Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cycle-phase supervision improves transferable physiological representations from wearable data for women's health

Rocio Mexia Diaz

r.mexiadiaz@hotmail.com

Terra Enabling Developers, Inc. (DBA TERRA API)

Faraaz Akhtar

Terra Enabling Developers, Inc. (DBA TERRA API)

Cameron Crawford

Terra Enabling Developers, Inc. (DBA TERRA API)

Alistair Brownlee

Terra Enabling Developers, Inc. (DBA TERRA API)

Kyriakos Eleftheriou

Terra Enabling Developers, Inc. (DBA TERRA API)

Article

Keywords: wearable data, menstrual cycle, reproductive health, representation learning, transfer learning

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10634632/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Cycle-phase supervision improves transferable physiological representations from wearable data for women's health

Rocío Mexía Díaz[1*], Faraaz Akhtar[1], Cameron Crawford[1], Alistair Brownlee[1], Kyriakos Eleftheriou[1]

[1]Terra Enabling Developers, Inc. (DBA TERRA API), Beaconsfield Street, London, United Kingdom.

*Corresponding author(s). E-mail(s): research@tryterra.co;

Abstract

The menstrual cycle can be tracked through physiological signals captured by consumer wearables overnight — skin temperature, breathing rate, heart rate and heart rate variability — which act as proxies for the underlying hormonal variations. Other reproductive states, such as pregnancy or menopause, are often studied in isolation rather than as perturbations of the normal menstrual hormone rhythm, even though they are all governed by the same endocrine system. Here, we present CHARM (Cyclic Hormone-Aware Representation of the Menstrual cycle), a framework that learns menstrual cycle representations from wearable data by combining two training objectives: reconstruction of masked sensor channels, and a menstrual cycle phase (luteal or follicular) supervised loss. Phase supervision improves recovery of the underlying hormone dynamics from the learned representation. Without further training, frozen CHARM embeddings also generalise to multiple reproductive states and symptoms not seen during model training. The phase objective substantially improves this generalisation and is highly label-efficient, matching a strong baseline while using around 5-fold fewer labels for perimenopause and outperforming the baseline on menopause across all data regimes.

Keywords: wearable data, menstrual cycle, reproductive health, representation learning, transfer learning

1 Introduction

The menstrual cycle is governed by the hypothalamic–pituitary–ovarian (HPO) axis. Gonadotropin-releasing hormone from the hypothalamus drives the pituitary gland to secrete luteinising hormone (LH) and follicle-stimulating hormone (FSH), which act on the ovary and regulate the secretion of the two hormones that shape the cycle, oestrogen and progesterone; these in turn feed back on the hypothalamus and pituitary [1]. Oestrogen levels rise through the late follicular phase to a pre-ovulatory peak and then drop to a smaller mid-luteal peak following ovulation. Progesterone, on the contrary, rises through the luteal phase to sustain the uterine lining; if implantation does not occur it collapses, triggering menses and a return to the follicular phase [2].

Different hormonal states such as pregnancy, menopause and birth control use have characteristic patterns that deviate from the HPO-dictated ovulatory rhythm [1, 3, 4]. In early pregnancy, the corpus luteum releases progesterone, maintaining its levels at a luteal-like high; the placenta then sustains this luteal-like state to term [3]. Hormonal contraception supplies exogenous progestin and oestrogen that signal the body is already in the luteal phase, thus suppressing the release of LH and FSH and inhibiting ovulation [4]. Exogenous hormones replace the natural follicular-to-luteal swing with a luteal-like progesterone plateau that drops during the placebo week [4]. The menopausal transition is characterised by high FSH and LH levels while oestrogen and progesterone become variable and then decline [1].

Although wearables cannot measure hormones directly, oestrogen and progesterone have well-characterised physiological effects that they do capture. Progesterone is thermogenic and raises core body temperature, breathing rate, and resting heart rate while suppressing heart-rate variability (HRV); oestrogen exerts the opposite, weaker effect [4–6]. Therefore, in the luteal phase, nocturnal skin temperature, heart rate, and breathing rate rise while HRV falls [7, 8]. Crucially, these variations are better described as a smooth, sinusoidal oscillation across the cycle rather than as a step change between follicular and luteal plateaus [9].

Reading the menstrual cycle from physiological signals captured by consumer wearables is now well validated [7, 10]. Meta-analyses have reported a pooled 0.88 accuracy for wearable fertility-window detection [10]. Established approaches fall into three families. The oldest and most widely used is the calendar method, which places menses and ovulation by cycle arithmetic alone, assuming a fixed fertile window and a constant luteal phase length [11]. The second family exploits the post-ovulatory temperature shift. Applied to ring-captured nocturnal temperature, a published filter-and-threshold algorithm locates menses onset and ovulation directly from the shape of the temperature curve [8], with a reported mean absolute error of 1.26 days on ovulation detection, against 3.44 days for the calendar method [12]. The third family applies machine learning, typically random forests, over engineered temperature, heart-rate and HRV features. A wrist multi-sensor model identified the six-day fertile window with 90% accuracy across 708 cycles [7]; a wrist temperature-and-heart-rate model reported a fertile-window ROC-AUC of 0.869 in regular cyclers, falling to 0.763 in irregular ones [13]; and a random forest over skin temperature, electrodermal activity and interbeat intervals separates three cycle phases at 87% accuracy (ROC-AUC

0.96), dropping to 68% (ROC-AUC 0.77) for day-level four-phase tracking, on 65 cycles from 18 participants [14].

Other reproductive states including pregnancy [15, 16], hormonal contraception [4], and the menopausal transition [16] have also been widely characterised by mapping the hormonal deviations from the HPO axis described earlier to the observed physiological effects of these hormones. However, no tool tracks these hormone-driven states together; a recent scoping review notes that no existing reviews concurrently cover events such as the menstrual cycle and the menopausal transition within one framework [17]. Representation learning offers a way to exploit the shared cyclic structure of the HPO-driven hormonal rhythm. Deep self-supervised models trained directly on consumer-wearable biosignals [18, 19] improve generalisation across cohorts and devices [20], and periodicity-aware objectives have been explored for other physiological and video signals [21, 22]; neither research direction targets the hormonal axis or the multi-week menstrual cycle.

The dominant data source for real-world reproductive health is self-reported tracking, which is inherently noisy and confounded by tagging and app engagement behaviours [23]. Objective ground truth data such as hormone measurements is scarce and the few datasets that pair this with wearable data are small (mcPHASES, 42 participants) [24]. More broadly, publicly available datasets for the different reproductive states (pregnancy, menopause, contraception) remain limited, which constrains any model trained for a single state [17]. Within reproductive health, models remain small-scale and single-task. These limitations motivate learning a menstrual cycle-aware representation of the wearable physiology that embeds hormone dynamics and transfers across reproductive states in low-data regimes, rather than training a separate model for each.

We present CHARM (**C**yclic **H**ormone-**A**ware **R**epresentation of the **M**enstrual cycle), a pretraining framework that learns a wearable representation of menstrual physiology from two objectives: a masked-channel self-supervised reconstruction loss and menstrual cycle phase (follicular versus luteal) supervision. The phase objective uses only the timing of menstruation and ovulation, yet grounds the learned representation in the cyclical hormonal rhythm and improves the recovery of measured hormone concentrations from wearable data. The resulting frozen representation transfers to multiple reproductive conditions and symptoms that are held out during pretraining, including pregnancy and menopause, and substantially improves label efficiency across downstream tasks. By learning a shared representation from normal cycling physiology, CHARM provides a unified foundation for modelling multiple reproductive health states rather than training separate models for each condition.

2 Results

2.1 Wearable signals characterise the menstrual cycle and its perturbations

Before evaluating the representation, we confirm that our wearable cohort reproduces the physiological signatures of the menstrual cycle and its perturbations reported in prior work (Figure 1).

Across the ovulatory cycle, night-time skin temperature, heart rate and breathing rate drop around menses while HRV rises, aligning with the follicular transition and the fall in progesterone (Figure 1a), reproducing the established cycle physiology [7, 8]. In the first 10 weeks of pregnancy, temperature, heart rate, and breathing rate all rise steeply while HRV drops, aligning with the high corpus luteum progesterone levels. Temperature and breathing rate then stabilise relative to pre-pregnancy levels while heart rate and HRV rise and decrease respectively up to term (Figure 1b). Hormonal contraception replaces the natural follicular-to-luteal oscillation with a luteal-like plateau that drops during the placebo (bleeding) week (Figure 1c) [4]. Finally, during the menopausal transition, there are two distinct cohorts: perimenopause and menopause. Perimenopausal users are still cycling but sit at a markedly lower HRV level than younger cycling users (Figure 1d), and established menopause loses the monthly rhythm altogether, visible as the disappearance of the ∼28-day autocorrelation peak (Figure 1e).

2.2 Phase representation improves period and ovulation detection

On held-out test users, a linear probe trained on CHARM embeddings places period onset within ±3 days of the self-reported event for 78.8% of onsets, and LH-confirmed ovulation within ±3 days for 75.9% of events, with mean absolute errors of 3.68 and 3.13 days respectively (Table 1). The performance lift attributed to the phase loss is quantified by the difference in performance between CHARM and CHARM (recon-only) — an ablation of our model trained without the phase loss, i.e. with a reconstruction objective only. On period onset, the phase objective lifts performance by 9.8 percentage points at ±3 days; on ovulation the corresponding gain is 15.3 points.

Compared to existing work, our method was the best performing for period detection, followed by the random forest fitted to engineered daily features [7, 14]. It also leads all other methods for mean absolute error in both tasks. The calendar and temperature methods, however, achieve the highest accuracy for ovulation detection and place more events inside a tight tolerance window (82.1% and 78.6% within ±3 days).

Table 1: Mean absolute error and accuracy within a tolerance window (±1d, ±2d, ±3d) of CHARM compared against established methods. Metrics are reported matched to a comparable prediction rate of approximately one event per 28 observed days.

	Period onset (N = 33,591)				Ovulation, LH-confirmed (N = 307)			
Method	MAE (d)	±1 d	±2 d	±3 d	MAE (d)	±1 d	±2 d	±3 d
CHARM	**3.68**	**0.594**	**0.721**	**0.788**	**3.13**	0.456	0.645	0.759
CHARM (recon-only)	5.04	0.480	0.612	0.690	4.57	0.300	0.443	0.606
RF on daily features [7, 14]	4.32	0.585	0.698	0.759	6.22	0.293	0.450	0.547
Maijala et al. [8]	5.45	0.375	0.551	0.659	3.56	0.509	0.698	0.786
Calendar [11]	8.76	0.291	0.434	0.538	4.84	**0.551**	**0.734**	**0.821**

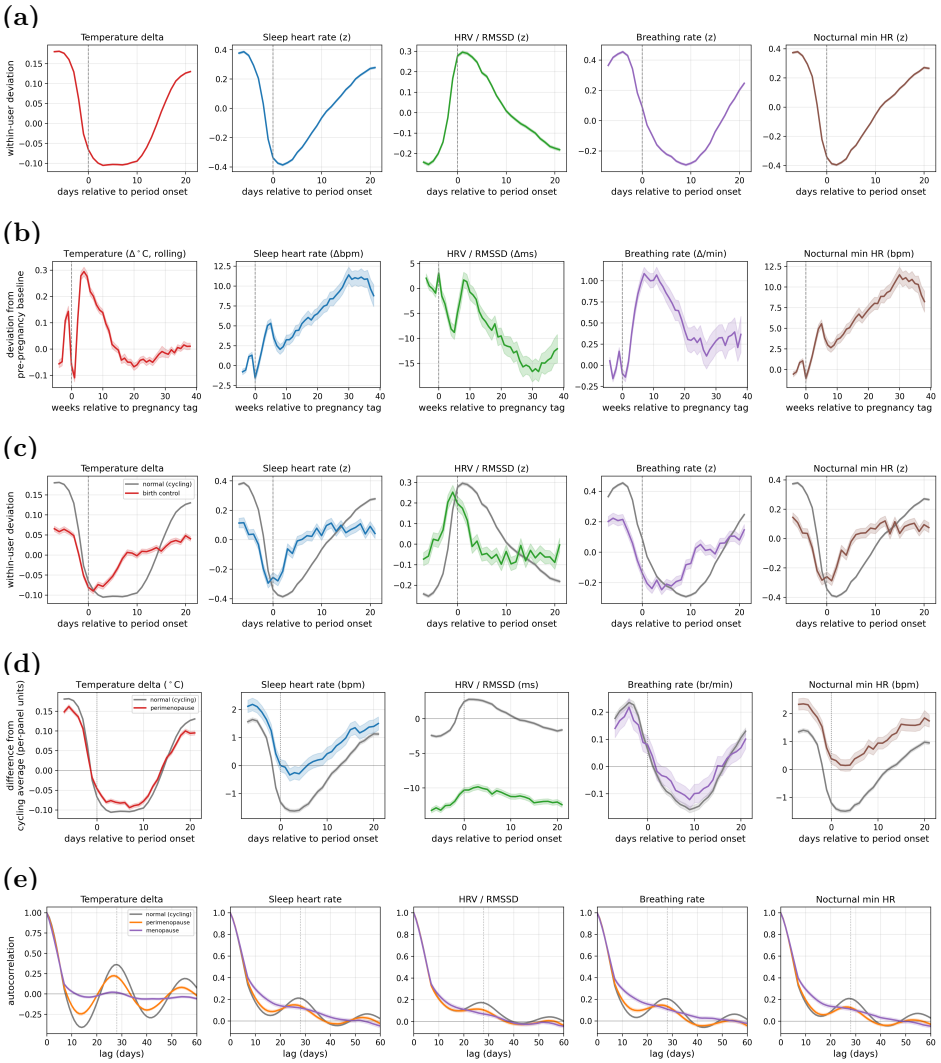


Fig. 1: Wearable physiology biomarkers (temperature, average and minimum heart rate, HRV and breathing rate) characterise the menstrual cycle and its perturbations (Terra API cohort). **(a)** The ovulatory cycle, within-user and period-aligned. **(b)** Pregnancy, as deviation from each user's pre-pregnancy baseline. **(c)** Hormonal contraception versus normal cycling physiology. **(d)** Perimenopause relative to the cycling average, period-aligned. **(e)** Autocorrelation of the 7-day-smoothed signals compared with cycling users; perimenopausal users in orange, menopausal users in purple. Lines are within-user-centred means and shaded bands show ± 1 standard error of the mean throughout **(a)**–**(d)**. In **(a)** (n = 24,825 cycles) the standard error is narrower than the plotted line.

2.3 Phase representation transfers to reproductive conditions and symptoms

Table 2 compares the downstream performance of a linear probe fitted on the frozen CHARM embeddings against a LightGBM probe on raw data and the phase ablation, CHARM (recon-only). On the held-out test set, our embedding produces the highest ROC-AUC across all tasks. The paired performance difference between CHARM and its phase ablation, bootstrapped over user clusters and corrected for multiple comparisons (Benjamini–Hochberg), demonstrated that the phase objective significantly improves the representation on all tasks (BH-corrected $p < 0.05$). The biggest lift provided by the phase loss is 8.7 percentage points, in the detection of spotting. It is the phase objective that also lifts the performance above raw-window LightGBM probe on every task except for the pooled menopausal transition, perimenopause and PMS.

Table 2: Comparison of condition and symptom (within-user) transfer ROC-AUC between CHARM, the phase ablation, CHARM (recon-only), and raw-window LightGBM (RAW-gbm). N is the number of tag-positive users in the cohort. * marks the tasks on which the phase loss significantly improves performance, bootstrapped over user clusters and Benjamini–Hochberg corrected.

Block	Task	N	RAW-gbm	CHARM (recon-only)	CHARM
Cond	Pregnancy	1177	0.940	0.913	**0.952***
Cond	Menopausal transition	845	0.768	0.784	**0.826***
Cond	Menopause	336	0.885	0.882	**0.930***
Cond	Perimenopause	425	0.678	0.704	**0.752***
Cond	Birth control	417	0.754	0.708	**0.762***
Sym	Cramps	963	0.688	0.651	**0.702***
Sym	PMS	352	0.500	0.553	**0.589***
Sym	Spotting	327	0.581	0.548	**0.635***
Sym	Hot flashes	341	0.496	0.452	**0.521***

2.4 Low data regime performance

Across pregnancy, menopause, and perimenopause CHARM achieves the highest ROC-AUC at every label budget compared to the raw data probe (Figure 2). The value added by the phase objective (the gap between the blue and orange lines) is large for every task, lifting performance by at least 4 percentage points at every budget except perimenopause at the smallest, where the gain is 1.3 points. The phase objective is also responsible for the lift of performance above the raw-data floor for pregnancy and menopause.

To evaluate the label efficiency of CHARM over the raw-data baseline, we take the performance of our embedding at the smallest budget of 10 labelled positive users and compare how many users the raw baseline needs to match it. For pregnancy, we achieve 0.913 with 10 labels, which the raw baseline does not match until ∼25 (a 2.5-fold difference). For perimenopause, our embedding performance starts at an

ROC-AUC of 0.614 at 10 labels, which the raw-data baseline requires ∼50 to match (a 5-fold difference). The raw-window data never achieves the performance obtained with our embedding for menopause, while birth control has overlapping performance throughout the data budgets.

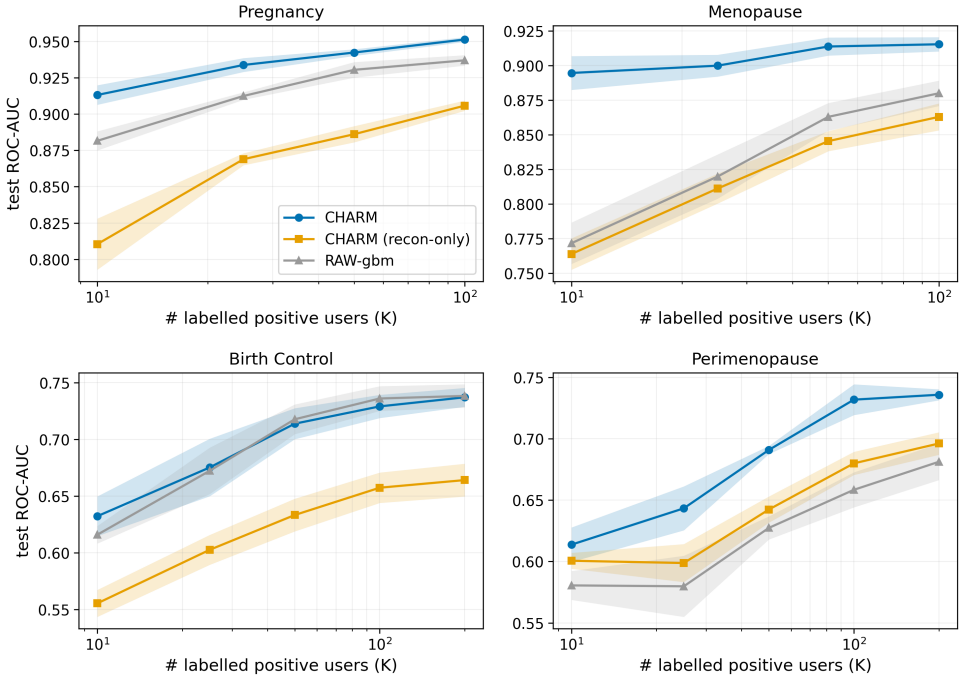


Fig. 2: Test ROC-AUC versus the number of labelled positive training users, across four conditions. Lines are means over 5 seeds and bands are ±1 standard error of the mean. A linear probe on the frozen CHARM embeddings (blue) is compared against its ablation (orange) — CHARM trained with a reconstruction objective only — and a raw-window LightGBM (grey).

2.5 Hormone recovery from the embeddings

The average hormone trajectories reconstructed by a linear probe fitted on CHARM embeddings and its phase ablation are shown in Figure 3. The nightly channels carry a strong progesterone signal and oscillate with the cycle. Averaging across cycles cancels the day-to-day noise, resulting in a Spearman ρ above 0.8 for both CHARM and its ablation on both oestrogen and progesterone. Therefore, we instead report the held-out coefficient of determination (R^2), computed per day on held-out subjects without cross-cycle averaging, which isolates the gain contributed by the phase objective. LH recovery scores a Spearman ρ of 0.6.

The phase objective significantly sharpens the recovery of all three hormones ($p < 0.004$). Oestrogen (E3G) R^2 rises from 0.052 to 0.078; progesterone (PdG) rises from 0.158 to 0.217; and luteinising hormone (LH) from 0.017 to 0.036.

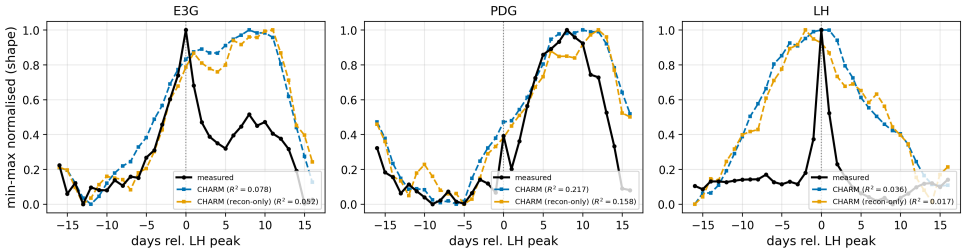


Fig. 3: mcPHASES cycle-averaged trajectories aligned to the LH-surge peak. Measured hormone (black) vs. recovered from the frozen representation by a linear probe, for CHARM (blue) and CHARM (recon-only) (orange), min-max normalised. Each legend gives the per-encoder held-out R^2.

2.6 Temperature masking robustness

The observed performance drop when masking the temperature channel at inference is shown in Table 3. Performance degrades most on birth control (-0.109) while hot flashes and PMS have very small variations. The remaining states drop $\sim$0.04–0.07. Even fully temperature-masked, pregnancy and menopause remain well above chance (> 0.80).

Table 3: Cost of removing temperature on downstream task accuracy. The temperature channel and its observation mask are zeroed at inference on the same test-split users; $\Delta = \text{AUC}_{\text{masked}} - \text{AUC}_{\text{full}}$.

Task	full	temp-masked	Δ
Birth control	0.762	0.654	-0.109
Menopause	0.930	0.858	-0.071
Pregnancy	0.952	0.894	-0.058
Menopausal transition	0.826	0.771	-0.055
Cramps	0.702	0.647	-0.055
Spotting	0.635	0.589	-0.045
Perimenopause	0.752	0.708	-0.044
PMS	0.589	0.565	-0.023
Hot flashes	0.521	0.535	$+0.014$

3 Discussion

Two important hormones that regulate the menstrual cycle and have an effect on physiology are oestrogen and progesterone. The luteal phase is the progesterone-driven state of the cycle, and has a clear thermogenic and cardiac signature in wearable data [4, 5]. In this study, we show that supervising on this cycle phase produces CHARM embeddings, a hormone-grounded frozen representation that improves downstream task performance.

A linear probe fitted on our frozen embeddings places both period onset and ovulation within ±3 days for over three quarters of events (Table 1). This performance ceiling can be expected for several reasons. Firstly, the model is not fine-tuned; the encoder is frozen and only a small probe is trained. Secondly, most cycle-tracking studies report higher accuracies because they exclude irregular participants [7, 11, 17], whereas here they are deliberately retained and the positives are noisy self-reported events. Consistent with this, a recent wearable fertile-window model reports a comparable regular-versus-irregular gap (AUC 0.87 vs 0.76) [13]. A pooled meta-analytic accuracy of 0.88 on curated, regularly-cycling cohorts [10] is a reasonable ceiling for this task, although it is measured against a different cohort and evaluation framework, so it is not directly comparable.

Measured against the established methods on identical events (Table 1), CHARM outperforms other methods on menstrual onset, locating 78.8% of onsets within ±3 days against 65.9% for the temperature algorithm and 53.8% for the calendar. A random forest over lags and rolling statistics of the same channels reaches 75.9% on the same task, under ∼3% behind, demonstrating the advantage of machine learning models over methods relying on brittle assumptions about homogeneous luteal phases. On LH-confirmed ovulation CHARM achieves the lowest mean absolute error of any method (3.13 days), yet the calendar and temperature methods place more events inside a tight window (82.1% and 78.6% within ±3 days, against 75.9%). Both of these observations are consistent: the calendar method is near-exact when the following menses onset is recorded but fails entirely for the confirmed ovulations occurring near the end of a user's record, where no subsequent onset exists to anchor to. CHARM requires no future onset, predicting from physiology alone.

The small sample size of LH-confirmed ovulation events makes machine learning methods likely to overfit and fail to generalise, which results in a higher accuracy for the calendar method compared to machine learning methods. The random forest behaved very differently across the two events, reaching 75.9% on menstrual onset but only 54.7% on ovulation, which corroborates our hypothesis. CHARM embeddings, however, did not behave differently across tasks. This contrast demonstrates that the benefit of pretraining is concentrated where labels are scarce, which is further validated by the transfer tasks.

The transfer tasks have no established benchmark or baseline model to compare against — a systematic review of wearable menstrual and fertility tracking covering 13 studies finds work on ovulation, menstruation and fertile-window detection only, and none on menopausal staging, contraception or symptom prediction [25]. However, CHARM's phase objective significantly increases the ROC-AUC by up to 8.7 percentage points over the reconstruction objective alone, which demonstrates the significance

of the phase loss. The practical payoff of a shared representation is that it makes these states learnable from very few labels: the CHARM probe is more label-efficient than LightGBM on raw data across conditions, most so where labels are scarcest (menopause; Figure 2).

The phase loss behaves like a hormonal prior, which was tested by probing our frozen representation against measured hormones. Of the three HPO hormones recorded on the mcPHASES dataset, only oestrogen and progesterone have physiologically measurable effects a wearable can record: progesterone is thermogenic and oestrogen has the opposite, weaker effect [4, 5]. The linear probe recovers both dynamics with over 80% Spearman correlation, and the phase objective significantly sharpens all three signals ($p < 0.004$) beyond reconstruction. Luteinising hormone is a transient pituitary signalling pulse: it triggers ovulation, but the autonomic and thermal signature around its surge is ultradian (2–5 h) rather than a sustained daily effect, and daily wearable summaries do not track it [26]. This is why the Spearman correlation for LH sits lower than oestrogen and progesterone.

Finally, because different devices expose different channels, robustness to a missing channel is a proxy for cross-device transfer, and the masked-channel reconstruction objective was chosen partly for this reason. Masking temperature at inference (Table 3) degrades the encoder gracefully rather than collapsing it. In the case of PMS and hot flashes, the performance degradation due to masking is almost negligible, likely because the encoder was already performing at near chance.

Prior wearable work models each reproductive state separately and treats the others as confounds to exclude [15, 16], while no review has yet summarised methods that track these hormone-driven states jointly [17]. The main contribution of this work is the CHARM pretraining framework rather than any specific architecture: a hormone-grounded phase objective that requires no hormone measurements and relies only on cycle timing. By combining this objective with masked-channel reconstruction, CHARM learns a shared representation that transfers across reproductive states and symptoms with substantially greater label efficiency and improved physiological interpretability.

Several limitations remain. Cross-device validation is needed to establish real-world generalisation beyond the missing-channel experiments presented here. The downstream labels are noisy self-reported events and should be validated against external ground truth, including hormone panels and clinician-confirmed pregnancy and menopausal staging.

4 Methods

4.1 Dataset

Our cohort consists of 38,064 users, comprising a total of 8,308,324 days. Users wore commercially available ring-based wearables connected through Terra API, which aggregates daily health metrics and per-session sleep data whenever the device is worn overnight. This sampling frame carries the demographic bias common to premium consumer wearables: owners skew toward higher socioeconomic status and greater health

Table 4: Terra API cohort composition. Each user is assigned one mutually-exclusive reproductive group from their self-reported events, so the group rows partition the cohort and sum to the total. Observed days are user-days of wearable data contributed by that group. Symptoms are single-day events logged by users in any group; their day counts are the number of days carrying that tag, not the length of the taggers' records.

Group	Users	Observed days
Cycling	35,625	7,600,817
Pregnancy	1,177	346,996
Menopausal transition	845	245,427
perimenopause only	425	125,059
menopause only	336	95,684
both stages logged	84	24,684
Birth control	417	115,084
Total	38,064	8,308,324
Symptom	Users	Days logged
Cramps	3,715	10,275
PMS	1,325	4,579
Spotting	1,227	3,323
Hot flashes	736	4,696

engagement, so the cohort is not representative of the general population. Terra API does not store age or demographics, so these are not controlled for directly.

Users are grouped into four mutually exclusive groups: cycling, pregnancy, menopausal transition or birth control. Table 4 lists the cohort groups' sample sizes and the number of symptom tags available. The menopausal group is broken down into perimenopause and menopause but users with both tags exist and are excluded from the smaller sub-groups. Cycling users are split into train, validation and test sets in a 70:15:15 ratio at the user level to pretrain the encoder. The test set is used to evaluate period and ovulation event detection, and is included in symptom detection evaluation. The whole pregnancy, menopause and birth control groups are used for downstream task performance evaluation (conditions and symptoms), and never used in pretraining.

The per-channel transforms are listed in Table 5: four of the five channels are converted to a per-user 30-day rolling z-score (minimum 7 valid days), so the model sees within-person deviations rather than absolute levels, while temperature delta is already a baseline-relative delta and is kept in native units. Each channel is paired with a binary observation mask. Phase labels for supervision were generated from the period and ovulation event tags such that the luteal phase begins at ovulation and the

Table 5: The five daily input channels: what each measures and how it is preprocessed before entering the encoder. All rolling statistics are causal (trailing) and computed per user.

Channel	What it measures	Preprocessing
Temperature delta	Night skin temperature as a deviation from the individual's own rolling (monthly) baseline (°C)	Kept in native units (already a baseline-relative delta)
Sleeping heart rate	Mean heart rate over the sleep period (bpm)	Per-user 30-day rolling z-score
Breathing rate	Mean respiratory rate during sleep (breaths/min)	Per-user 30-day rolling z-score
HRV (RMSSD)	Mean overnight heart-rate variability during sleep (ms)	Per-user 30-day rolling z-score
Resting heart rate	Lowest (nocturnal-minimum) heart rate during sleep (bpm)	Per-user 30-day rolling z-score

follicular phase at menses. Ovulation is rarely labelled, and in the absence of a tag is fixed 13 days before the next period.

Throughout, we call a cycle regular when its length falls within the conventional 21–35 day range, and irregular otherwise [27]. The fixed 13-day luteal assumption above is not supportable outside that range, so phase labels are only assigned on regular cycles; days belonging to irregular cycles carry no phase label and are masked out of the phase loss, while still contributing to the reconstruction objective. Irregular cyclers are not excluded anywhere else: they remain in the pretraining cohort, in the train/validation/test split, and in every downstream evaluation.

The external, device-independent hormone reference is mcPHASES [24] (daily urinary E3G/PdG/LH metabolites paired with Fitbit wearable data, 42 participants and 184 LH-aligned cycles, of which the 161 with a resolvable LH surge enter the trajectory analysis). Because E3G is a urinary metabolite of oestradiol, its excretion lags the serum oestradiol peak, so in our LH-aligned data the E3G maximum coincides with the LH-surge day rather than preceding it as serum oestradiol does [28]. It is never used for training; it serves only offline, to probe how recoverable each hormone is from the learned embedding.

4.2 Architecture and Pretraining

The encoder is a causal temporal convolutional network [29, 30] operating on a 90-day window of the 10 input channels. The 90-day window yields a receptive field of 61 days which covers roughly two menstrual cycles — long enough to expose the repeating follicular-to-luteal structure the phase objective supervises, while also reaching back to capture the slower progression of states such as early pregnancy and the menopausal transition. A 1×1 convolution projects the 10 input channels to 128 channels, which then pass through four residual temporal blocks with exponentially increasing dilations $1, 2, 4, 8$. Each block applies two dilated 1-D convolutions (kernel size 3) with ReLU, dropout, and a residual connection. Convolutions are left-padded to avoid future leakage if the input sequence is shorter than 90 days. A final 1×1 convolution produces the per-day embedding $z_t \in \mathbb{R}^{128}$. Phase classification and reconstruction heads are

attached to this 128-dim embedding which project it into a 1-dim phase output, and a 5-dim reconstruction output.

CHARM is pretrained for up to 50 epochs with AdamW (learning rate 3×10^{-4}, batch size 256) and early stopping on the validation loss (patience 4). Each temporal block uses dropout 0.2, and training windows are drawn from the 90-day input at a 4-day stride. The self-supervised signal is a masked-channel reconstruction loss: a whole channel per sample is hidden with probability 0.3 and the reconstruction head recovers it from the remaining channels under a masked MSE objective on the hidden entries, forcing z_t to encode cross-channel structure and improving robustness to missing inputs. The supervised signal is a menstrual-cycle phase head: each day is labelled luteal or follicular and trained with a binary cross-entropy objective. The full recon+phase objective is a weighted sum of the reconstruction loss (weight 0.5) and the phase cross-entropy. The downstream probes are an L_2-regularised logistic regression on the frozen representation, and a LightGBM probe on the raw-window. A secondary pretraining regime or ablation, CHARM (recon-only), pretrains the encoder without the phase loss, i.e. with the reconstruction objective alone.

4.3 Experiments

We begin by confirming that the cohort reproduces the physiological signatures of the menstrual cycle and its perturbations that are already established in the literature. For the ovulatory cycle, each channel is centred within user, by subtracting that user's own mean, and then averaged over cycles aligned at menses onset, so the curve shows the population's cyclical deviation shape rather than absolute levels. Pregnancy is aligned instead at the first pregnancy tag and expressed as a deviation from each user's pre-pregnancy baseline, restricted to gestations with a recorded birth or loss tag so that the window closes at delivery. Hormonal contraception and perimenopause are period-aligned and contrasted against the cycling population average. Finally, cyclicality itself is quantified as the autocorrelation of the 7-day-smoothed channels, comparing cycling, perimenopausal and menopausal users; the loss of the $\sim$28-day autocorrelation peak indexes the loss of the monthly rhythm. Bands are ± 1 standard error of the mean throughout.

The encoder representation, CHARM, is then frozen and evaluated on a set of downstream tasks. We first measure how accurately period onset and ovulation can be located within the cycle, and compare this against the established methods used in the wearable literature. A small L_2-regularised logistic-regression probe is attached to the frozen representation and emits a per-day event probability, evaluated on the held-out test set. We convert these probabilities into discrete predicted event days by rank-and-suppress peak picking. For each user, a budget of one event per 28 observed days is set from the length of their record. The highest-probability remaining day is taken as a predicted event, a ± 10-day window around it is then suppressed so that neighbouring days belonging to the same peak cannot also be selected, and the procedure repeats until the budget is spent.

The metrics used to score period and ovulation detection are mean absolute error (MAE), which scores the number of days separating the prediction from the real event, and the fraction of events located within ± 1, ± 2 and ± 3 days. Ovulation tags are

restricted to days on which the user logged a positive urinary luteinising-hormone test, which is the reference standard used by the wearable studies we compare against.

CHARM is compared against the three families of established cycle-tracking methods on the same test users. The calendar baseline uses cycle arithmetic alone; it predicts the next menses onset as the previous onset plus the population mean cycle length, and ovulation is placed 14 days before the following menses. The temperature baseline reimplements the algorithm of Maijala et al. [8], the only published method specified in full and developed on the same device. The daily temperature series is linearly interpolated and low-pass filtered with a second-order zero-phase Butterworth filter at a cutoff of 1.5/28 samples per day, menses onset is placed at the midpoint of the falling edge between adjacent extrema, and ovulation at the midpoint of the rising edge. Finally, a random forest is trained using lags, first differences, rolling means and rolling dispersions of the same five daily channels. This last baseline follows the third family of methods, in which a random forest is fitted to engineered daily features [7, 14], and is motivated by the meta-analytic finding that random forests are the best-performing algorithm class for this task [10]. It is an analogue of those approaches rather than a reproduction of either: both rely on proprietary feature sets and on channels unavailable in this cohort, such as electrodermal activity and sub-daily interbeat intervals.

All methods, including ours, are scored retrospectively over the complete record, as in the studies being compared against. The encoder is causal, but neither the peak picking, which selects a global maximum, nor the temperature baseline's zero-phase filter, nor the calendar's use of the observed next onset, is causal.

We then measure transfer to the reproductive conditions and symptoms that were never seen during pretraining, and test whether the phase objective contributes to that transfer beyond reconstruction alone. To evaluate the downstream performance of our representation embedding and the isolated gain provided by the phase loss, both CHARM and CHARM (recon-only) are frozen and a small linear (logistic-regression) probe is attached. The probe is trained for different downstream tasks that had not been seen by the model during pretraining, and benchmarked against a LightGBM probe trained directly on the flattened raw window — so even a linear probe on the frozen representation is compared against a stronger, non-linear model on the raw signals.

This experiment measures detection of reproductive states (pregnancy, menopause, birth control) and symptoms (PMS, cramps, hot flashes, spotting) never seen during pretraining, reported as ROC-AUC, from a 5-fold GroupKFold pass. Negative samples are obtained from the held-out test set of cycling cohort for the conditions, and from non-tagged days from the same user for the symptoms. Positive samples are 90-day windows, stepped every four days, where at least the last day is positive. The paired difference $\Delta\text{AUC} = \text{CHARM} - (\text{recon-only})$ is computed per task, testing whether the phase objective adds anything beyond reconstruction, with a user-cluster bootstrap and Benjamini–Hochberg correction across the nine reported tasks.

Three further experiments characterise the representation rather than its accuracy: how efficiently each task can be learned from few labels, how much the encoder depends

on the temperature channel, and how much of the underlying hormonal signal the latent space retains.

The same small linear probe as the transfer tasks is attached to the frozen CHARM and CHARM (recon-only). Label efficiency is reported as ROC-AUC against the number of labelled positive training users, averaged over five seeds at a fixed held-out test set.

Masking the temperature channel measures how much the encoder relies on temperature by zeroing that channel and setting its observation mask to missing at inference, and reporting the change in condition and symptom ROC-AUC, $\Delta = \mathrm{AUC}_{\mathrm{masked}} - \mathrm{AUC}_{\mathrm{full}}$.

Finally, a linear probe is fitted to the frozen CHARM and CHARM (recon-only) encoders and trained to recover the measured mcPHASES hormones E3G, PdG and LH. These hormones are z-scored within person, and we report the cycle-averages spearman correlation and held-out coefficient of determination under subject-level GroupKFold.

4.4 Statistics and reproducibility

All evaluations use user-level GroupKFold splits so that no participant appears in more than one fold. Uncertainty is estimated by non-parametric user-cluster bootstrap resampling, with two-tailed tests at $\alpha = 0.05$. Period and ovulation detection are evaluated on held-out test users. Reproductive-state transfer is evaluated by user-level GroupKFold cross-validation on encoder-unseen users, because pretraining uses only cycling users. The effect of the phase objective is measured as the paired difference $\Delta\mathrm{AUC} = (\mathrm{recon+phase}) - (\mathrm{recon})$, tested per task by user-cluster bootstrap, with the nine task-level P values corrected using the Benjamini–Hochberg procedure. Hormone recovery is evaluated by held-out R^2 under subject-level GroupKFold on mcPHASES (42 subjects, 161 cycles); phase-versus-reconstruction differences (ΔR^2) are tested by subject-level bootstrap resampling (2,000 iterations).

Declarations

Funding

This research was conducted as part of the authors' employment at Terra Enabling Developers. No external funding was received.

Competing interests

All authors are employees of Terra Enabling Developers, the wearable-data infrastructure (Terra API) provider whose data were used in this study.

Ethics approval and consent to participate

This study is a secondary analysis of fully de-identified, aggregated wearable data obtained through Terra API. The underlying data is collected from end users who provide informed consent to data collection and to its use for research through the terms

and conditions of the third-party applications to which they connect their devices. The study did not require additional institutional review board approval. This was confirmed using the Medical Research Council and Health Research Authority which returned that NHS Research Ethics Committee review is not required for this study. The study recruits no participants through the NHS or adult social care, collects no tissue or prospective information, and processes no confidential patient information.

Data availability

The wearable data analysed in this study was provided by Terra API and is not publicly available due to user-privacy and data protection restrictions. Aggregated data statistics are available from the corresponding author on reasonable request and with the permission of Terra Enabling Developers. The external mcPHASES reference dataset is publicly available on PhysioNet (https://physionet.org/content/mcphases/1.0.0/).

Code availability

The code used to train and evaluate the models is proprietary to Terra Enabling Developers and is available from the corresponding author on reasonable request.

Author contributions

R.M.D. conceived and designed the study, implemented the models and experiments, performed the analyses, and wrote the manuscript. F.A. contributed to the study design and revised the manuscript. C.C., A.B., and K.E. read and approved the final manuscript.

References

[1] Valera, H., Chen, A. & Grive, K. J. The hypothalamic-pituitary-ovarian axis, ovarian disorders, and brain aging. *Endocrinology* **166** (2025).

[2] Anckaert, E. *et al.* Extensive monitoring of the natural menstrual cycle using the serum biomarkers estradiol, luteinizing hormone and progesterone. *Practical Laboratory Medicine* **25**, e00211 (2021).

[3] Ku, C. W. *et al.* Gestational age-specific normative values and determinants of serum progesterone through the first trimester of pregnancy. *Scientific Reports* **11**, 4161 (2021).

[4] Baker, F. C., Siboza, F. & Fuller, A. Temperature regulation in women: Effects of the menstrual cycle. *Temperature* **7**, 226–262 (2020).

[5] Jasinski, S. R., Presby, D. M., Grosicki, G. J., Capodilupo, E. R. & Lee, J. A novel method for quantifying fluctuations in wearable-derived daily cardiovascular parameters across the menstrual cycle. *npj Digital Medicine* **7**, 373 (2024).

[6] Slatkovska, L., Jensen, D., Davies, G. A. L. & Wolfe, L. A. Phasic menstrual cycle effects on the control of breathing in healthy women. *Respiratory Physiology & Neurobiology* **154**, 379–388 (2006).

[7] Goodale, B. M. *et al.* Wearable sensors reveal menses-driven changes in physiology and enable prediction of the fertile window: Observational study. *Journal of Medical Internet Research* **21**, e13404 (2019).

[8] Maijala, A., Kinnunen, H., Koskimäki, H., Jämsä, T. & Kangas, M. Nocturnal finger skin temperature in menstrual cycle tracking: ambulatory pilot study using a wearable oura ring. *BMC Women's Health* **19**, 150 (2019).

[9] Gombert-Labedens, M. *et al.* Using wearable skin temperature data to advance tracking and characterization of the menstrual cycle in a real-world setting. *Journal of Biological Rhythms* **39**, 331–350 (2024).

[10] Shi, Y. *et al.* The diagnostic accuracy of wearable digital technology in detecting fertility window and menstrual cycles: a systematic review and bayesian network meta-analysis. *npj Digital Medicine* (2026).

[11] Arévalo, M., Jennings, V. & Sinai, I. Efficacy of a new method of family planning: the standard days method. *Contraception* **65**, 333–338 (2002).

[12] Thigpen, N., Patel, S. & Zhang, X. Oura ring as a tool for ovulation detection: Validation analysis. *Journal of Medical Internet Research* **27**, e60667 (2025).

[13] Luo, C. *et al.* Prediction of the fertile window and menstruation with a wearable device via machine-learning algorithms. *Reproductive BioMedicine Online* **51**, 104795 (2025).

[14] Kilungeja, G., Graham, K., Liu, X. & Nasseri, M. Machine learning-based menstrual phase identification using wearable device data. *npj Women's Health* **3**, 29 (2025).

[15] Adaimi, R. *et al.* Temporal trajectories in sleep, temperature trends, cardiorespiratory, and activity metrics measured via oura ring during pregnancy: Large-scale observational analysis. *JMIR mHealth and uHealth* **13**, e80213 (2025).

[16] de Jager, E. *et al.* Wearable-derived heart rate variability across the menstrual cycle, hormonal contraceptive use, and reproductive life stages in females: A living systematic review. *Sports Medicine* **56**, 1153–1173 (2026).

[17] Johnson, S. C., O'Day, J., Kraus, E., Delp, S. L. & Hicks, J. L. Decoding menstrual health across the lifespan: a scoping review of digital health tools in research. *npj Women's Health* (2026).

[18] Abbaspourazad, S. *et al.* Large-scale training of foundation models for wearable biosignals. *International Conference on Learning Representations (ICLR)* (2024).

[19] Narayanswamy, G. *et al.* Scaling wearable foundation models. *International Conference on Learning Representations (ICLR)* (2025).

[20] Yuan, H. *et al.* Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine* **7**, 91 (2024).

[21] Yang, Y. *et al.* Simper: Simple self-supervised learning of periodic targets. *International Conference on Learning Representations (ICLR)* (2023).

[22] Destro, M. & Gygli, M. Cyclecl: Self-supervised learning for periodic videos. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 2849–2858 (2024).

[23] Li, K. *et al.* Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile-health data. *npj Digital Medicine* **3**, 79 (2020).

[24] Lin, G., Li, J. Y., Kalani, K., Truong, K. N. & Mariakakis, A. A longitudinal dataset of physiological, hormonal, metabolic, and self-reported menstrual health data. *Scientific Data* **13**, 411 (2026).

[25] Lyzwinski, L., Elgendi, M. & Menon, C. Innovative approaches to menstruation and fertility tracking using wearable reproductive health technology: Systematic review. *Journal of Medical Internet Research* **26**, e45139 (2024).

[26] Grant, A. D., Newman, M. & Kriegsfeld, L. J. Ultradian rhythms in heart rate variability and distal body temperature anticipate onset of the luteinizing hormone surge. *Scientific Reports* **10**, 20378 (2020).

[27] Bull, J. R. *et al.* Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *npj Digital Medicine* **2**, 83 (2019).

[28] Usala, S. J., Vineyard, D. D., Kastis, M., Trindade, A. A. & Gill, H. S. Comparison of day-specific serum LH, estradiol, and progesterone with Mira monitor urinary LH, estrone-3-glucuronide, and pregnanediol-3-glucuronide levels in ovulatory cycles. *Medicina* **60**, 1207 (2024).

[29] Lea, C., Vidal, R., Reiter, A. & Hager, G. D. Temporal convolutional networks: A unified approach to action segmentation. *European Conference on Computer Vision (ECCV) Workshops* (2016).

[30] Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint* (2018).