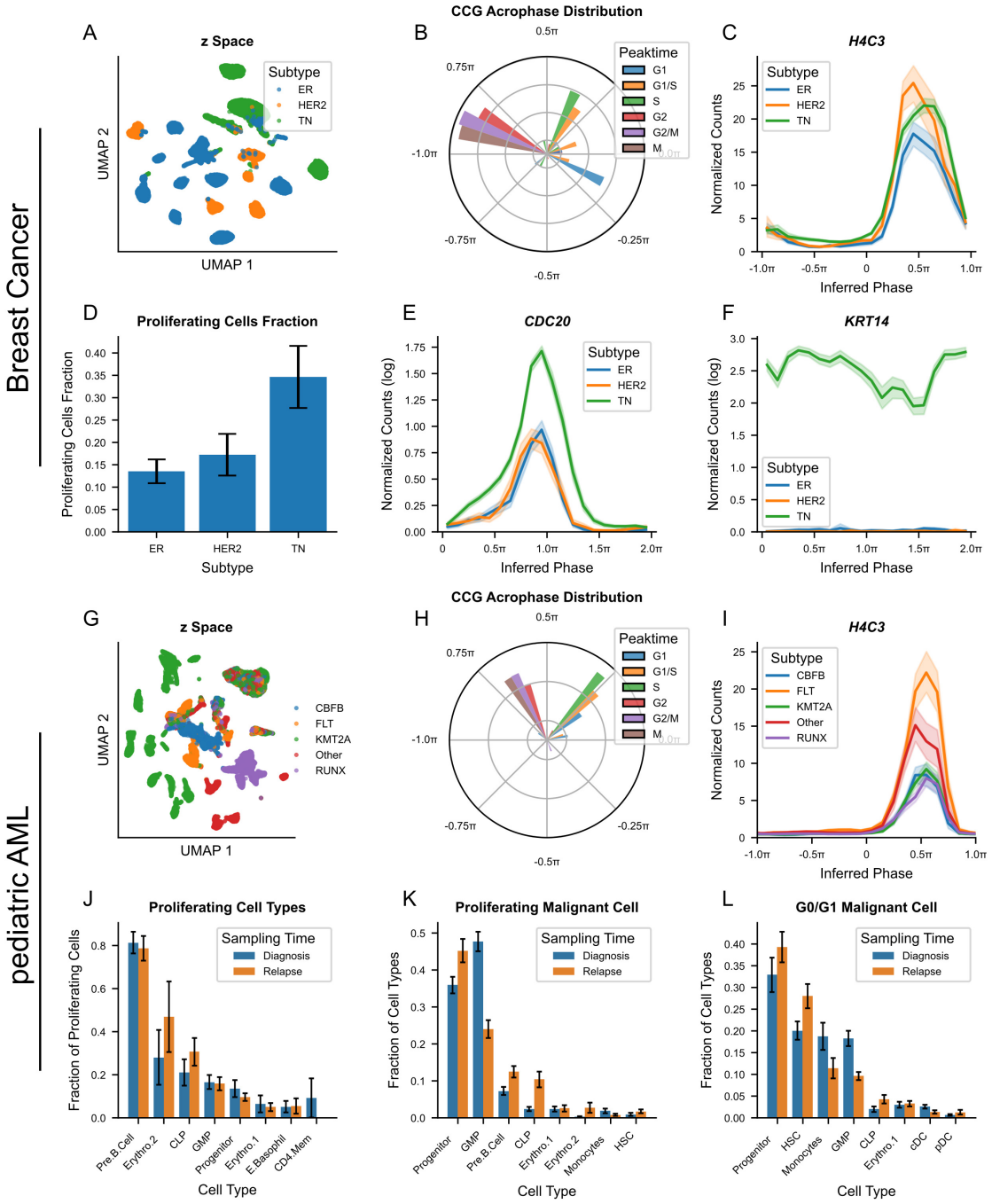

Fig. 3 CoPhaser reveals subtype-specific proliferation patterns in breast cancer and AML. A–F. Analysis of breast cancer scRNA-seq data from [30]. **A.** UMAP projection of the inferred contextual z-space colored by clinical subtype (ER: Estrogen receptor positive; HER2: human epidermal growth factor receptor 2 positive; and TN: triple-negative) **B.** Polar histogram of inferred peak phases for cell-cycle genes, colored by their annotated peak phase according to Cyclebase [27]. **C.** Reconstructed expression profile of *H4C3* as a function of inferred phase, colored by clinical subtype. **D.** Fraction of proliferating cells across tumor subtypes. **E,F.** Reconstructed expression profiles of *CDC20* (**E**) and *KRT14* (**F**) as a function of inferred phase, colored by clinical subtype. Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression ($\log(1 + \text{counts per } 10{,}000)$). **G–L.** Analysis of pediatric acute myeloid leukemia (AML) scRNA-seq data from [34] **G.** UMAP projection of the inferred contextual z-space, colored by tumor subtype. **H.** Polar histogram of inferred peak phases for cell-cycle genes, colored by their annotated peak phase according to Cyclebase [27]. **I.** Reconstructed expression profile of *H4C3* as a function of the inferred phase. Profiles were obtained by binning cells according to their inferred phase and averaging normalized expression (counts per 10,000) for each tumor subtype. **J.** Fraction of proliferating malignant cells for each cell type in diagnosis and relapse patients. Erythro.1: early erythrocytes; Erythro.2: late erythrocytes. Error bars represent the standard error of the mean (SEM) across patients. **K,L.** Cell-type composition of the proliferating malignant cells (**K**) and G0/G1 malignant cells (**L**) . Error bars represent the SEM across patients.

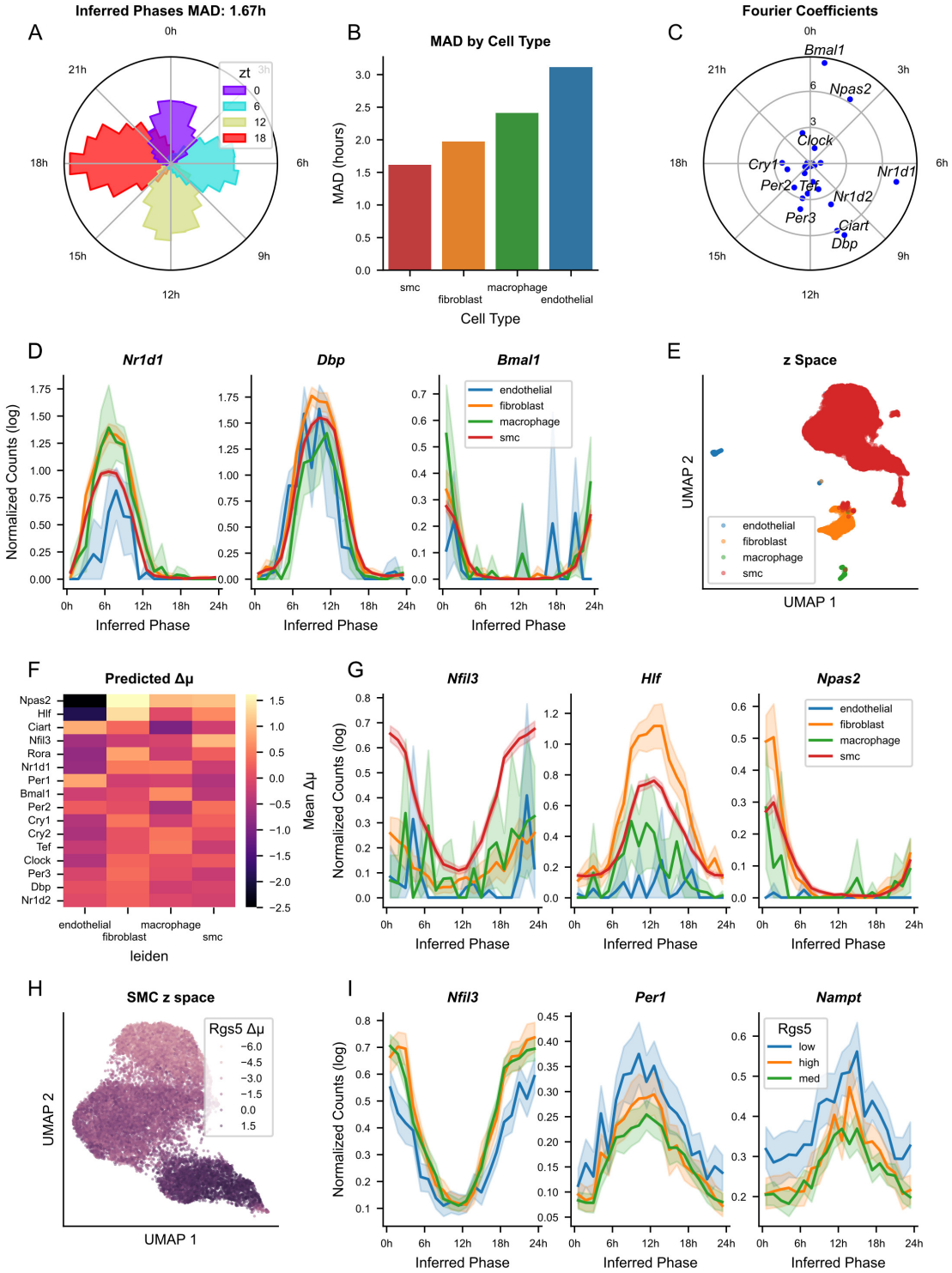

Fig. 4 CoPhaser accurately predicts circadian phases and context-modulated gene expression profiles in the mouse aorta. Analysis of mouse aorta scRNA-seq data from [20] collected every 4h over a 24h light-dark cycle. **A.** Polar histogram of inferred phases for all cells, colored by Zeitgeber Time (ZT). The Median Absolute Deviation (MAD) between the predicted phase and the ZT is 1.67h. **B.** MAD between the inferred phase and the ZT for each cell type. **C.** Polar scatter plot of inferred Fourier coefficients for rhythmic input genes suggests a *bona fide* circadian clock, showing peak-to-trough amplitude (log) as radius and peak phase as angle. **D.** Reconstructed expression profiles of core clock genes (*Nr1d1*, *Dbp*, *Bmal1*) as a function of inferred phase, colored by cell type. Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression (log(1+counts per 10,000)). **E.** UMAP projection of the inferred contextual z-space colored by cell type. **F.** Heatmap of inferred mean shifts ($\Delta\mu$) for rhythmic input genes across cell types. **G.** Reconstructed expression profiles of *Nfil3*, *Hlf*, and *Npas2* as a function of inferred phase, colored by cell type. Profiles computed as in **D**. **H.** UMAP projection of the inferred z-space colored by the inferred mean shift ($\Delta\mu$) of *Rgs5*, highlighting heterogeneity within smooth muscle cells (SMCs). **I.** Reconstructed expression profiles of *Nfil3*, *Per1*, and *Nampt* as a function of inferred phase, colored by SMC subtypes (Leiden clusters), corresponding to low, medium and high *Rgs5* expression levels. Profiles computed as in **D**.

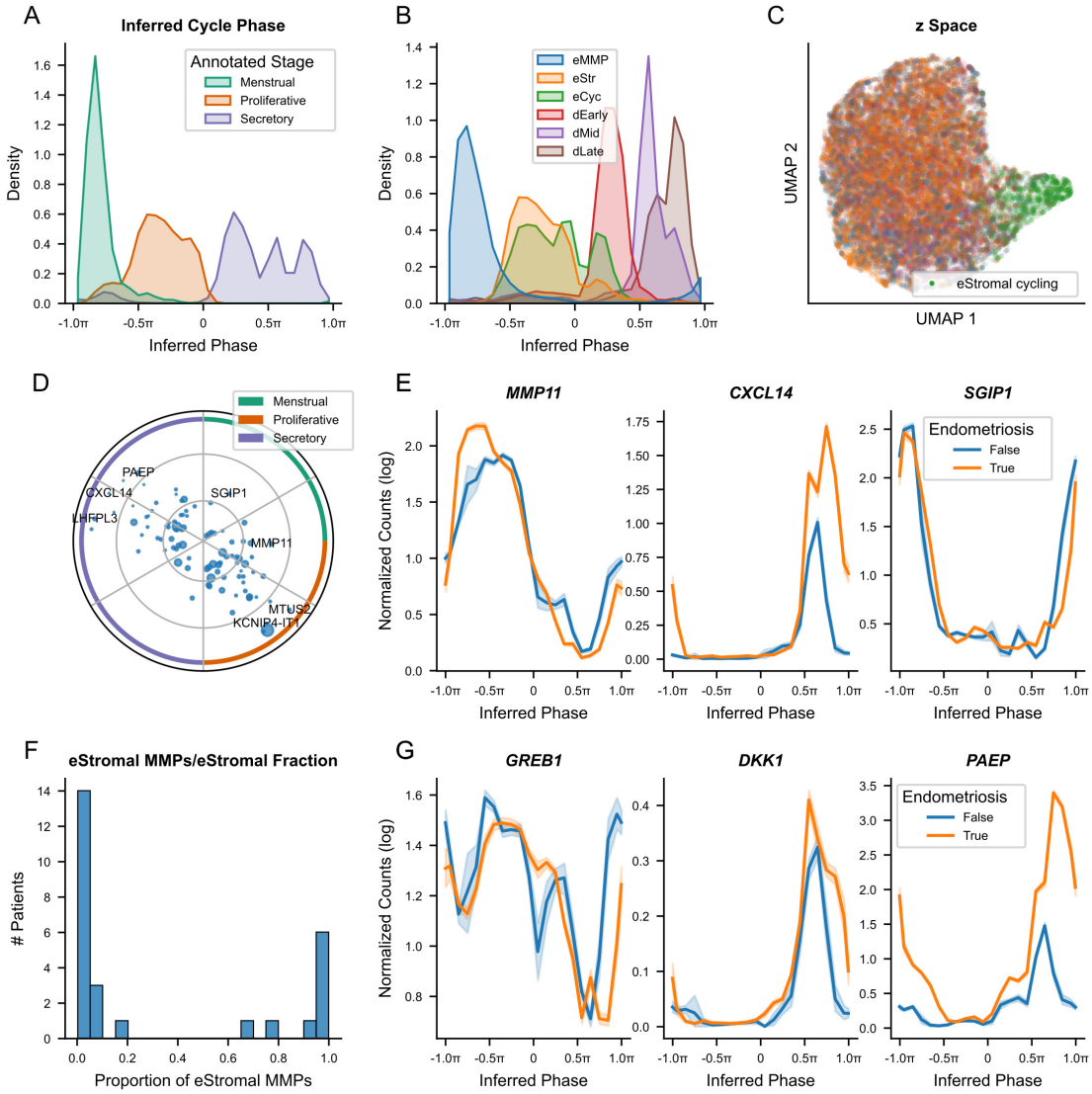

Fig. 5 CoPhaser reveals menstrual cycle gene expression dynamics and endometriosis-associated alterations. Analysis of endometrium single-nucleus RNA-seq (snRNA-seq) data [36]. **A.** Distribution of inferred phases for stromal cells from donors without hormonal treatment, colored by the annotated menstrual cycle stage. **B.** Distribution of inferred phases for all stromal cells, colored by cell state. **C.** UMAP projection of the inferred contextual z-space, colored by cell state. **D.** Polar scatter plot of inferred amplitudes and phases for the set of rhythmic input genes, showing peak-to-trough amplitude (log) as radius and peak phase as angle. Point size is proportional to mean expression. Colored rings indicate the approximate menstrual cycle stages, corresponding to the color scheme of **A**. **E.** Reconstructed expression profiles of *MMP11*, *CXCL14*, and *SGIP1* as a function of inferred phase, comparing cells from healthy donors and donors with endometriosis. Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression ($\log(1+\text{counts per } 10{,}000)$). **F.** Distribution of the proportion of eStromal MMPs cells (relative to the total eStromal population) across patients. **G.** Reconstructed expression profiles of *GREB1*, *DKK1*, and *PAEP* as a function of inferred phase, comparing cells from healthy donors and donors with endometriosis. Profiles computed as in **E**.

Supplementary Figures

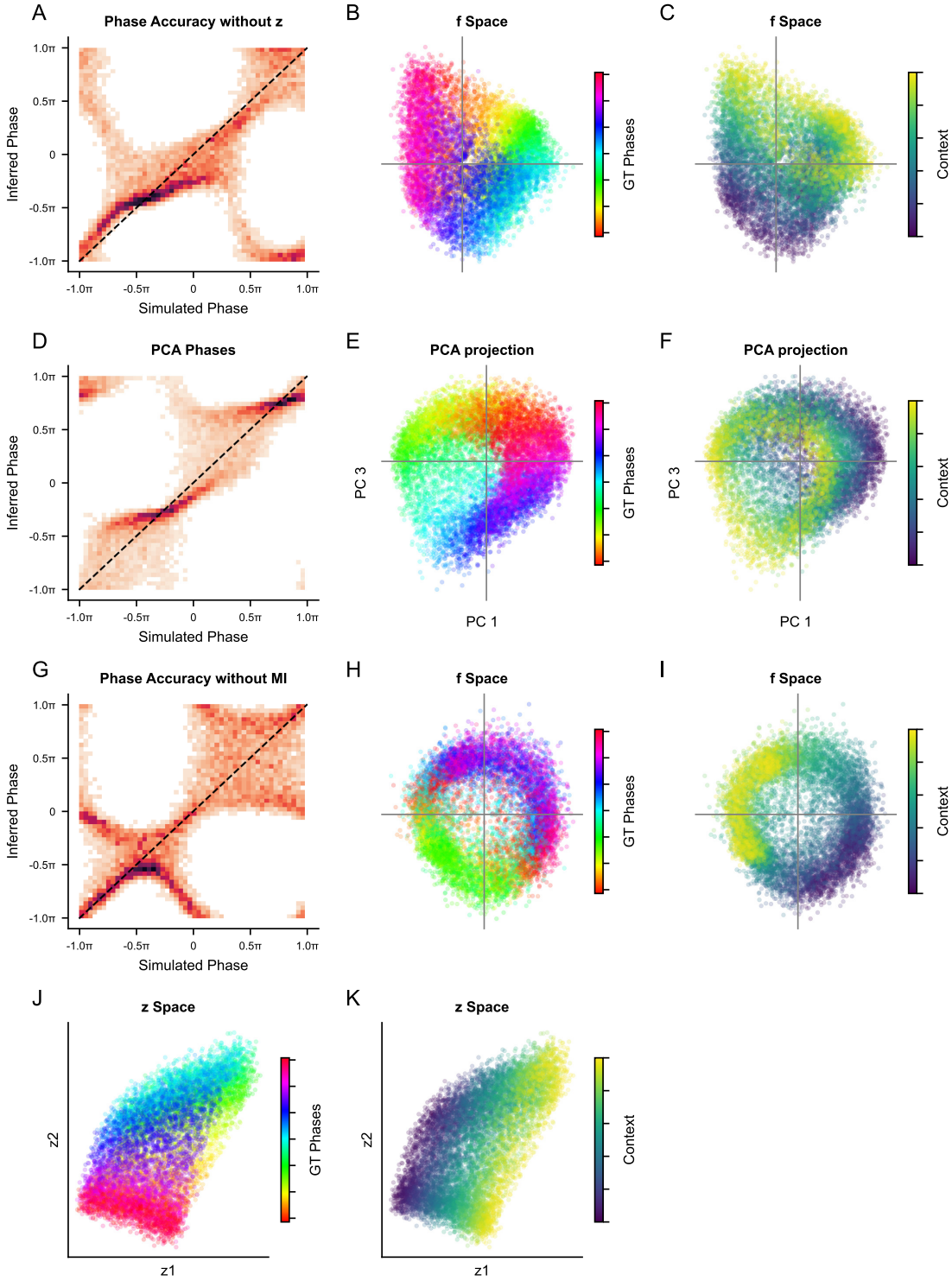

Fig. S1 Ablation study validates the dual latent space architecture and independence constraints. **A–C.** Results on simulated data when the contextual z-space is removed from the model. **A.** Comparison between inferred phases and the simulated ground truth (GT), colored by local density. **B, C.** Projected f-space colored by the GT phase (**B**) and by the one-dimensional contextual variable (**B**). The axis intersection indicates the (0,0) center. **D–F.** Principal component analysis (PCA) of the simulated data, projected onto the two components capturing most of the cyclic variance (PC1 and PC3). Panels are colored by local density (**D**), GT phases (**E**) and contextual variable (**F**). **G–K.** Results on simulated data without the Mutual Information Neural Estimation (MINE) to enforce independence between latent spaces. **G.** Comparison between inferred phases and GT phases, colored by local density. **H, I.** Projected f-space colored by GT phase (**H**) and contextual variable (**I**). **J, K.** z-space colored by GT phase (**J**) and contextual variable (**K**).

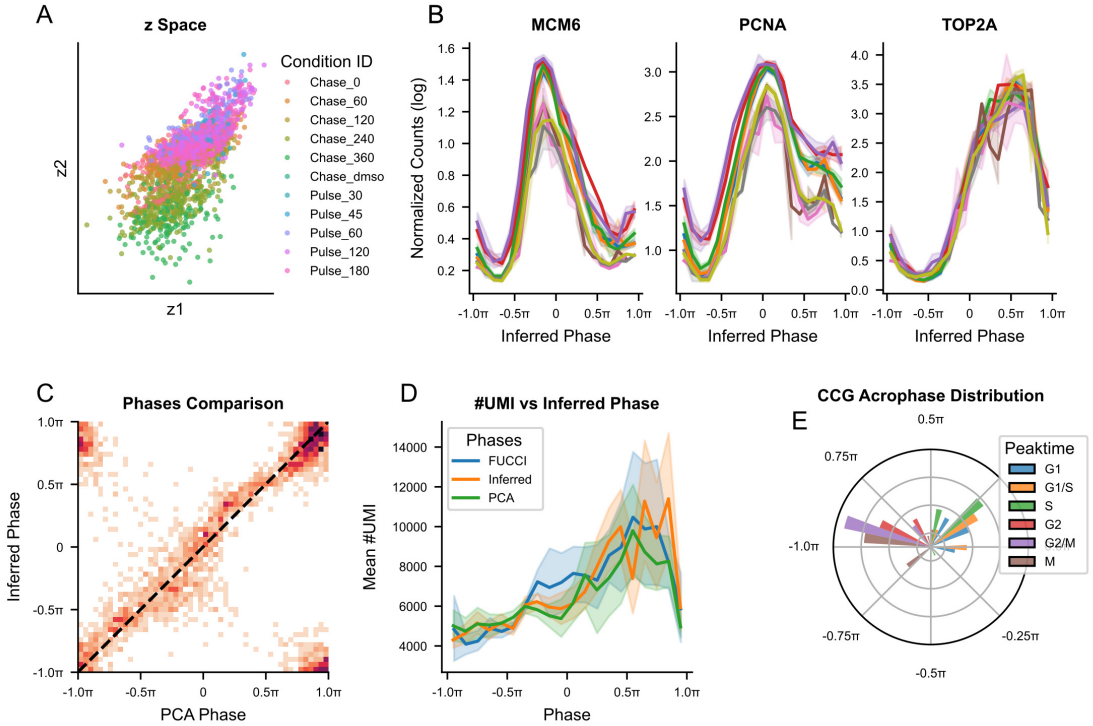

Fig. S2 Validation of context-corrected inferred cell-cycle phases. Extended analysis of the RPE1 scEU-seq dataset [8]. **A.** Inferred contextual z-space, colored by experimental condition (pulse/chase durations). **B** Reconstructed gene expression profiles for selected cell-cycle genes (*MCM6*, *PCNA*, *TOP2A*) as a function of inferred phase, colored by condition. Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression ($\log(1 + \text{counts per } 10{,}000)$). **C** Comparison between phases inferred by CoPhaser and phases obtained from principal component analysis (PCA), colored by local density. **D** Mean number of unique molecular identifiers (UMIs) per cell as a function of phase for FUCCI-derived, model-inferred, and PCA-derived phases in the chase condition. **E** Polar histogram of inferred peak phases for cell-cycle genes, colored by their annotated peak phase according to Cyclebase [27].

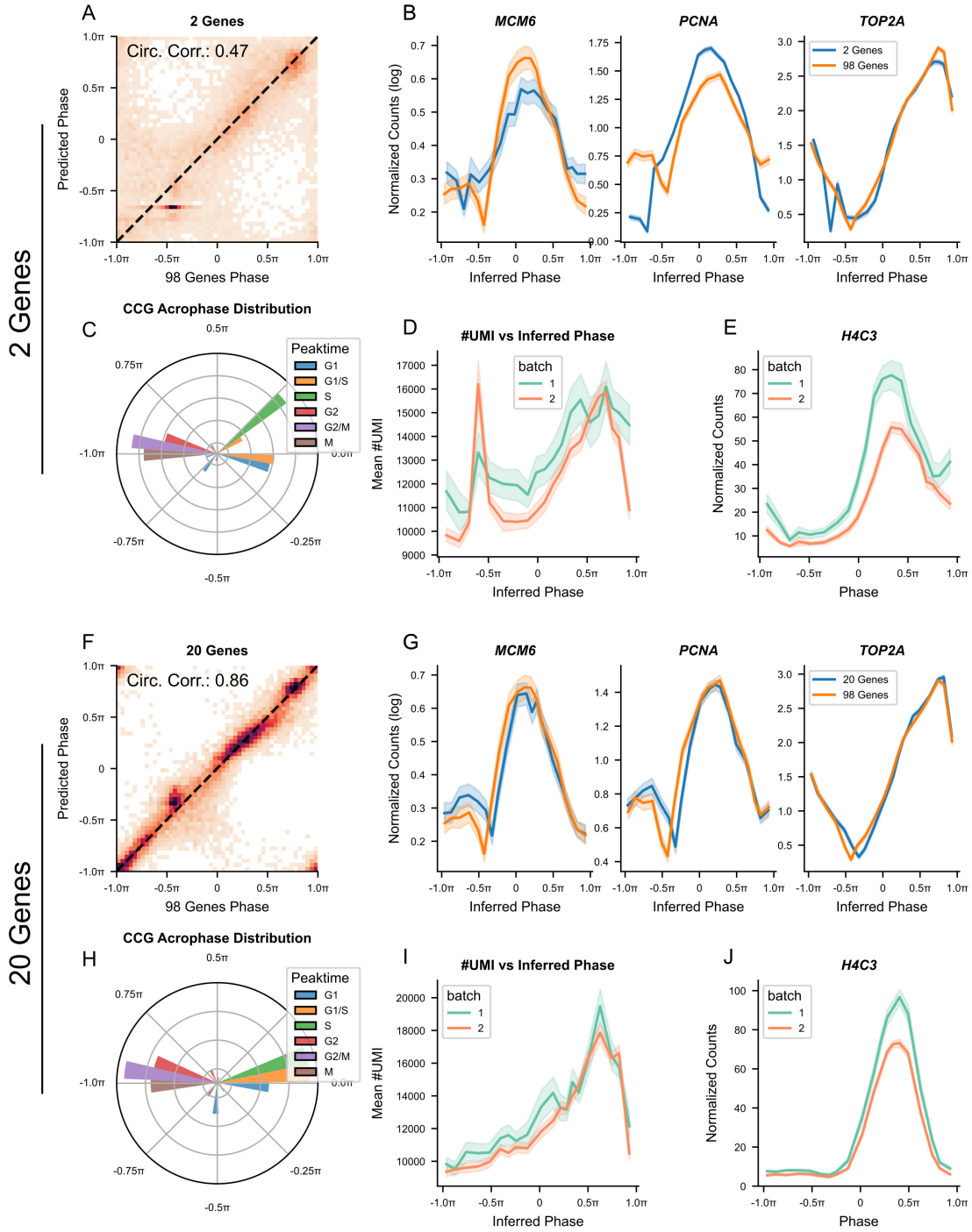

Fig. S3 Iterative refinement of the cycling gene set enables robust phase inference. Extended analysis of the RPE1 scRNA-seq dataset [28]. **A–E.** Initial model performance when fitted using only two rhythmic input genes (*TOP2A* and *PCNA*). **A.** Comparison between phases inferred by the 2-gene model and phases inferred by the full 98-gene model, colored by local density. Circular correlation = 0.47. **B.** Reconstructed expression profiles of selected cell-cycle genes (*MCM6*, *PCNA*, *TOP2A*) as a function of phase, comparing the 2-gene model and the full 98-gene model. Profiles were obtained by binning cells according to the inferred phase and averaging log-normalized count $\log(1 + \text{counts per } 10{,}000)$. **C.** Polar histogram of inferred peak phases for cell-cycle genes, colored by their annotated peak phase according to Cyclebase [27]. **D.** Mean number of unique molecular identifiers (UMIs) per cell as a function of the inferred phase, colored by batch. **E.** Mean fraction of counts computed from histone genes as a function of phase inferred, colored by batch. **F–J.** Improved model performance using a refined set of 20 genes. Genes were selected based on a score combining mean expression and inferred amplitude (see Methods). Panels are analogous to **A–E** and show the removal of artifacts observed in the initial model.

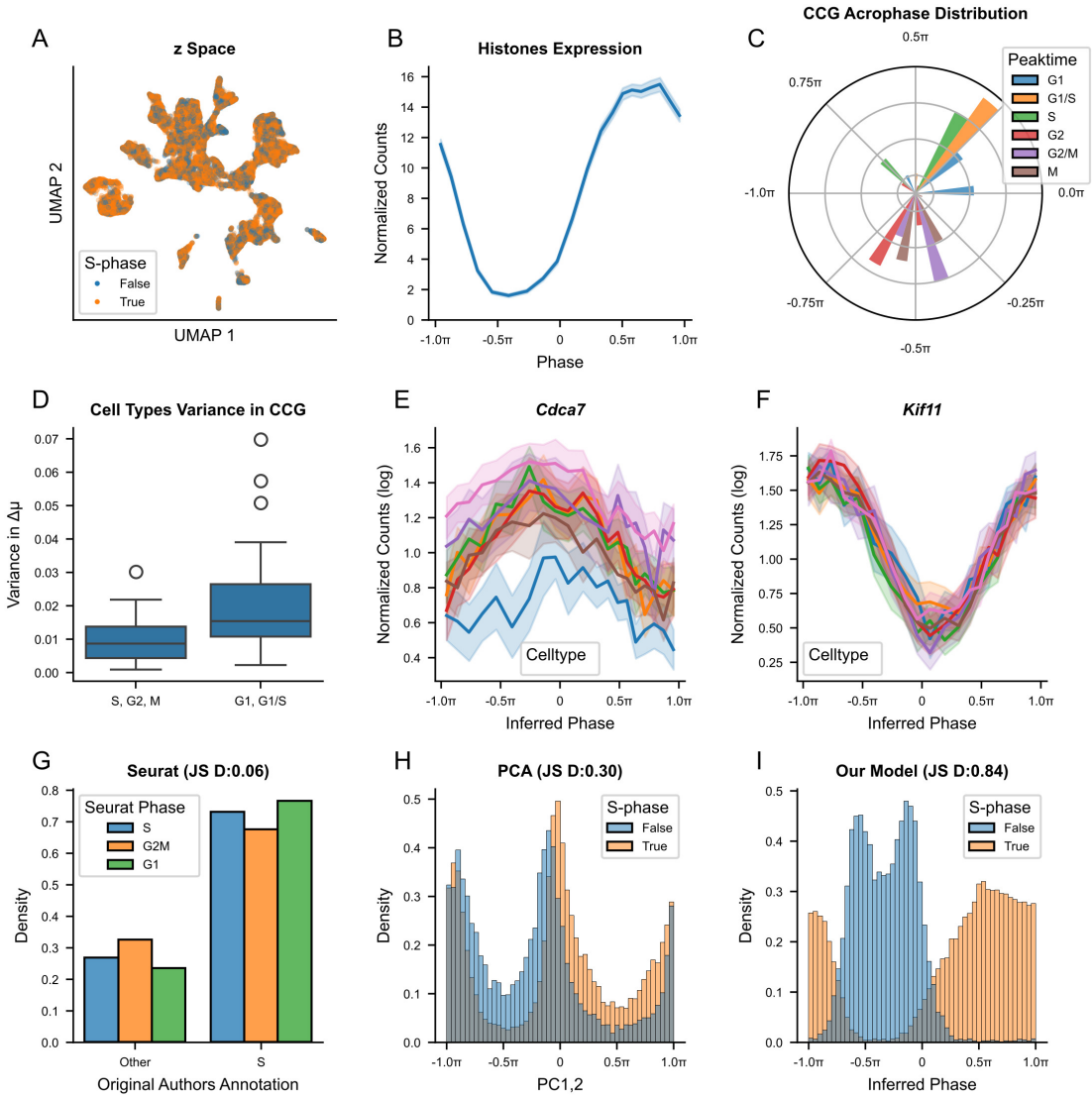

Fig. S4 Validation of context-corrected cell-cycle inference in mouse embryonic cells. Extended analysis of the mouse embryonic VASA-seq from [15]. **A** Projected contextual z-space, colored by original S-phase annotation (True/False). **B** Mean fraction of counts computed from histone genes as a function of inferred phase. **C** Polar histogram of inferred peak phases for cell-cycle genes, colored by their annotated peak phase according to Cyclebase [27]. **D** Boxplot displaying the variance between the mean inferred $\Delta\mu$ of the principal cell types (>1000 cells) for the cell cycle genes with a mean fraction of counts >1e-4, depending on their peaking cell cycle phase according to **C**. **E–F** Reconstructed expression profile as a function of inferred phase, comparing the different principal cell types. Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression ($\log(1 + \text{counts per } 10{,}000)$), for *Cdca7* (E), and *Kif11* (F). **G–I** Distribution of cells annotated as S phase along discrete phases obtained from Seurat Cell-Cycle Scoring (G), and continuous phases derived from the principal components 1 and 2 (H), and CoPhaser (I); the Jensen–Shannon divergence between the observed phase distribution and the original S-phase distribution is reported in each panel.

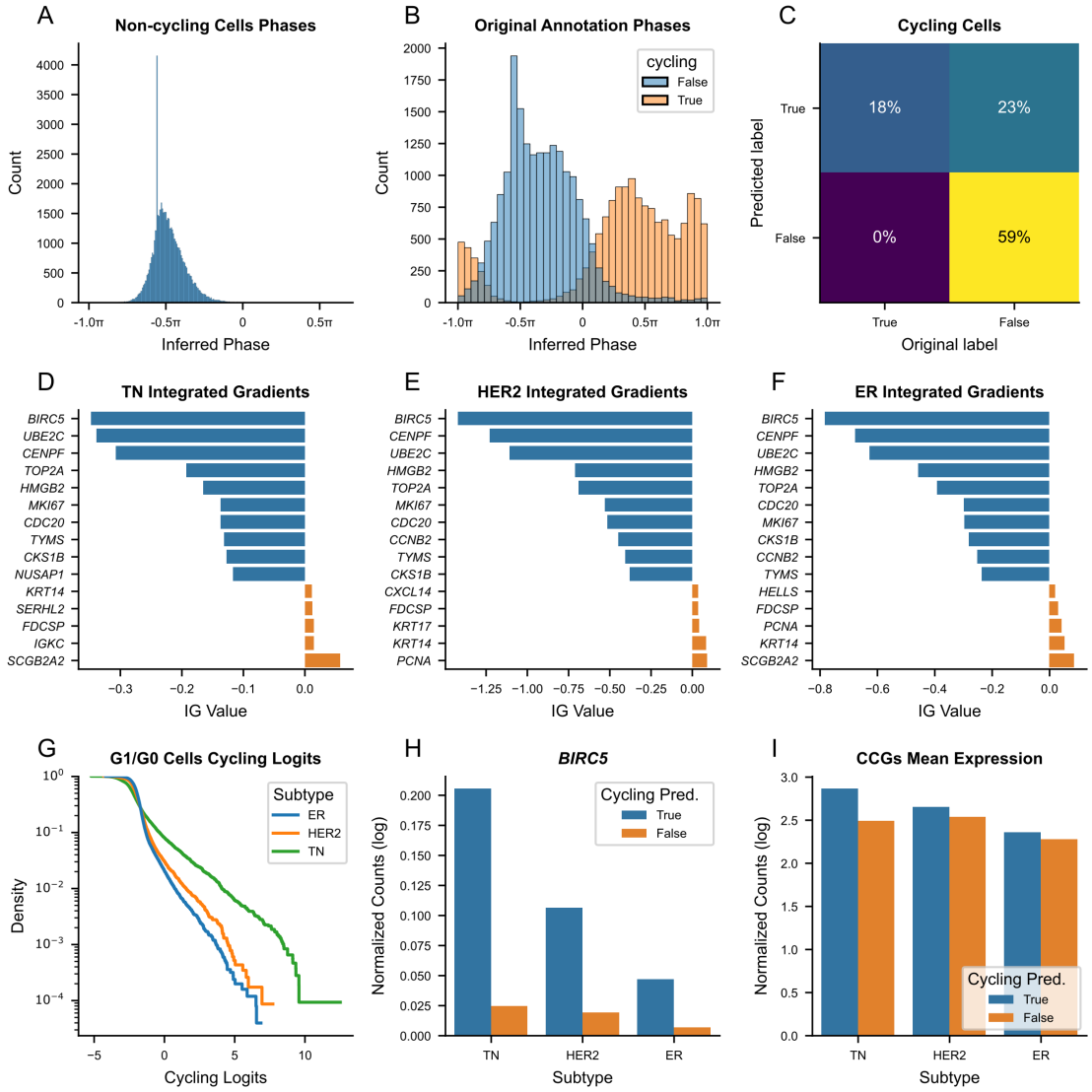

Fig. S5 Validation of the cycling state classifier in breast cancer. A. Distribution of inferred phases for cells classified as non-cycling by the Bernoulli VAE. **B.** Distribution of inferred phases for cells classified as cycling by our model, colored by the original cycling status annotation. **C.** Confusion matrix comparing our model-derived cycling status with original annotations. **D–F.** Top genes contributing to the cycling classification (identified via Integrated Gradients) for three breast cancer subtypes. **G.** Empirical cumulative distribution function (ECDF) of cycling logits for cells in the G0/G1 phase, colored by subtype. **H,I.** Mean expression of *BIRC5* (**H**) and aggregate cell-cycle genes (**I**) in mid-G1 cells ($[-0.7\pi, -0.4\pi]$), comparing cells with high versus low cycling probability (weighted average) across subtypes.

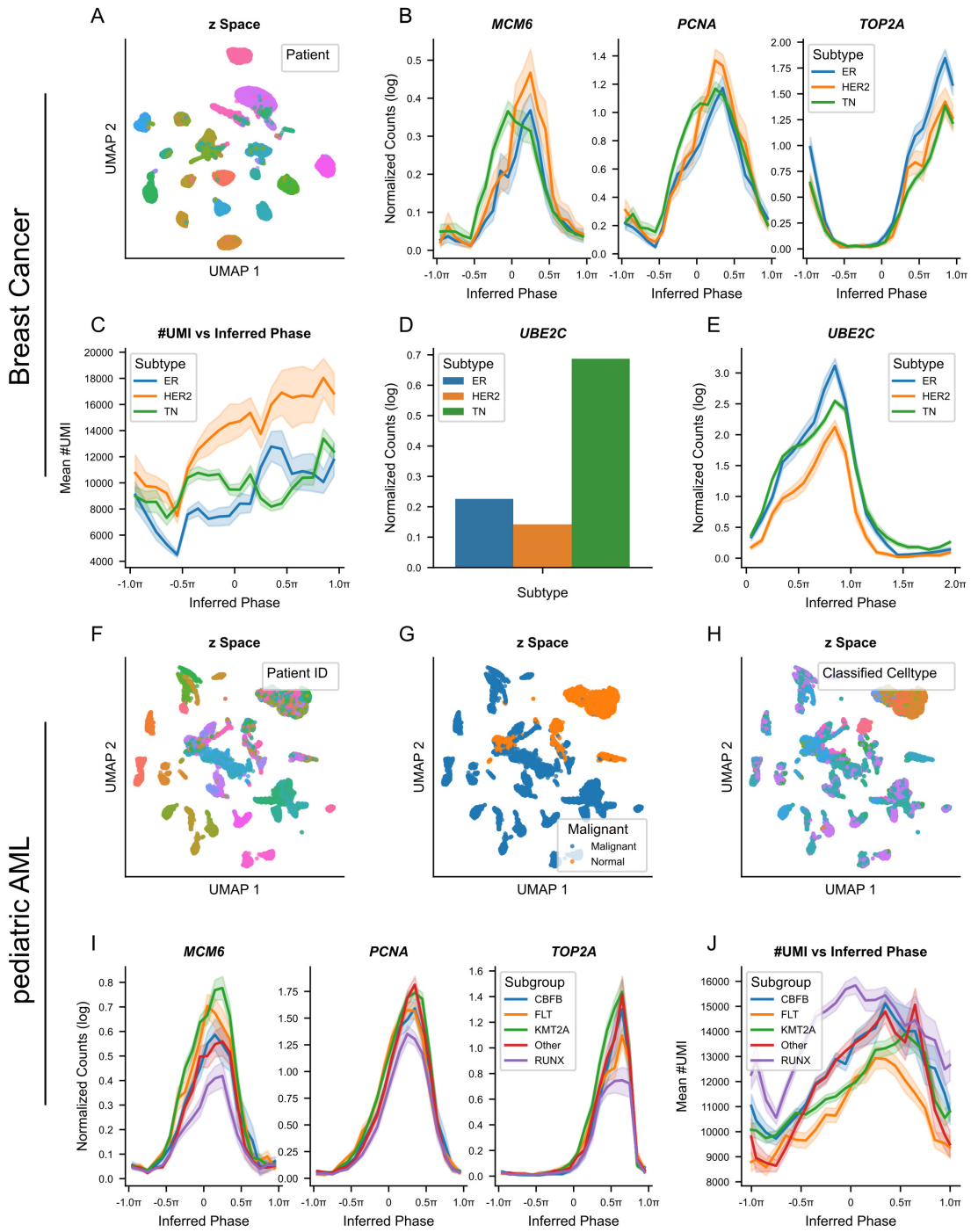

Fig. S6 Characterization of cell-cycle dynamics in cancer. A–E Extended analysis of the breast cancer scRNA-seq dataset from [51]. **A** Inferred contextual z-space colored by patient ID. **B** Reconstructed gene expression profiles for selected cell-cycle genes (*MCM6*, *PCNA*, *TOP2A*) as a function of inferred phase, colored by cancer subtype. Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression (log(1+counts per 10,000)). **C** Mean number of unique molecular identifiers (UMIs) per cell as a function of inferred phase, colored by cancer subtype. **D** Pseudobulk mean expression of *UBE2C* across cancer subtypes. **E** Reconstructed expression profile of *UBE2C* as a function of inferred phase, colored by cancer subtype. Profile were obtained as in **B**. **F–J** Extended analysis of the pediatric AML dataset from [34]. **F–H** Inferred contextual z-space colored by patient ID (**F**), malignancy status (**G**), and classified cell type (**H**). **I** Reconstructed gene expression profiles for selected cell-cycle genes (*MCM6*, *PCNA*, *TOP2A*) as a function of inferred phase, colored by tumor subtype (ER: Estrogen receptor positive, HER2: human epidermal growth factor receptor 2 positive, and TN: triple-negative). Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression (log(1+counts per 10,000)). **J** Mean number of unique molecular identifiers (UMIs) per cell as a function of inferred phase, colored by tumor subtype.

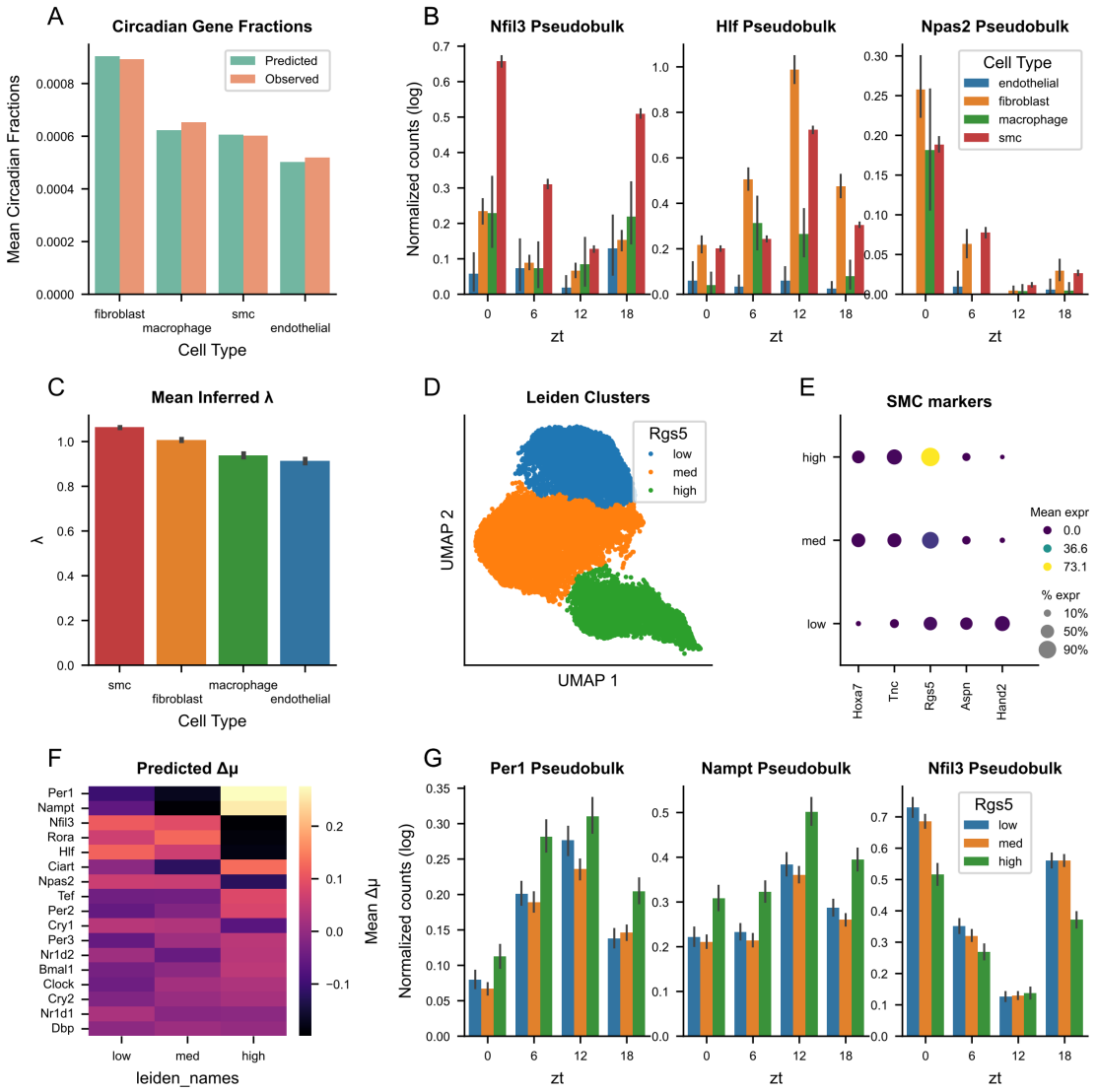

Fig. S7 CoPhaser enables robust circadian inference in rare cell types and reveals intra-population heterogeneity. Extended analysis of mouse aorta scRNA-seq data from [20]. **A,** Comparison of the mean fraction of counts derived from circadian genes (observed) versus the model prediction, across cell types. **B,** Pseudobulk expression profiles of *Nfil3*, *Hlf*, and *Npas2* as a function of Zeitgeber Time (ZT), colored by cell type. **C,** Mean inferred rhythmic amplitude correction (λ) across cell types. **D,** Inferred contextual z-space for the smooth muscle cell (SMC) population, colored by Leiden cluster. **E,** Dot plot showing the expression of regional and SMC markers across the SMC leiden clusters, annotated according to their mean *Rgs5* expression. **F,** Heatmap of inferred mean shifts ($\Delta\mu$) for circadian rhythmic input genes across the SMC Leiden clusters. **G,** Pseudobulk expression profiles of *Per1*, *Nampt*, and *Nfil3* as a function of ZT, colored by cell type.

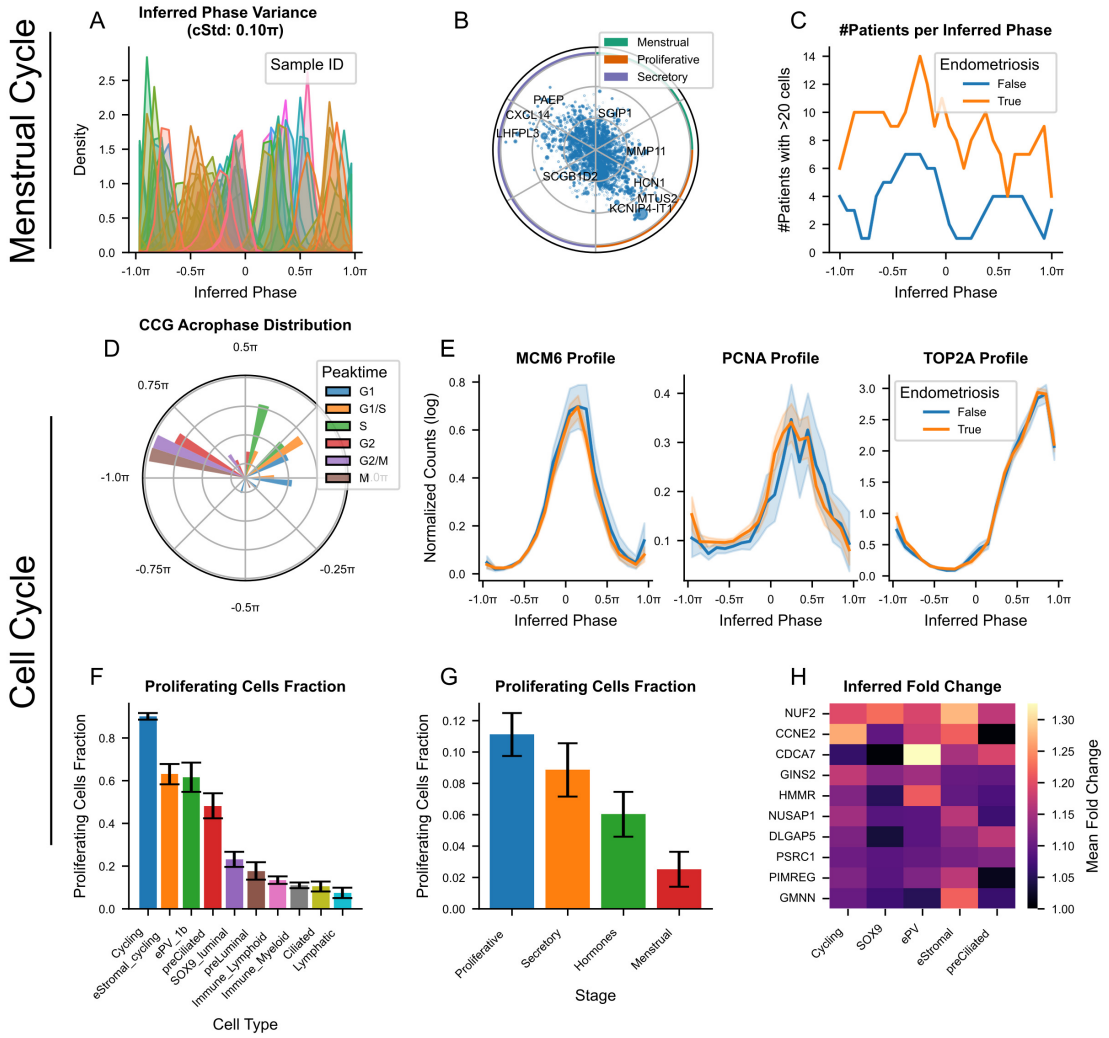

Fig. S8 Validation of menstrual cycle reconstruction and analysis of proliferation dynamics in the human endometrium. Extended analysis of the endometrium single-nucleus RNA-seq (snRNA-seq) data from [36]. **A,** Distribution of inferred phases colored by patient ID. The mean circular standard deviation (cSTD) of phases within patients is reported (0.10π). **B.** Polar scatter plot of inferred amplitudes and phases for the set of context genes, showing peak-to-trough amplitude (log) as radius and peak phase as angle. Point size is proportional to mean expression. Colored rings indicate the approximate menstrual cycle stages. **C,** Number of patients with more than 20 cells in each bin of the inferred phase, comparing healthy donors and donors with endometriosis. **D,** Polar histogram of inferred peak phases for cell-cycle rhythmic input genes, from all celltypes, colored by their annotated peak phase according to Cyclebase [27]. **E,** Reconstructed expression profiles of cell-cycle genes (*MCM6*, *PCNA*, *TOP2A*) as a function of inferred cell-cycle phase, comparing cells from healthy donors and donors with endometriosis (True/False). Profiles were obtained by binning cells according to their inferred phase and averaging log-normalized expression ($\log(1 + \text{counts per } 10{,}000)$). **F,** Fraction of proliferating cells across different cell types with error bars indicating the standard error of the mean (SEM) across patients. **G,** Fraction of proliferating cells across menstrual cycle stages, with error bars indicating the SEM across patients. **H,** Heatmap of the inferred mean absolute fold change in expression of cell-cycle genes between donors with endometriosis and healthy donors, across cell types.