TRAINING AND SCALING COMPUTE-OPTIMAL PHYSIOLOGICAL WAVEFORM FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the scaling laws and compute-optimal training of physiological waveform foundation models (FMs). We train `Aether`, a family of over one hundred FMs ranging from 20M to 2.1B parameters, on up to 36.3M hours of physiological waveforms. We construct eight clinical prediction tasks from MIMIC-III and evaluate the FMs through linear probing. The 720M FM outperforms all existing baseline FMs across all eight tasks. A scaling law of model size, pretraining hours, and labeled patients predicts downstream ranking error, *i.e.* $1 - \mathrm{AUROC}$, effectively with 0.5% prediction MAE at held-out resource scales and 0.9% MAE when extrapolating to 2.1B parameters. We present three findings: (1) Compute-optimal training scales both FM size and pretraining hours. Under the fitted law, a $10.0\times$ increase in compute FLOPs scales model size by $1.2\times$ and pretraining hours by $8.2\times$. (2) Larger FMs use waveform data more efficiently, and greater pretraining exposure increases the benefit of model scaling. Starting from 25M parameters and 4.8M pretraining hours, doubling FM size reduces the predicted hours needed for the same performance by 51.8%. (3) Pretraining and clinical supervision reinforce each other: more labeled patients increase the return to pretraining, while larger FMs and longer pretraining reduce labeling requirements. For the example of the 720M FM, extending pretraining from 120K to 36.3M hours reduces the predicted patient requirement by 61% at a target ranking error. These findings provide a quantitative training recipe and a promising and durable scaling path for physiological waveform modeling and downstream clinical prediction.

1 INTRODUCTION

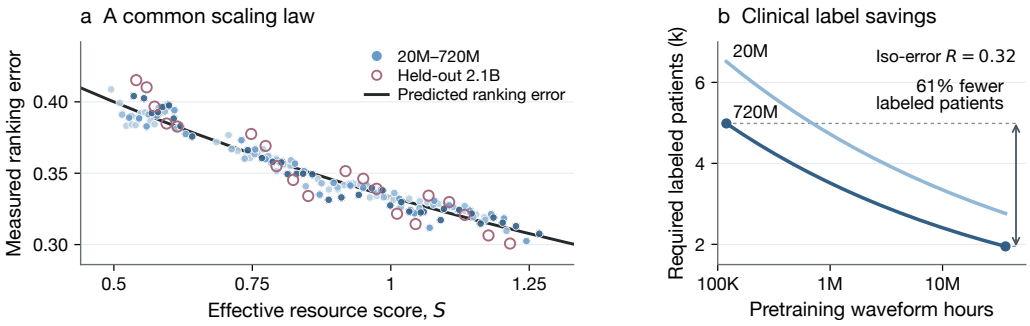

Figure 1: Scaling clinical prediction and label efficiency. (a) Measured eight-task mean ranking error against $S = P^g/(AN^{-a} + BD^{-b})$. (b) Predicted labeled patient requirements at $\mathcal{R} = 0.32$. Extending 720M pretraining from 120K to 36.3M hours reduces required patients by 61%.

Scaling laws for large language models describe how held-out loss follows power-law relationships with model size and training data (Kaplan et al., 2020), and compute-optimal training scales model size and token count in approximately equal proportion as compute grows (Hoffmann et al., 2022). The practical value is that a large training run can be configured in advance, allocating its compute before committing to the most expensive experiments. Physiological waveforms bring this resource-allocation question to continuous measurements and clinical supervision. Existing work offers two complementary starting points. Foundation models demonstrate transferable representations for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

electrocardiography, photoplethysmography, electroencephalography, sleep polysomnography, and multimodal physiological signals (McKeen et al., 2024; Pillai et al., 2025; Wang et al., 2025; Thapa et al., 2026; Guo et al., 2025). Scaling studies characterize resource-performance relationships in structured clinical event records (Zhang et al., 2025), ECG (Li et al., 2026; Al-Masud & Strodthoff, 2026), and wearable and intracranial recordings (Narayanswamy et al., 2024; Han et al., 2026). These advances motivate a scaling law of FM size, waveform exposure, and labeled patients.

We ask: *Given a pretraining compute budget and a labeled-patient budget, how should model size and waveform exposure scale to achieve the best downstream clinical performance?* To study this question, we train `Aether`, a family of over one hundred foundation models (FMs) ranging from 20M to 2.1B parameters, on up to 36.3M cumulative hours of physiological waveforms. To efficiently model long waveform contexts, we use a bidirectional hybrid backbone based on Gated DeltaNet (GDN) (Yang et al., 2025), arranging GDN and full-attention blocks in a repeating 3:1 pattern as in Qwen Team (2026). We construct eight clinical prediction tasks from MIMIC-III waveforms and clinical records (Moody et al., 2020; Johnson et al., 2016). Each pretrained FM is evaluated with a linear probe on frozen representations, following established protocols in visual representation learning (Chen et al., 2020; He et al., 2020; 2022) and ECG representation learning (Kiyasseh et al., 2021; Diamant et al., 2022). Overall, we evaluate over one thousand combinations of model size, pretraining hours, labeled-patient counts, and clinical task, studying how resources invested both before and after pretraining translate into downstream physiological performance. Our `Aether`-720M outperforms the existing baseline FMs across all eight tasks.

We utilize this resource study to derive a quantitative training recipe. A scaling law relates the eight-task mean ranking error to FM size N, pretraining waveform hours D, and labeled patients per task P (Figure 1a). It predicts held-out resource configurations with 0.5% MAE and generalizes to larger FMs and longer pretraining runs. We use this law to derive compute-optimal model–data allocations and quantify the waveform and labeling requirements for a target clinical performance.

We present three findings: (1) Compute-optimal training scales FM size and pretraining hours together, with hours growing faster than model size. Under the fitted law, a $10.0\times$ increase in compute scales FM size by $1.2\times$ and pretraining hours by $8.2\times$. (2) Larger FMs are more data-efficient, and longer pretraining amplifies model-scaling gains. At fixed P and ranking error, doubling a 25M FM from a 4.8M-hour reference reduces predicted pretraining hours by 51.8%.

(3) More labeled patients amplify pretraining gains, while scaling pretraining reduces labeling needs. For example, the law predicts 27.2% fewer labeled patients with larger FMs and 61% fewer with longer pretraining at fixed target performance (Figure 1b; Section 5). In practice, clinical labels add expert curation to the access requirements of raw waveforms, making both harder to expand than compute (Krishnan et al., 2022; Rieke et al., 2020). This suggests a promising and durable scaling path: scale pretraining on available waveforms to improve clinical prediction with fewer labels.

Our contributions and findings are summarized as follows: (1) We establish a predictive scaling law for clinical performance and derive compute-optimal allocations for scaling FM size and pretraining hours with compute. (2) We quantify the data-efficiency benefits of model scaling: larger FMs reach the same predicted performance with fewer waveform hours, while greater pretraining exposure increases the benefit of scaling model size. (3) We characterize how pretraining and clinical supervision reinforce each other: more labeled patients increase the return to pretraining, while larger FMs and longer pretraining reduce the predicted labeling requirements for a target clinical performance.

2 Related work

Scaling and downstream transfer. Empirical studies characterize predictable relationships between dataset size, model capacity, and generalization error across learning domains (Hestness et al., 2017; Rosenfeld et al., 2020). Language-model scaling laws connect model size, data, and compute (Kaplan et al., 2020), with compute-optimal training jointly allocating parameters and training tokens (Hoffmann et al., 2022). Related analyses extend to vision transformers (Zhai et al., 2022) and neural machine translation (Ghorbani et al., 2022). Transfer scaling expresses pretraining benefits through downstream data requirements (Hernandez et al., 2021), while data-constrained scaling distinguishes new tokens from repeated exposure (Muennighoff et al., 2023). Studies of extrapolation and changes in scaling regimes further examine which functional forms predict performance beyond

the fitted range (Alabdulmohsin et al., 2022; Caballero et al., 2023). These works motivate our scaling model and its evaluation at unseen model scales and compute allocations for clinical prediction.

Scaling in physiological and clinical data. Li et al. (2026) jointly examine ECG architecture, pretraining paradigm, and model–data scaling, while Al-Masud & Strodthoff (2026) compare self-supervised objectives as ECG pretraining data grows. Wearable scaling studies distinguish subject count from recording duration (Narayanswamy et al., 2024); DIVER-1 studies model size, unique data, subject count, and training duration for iEEG (Han et al., 2026). Complementary analyses consider modality diversity and model capacity in nocturnal biosignals (Yuan et al., 2026), data per subject in brain-to-image decoding (Banville et al., 2025), and compute-optimal scaling in EHR sequences (Zhang et al., 2025). Cohort-composition experiments also show that ECG transfer depends on the distribution of pretraining patients (Khattak et al., 2026). Building on these precedents, we study a joint model–exposure–supervision law for clinical waveform prediction, with labeled patients as an explicit resource coordinate. We compare patient- and window-based label acquisition under a common model ladder and evaluation protocol to test whether that coordinate transfers across acquisition designs while keeping task definitions, validation cohorts, and selection criteria fixed.

Physiological waveform foundation models. Early ECG pretraining methods, including CLOCS (Kiyasseh et al., 2021) and PCLR (Diamant et al., 2022), show that unlabeled cardiac recordings can support label-efficient downstream learning. ECG-FM combines contrastive and generative pretraining for electrocardiogram analysis (McKeen et al., 2024). PaPaGei uses waveform morphology to learn transferable photoplethysmography representations (Pillai et al., 2025). For EEG, BENDR learns transferable features through contrastive pretraining (Kostas et al., 2021). LaBraM learns by predicting masked neural tokens (Jiang et al., 2024), while CBraMod uses masked reconstruction with separate spatial and temporal attention (Wang et al., 2025). BIOT supports learning across biosignal datasets with heterogeneous channels, sample lengths, and missing values (Yang et al., 2023). SleepFM learns multimodal polysomnography representations for disease prediction (Thapa et al., 2026), and QualityFM addresses variable ECG and PPG signal quality through self-distillation (Guo et al., 2025). These studies establish the utility of pretrained waveform representations across modalities and clinical tasks. Our focus is to quantify how resources allocated to pretraining and downstream supervision translate into clinical performance under a common linear-probe protocol across all model sizes, pretraining exposures, and labeled-patient budgets.

3 Experimental Design

We study scaling by varying FM size, pretraining hours, and labeled-patient coverage within a common training and evaluation framework. Section 3.1 presents the `Aether` architecture and scaling ladder. Section 3.2 describes the waveform data and pretraining recipe. Section 3.3 introduces the eight clinical tasks, downstream linear-probing protocol, and evaluation metrics.

3.1 Model Architecture

Figure 2 summarizes the training and adaptation of `Aether`, which tokenizes one hour of physiological waveforms into a sequence of continuous representations. Inputs are sampled at 125 Hz and 62 channels. We normalize each channel within its input window and jointly project patches of 25 samples across channels, giving 18,000 temporal tokens per hour. To model this long context efficiently, the FM interleaves bidirectional Gated DeltaNet blocks (Yang et al., 2025) and full-attention blocks in a repeating 3:1 pattern (Qwen Team, 2026). Full-attention blocks use no positional encoding (NoPE) (Kazemnejad et al., 2023). Each block uses RMSNorm (Zhang & Sennrich, 2019) and a SwiGLU feed-forward layer (Shazeer, 2020).

Table 1: `Aether` scaling ladder. FFN: Dimension of FFN; Q/KV: full-attention head counts; GDN: recurrent heads.

Size	Width	Depth	FFN	Q/KV	GDN
20M	640	5	2,048	5/1	1
60M	896	7	2,816	7/1	2
180M	1,280	10	4,096	10/2	3
720M	1,920	15	6,144	15/3	4
2.1B	2,816	22	8,960	22/2	6

The scaling ladder derives hidden width, FFN width, and head counts from depth, as shown in Table 1, with a same 3:1 hybrid block pattern. Full-attention and GDN heads have dimension 128, and the GDN convolution width is 4. We use the four smaller FMs to establish the scaling laws, while `Aether`-2.1B provides the held-out model-size evaluation. Scaling fits use the exact FM parameter count N, excluding the reconstruction MLP head.

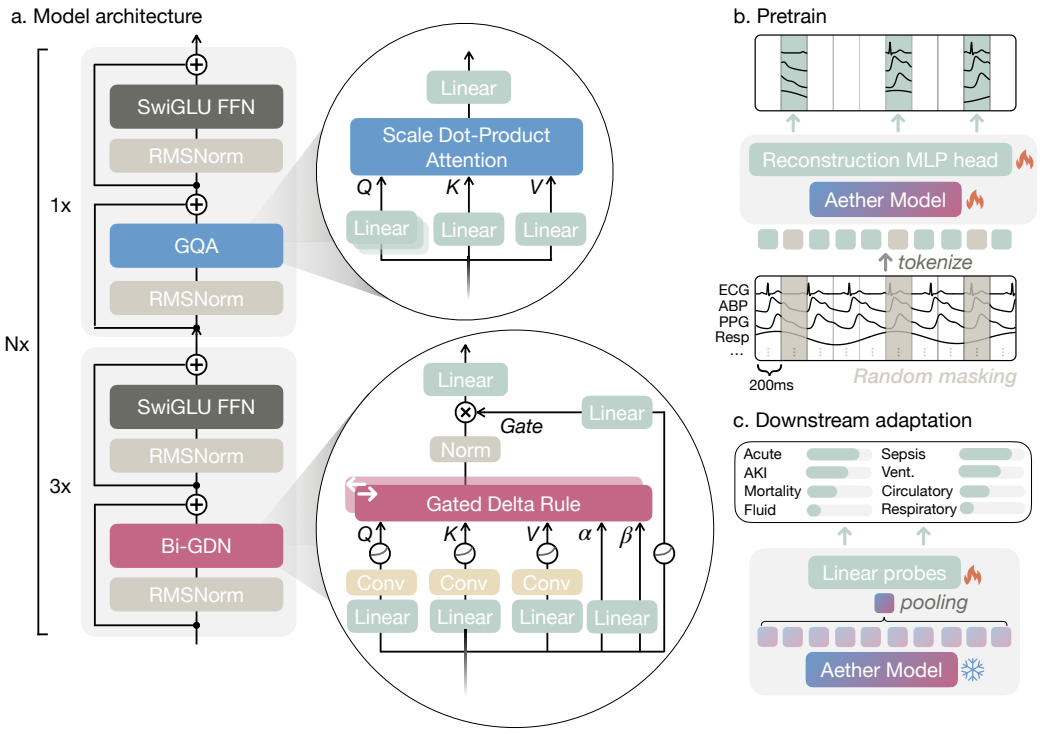

Figure 2: Overview of `Aether` architecture, pretraining, and downstream adaptation. (a) Bidirectional Gated DeltaNet (Bi-GDN) and grouped-query attention (GQA) blocks form a repeating 3:1 hybrid pattern. (b) Multichannel waveforms are divided into 200 ms patches, and the FM learns to reconstruct randomly masked patches. (c) The pretrained FM is frozen, and its mean-pooled representations support linear probing for downstream clinical prediction.

3.2 Pretraining

Data. The pretraining corpus comprises 398,204 hours of inputs from MIMIC-III waveform recordings (Moody et al., 2020; Johnson et al., 2016). We pretrain by random masked reconstruction (He et al., 2022), where each temporal patch is independently masked with probability 0.30, replacing its projected token with a learned mask embedding. A reconstruction head predicts the normalized waveform, and the loss averages squared error over observed values in masked patches:

$$\mathcal{L}_{\mathrm{PT}} = \frac{\sum_j m_j o_j (\hat{x}_j - x_j)^2}{\sum_j m_j o_j}, \tag{1}$$

where j indexes scalar samples across the minibatch, o_j indicates an observed value, and m_j indicates that its temporal patch is masked. We use AdamW (Loshchilov & Hutter, 2019) with a global batch of 512, learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.05, and gradient-norm clipping at 1.0. We follow the warmup-stable-decay (WSD) schedule (Hu et al., 2024), with 100 warmup steps, a stable learning rate of 10^{-4}, and 100 cosine-decay steps to 10^{-5}. Every pretraining run has its own terminal decay to ensure convergence.

Training. We measure pretraining exposure D in waveform-hours and evaluate 119K, 239K, 398K, 796K, 1.59M, and 2.39M hours. To extend pretraining hours beyond the available data, we repeat the pretraining waveforms when needed, with E denoting the pretraining repeat ratio. Compute budgets include forward and backward FLOPs accumulated over pretraining updates, following Hoffmann et al. (2022). We also built a JAX-based (Bradbury et al., 2018) distributed training infra, named FoXus, for our internal heterogeneous clusters spanning AWS Trainium2, Google TPU v5p, and NVIDIA B200. We built FoXus from scratch to give us full end-to-end control over the training stack, including compiler, kernels, parallelism and scheduling [1].

[1]We achieve MFU at 14% for Trainium2, 40% for v5p, and 34% for B200.

3.3 Downstream linear probing

Data. We construct eight held-out clinical prediction tasks from MIMIC-III waveforms and clinical records (Moody et al., 2020; Johnson et al., 2016) to evaluate the FMs' performance: acute deterioration, circulatory deterioration, dynamic mortality, respiratory escalation, sepsis onset, fluid response, KDIGO stage 2/3 acute kidney injury, and ventilator readiness within 24 hours. Acute deterioration has three binary targets, fluid response and acute kidney injury have two each, and the other tasks have one. Patients in all downstream tasks are held out from the pretraining cohort. Table 2 summarizes the training, validation and test cohorts. Details of task definitions and construction are in Appendix A.

Table 2: Overview of downstream task datasets.

	Train		Validation		Test	
Task	Hours	Patients	Hours	Patients	Hours	Patients
Acute	112,727	5,088	13,852	623	14,683	637
Circulatory	316,881	4,785	37,109	595	463,197	877
Mortality	593,154	7,112	72,226	885	77,809	887
Respiratory	122,597	4,334	14,937	545	16,356	542
Sepsis	152,486	3,892	22,089	480	214,884	726
Fluid response	4,742	1,491	552	176	692	205
KDIGO 2/3 AKI	124,623	3,800	15,315	461	14,542	481
Vent. readiness	9,276	2,069	1,112	240	1,215	245

Training. For linear probing, we sample {10%, 30%, 60%, 100%} of the full training patients, including all inputs from each selected patient. We record the labeled-patient count P as the patient-scaling coordinate. Together, the FM sizes, pretraining hours, patient counts, and tasks yield about one thousand combinations. A companion analysis samples the same fractions of waveform windows.

Following standard linear probing protocols (Chen et al., 2020; He et al., 2020; 2022), we freeze the FM, mean-pool its representations, and fit an L_2-regularized logistic probe for each target. Features are standardized using the selected training subset. Five candidates use fixed regularization strengths $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$. We select one complete multi-target probe per condition by validation patient-balanced macro AUROC and use it for every reported metric.

Metrics. Our primary outcome is ranking error, $\mathcal{R} = 1 - \bar{A}$ (Agarwal, 2014), where $\bar{A}$ is patient-balanced AUROC averaged across the task's targets. Each patient's observed inputs receive equal total weight per target (Williamson et al., 2003; Kahan et al., 2023). Changes in ranking error are absolute shifts on the percentage scale. We estimate uncertainty with 10,000 paired patient-bootstrap draws, retaining all inputs from each sampled patient (Obuchowski, 1997; Li & Zhou, 2008).

4 Scaling Laws and Evaluation

Table 3 summarizes the scaling-law notation. This section connects the clinical performance of `Aether` to model capacity, pretraining exposure, and downstream supervision. Section 4.1 establishes the effectiveness of our pretraining recipe through comparisons with random initialization and existing FMs. Section 4.2 fits a scaling law and evaluates its predictions at held-out resource levels. Section 4.3 evaluates predictions at 2.1B and longer pretraining durations.

Table 3: Summary of notations.

Symbol	Description
$\bar{A}$	Patient-level macro AUROC
$\mathcal{R}$	Ranking error, $1 - \bar{A}$
$\widehat{\mathcal{R}}$	Predicted ranking error
N	FM parameter count
D	Pretraining waveform hours
P	Labeled patients per task
C	Pretraining compute (FLOPs)
q	Waveform tokens per hour

4.1 Comparison with Existing FMs

We first evaluate whether our pretraining recipe yields useful representations for clinical prediction and how `Aether` compares with existing FMs. This comparison establishes the clinical benefit of our training recipe before examining its scaling behavior.

Baseline waveform FMs. We compare with thirteen pretrained models from six families: TimesFM 2/2.5/3 (Das et al., 2024) and Chronos-Bolt (Ansari et al., 2024), developed for time-series forecasting; MOMENT (Goswami et al., 2024), which learns general-purpose time-series representations through masked reconstruction; SOTER (Chen et al., 2026), a generative FM for wearable physiological signals; and NormWear 1/2 (Luo et al., 2026b;a), which learn wearable-signal representations and

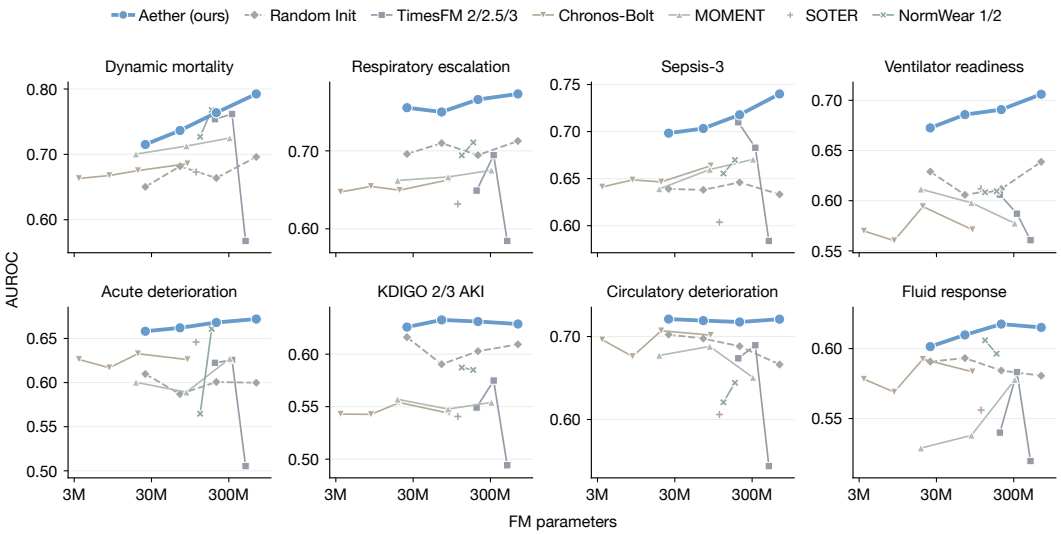

Figure 3: Comparison with existing waveform FMs across eight clinical prediction tasks. Each panel shows macro AUROC against FM size using 100% of labeled patients. To ensure a fair comparison, we report each model's best linear-probe AUROC across the same regularization sweep, $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$. Higher AUROC corresponds to lower ranking error.

physiological dynamics. We also include randomly initialized `Aether` FMs at each size to measure the contribution of our pretraining recipe. All FMs are frozen and evaluated with linear probes using the full training cohorts and the same patient-balanced validation metric.

Our pretraining improves clinical prediction. As shown in Figure 3, pretrained `Aether` outperforms its randomly initialized counterpart across all eight tasks and four model sizes. At 720M, our pretraining recipe improves eight-task mean AUROC by 6.4%, including gains of 9.6% for mortality and 10.7% for sepsis. These gains demonstrate the clinical value of our pretraining recipe.

`Aether` outperforms the evaluated baselines. The 720M FM achieves higher AUROC than every public baseline on each of the eight tasks in Figure 3. Relative to the strongest baseline within each task, ranking error decreases by 0.9%–9.4%, averaging 3.6% across tasks. The reductions reach 4.1% for AKI, 6.2% for respiratory escalation, and 9.4% for ventilator readiness. These results establish the effectiveness of the `Aether` training recipe across diverse clinical endpoints. We next quantify how its performance changes with FM size, pretraining hours, and labeled patients.

4.2 Fitting Scaling Laws

We fit a scaling law to the ranking error across the eight downstream tasks (Kaplan et al., 2020; Hoffmann et al., 2022):

$$\widehat{\mathcal{R}} = \left[2 + \frac{P^g}{AN^{-a} + BD^{-b}}\right]^{-1}, \tag{2}$$

where N is the FM parameter count, D is pretraining waveform hours, and P is the number of labeled patients per task. Nonlinear least squares over the resource grid gives $A = 6931.568$, $B = 19.421$, $a = 0.513$, $b = 0.050$, and $g = 0.296$. Appendix B develops the expected scaling terms and describes the fitting procedure. Figure 4(a) shows that configurations with different N, D, and P follow a common response curve when expressed through the resource score $P^g/(AN^{-a} + BD^{-b})$. Positive exponents give consistent improvements with each resource. Across the fitting grid, ranking error responds most strongly to proportional increases in labeled patients, identifying supervision as the leading downstream scaling factor after pretraining.

Figure 4(b,c) presents the effects of pretraining hours and labeled patients respectively. The fitted ranking error curves consistently decrease as either resource grows across all FM sizes, with larger FMs reaching lower ranking error at the same D and P. The law also quantifies their greater data efficiency: at fixed P and target ranking error, $AN^{-a} + BD^{-b}$ remains constant. Starting from

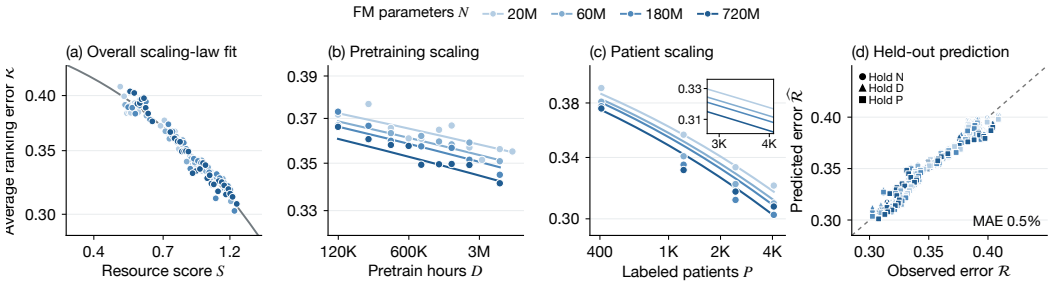

Figure 4: Scaling of downstream average ranking error $\mathcal{R}$ across eight tasks. (a) Fitted response to the resource score $S = P^g/(AN^{-a} + BD^{-b})$. (b) Pretraining scaling at $P = 1{,}222$. (c) Patient scaling at 4.8M pretraining hours. (d) Predictions at held-out resource levels: we hold out all configurations at one value of N, D, or P, fit on the rest, and predict the held-out configurations. MAE averages absolute prediction errors equally across the three resource axes.

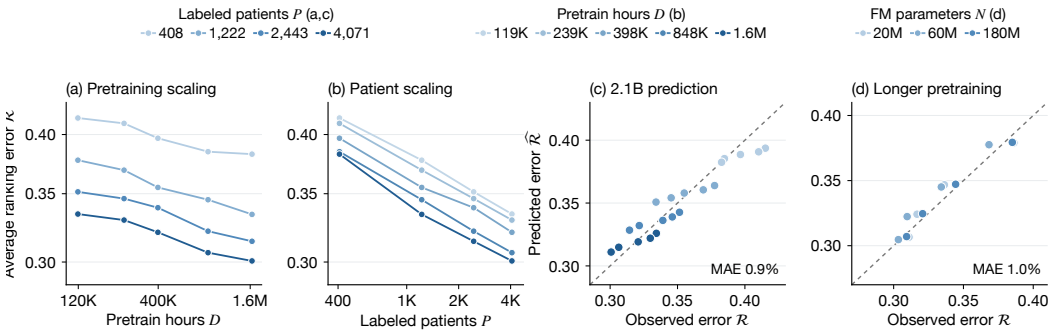

Figure 5: Scaling and prediction at larger resources. (a) Pretraining scaling at 2.1B for all four values of P. (b) Patient scaling at 2.1B for all five values of D. (c) Predictions for all 20 configurations at 2.1B. (d) Predictions for 20M, 60M, and 180M after 36.3M, 17.0M, and 5.9M pretraining waveform hours, respectively. Ranking errors and absolute patient counts are averaged across the eight tasks. Solid lines connect measurements in (a,b); dashed diagonals in (c,d) mark $\widehat{\mathcal{R}} = \mathcal{R}$.

25M parameters and 4.8M pretraining hours, the fit predicts that doubling FM size reduces the hours needed for the same error by 51.8%. These trends demonstrate consistent improvements from scaling FM size, pretraining waveform hours, and labeled patients.

We evaluate predictions at unseen resource levels by leaving out all configurations at one value of N, D, or P in turn and fitting on the remaining configurations. Figure 4(d) shows close agreement between predicted and observed ranking error, with a development MAE of 0.5%. This reduces MAE by 50.1% relative to a baseline that uses the measured error of the nearest configuration available during fitting, with distance measured in $(\log N, \log D, \log P)$ space. These results demonstrate accurate prediction of ranking error at held-out resource levels.

4.3 Evaluation on Predicting Larger Scales

To further evaluate the generalizability of our scaling law, we apply Equation 2, fitted on 20M–720M FMs, to `Aether`-2.1B and to longer pretraining runs at 20M, 60M, and 180M. Figure 5 summarizes all 20 configurations at 2.1B and all 12 configurations at the longer training durations, each evaluated across the same eight clinical tasks.

The gains from waveform pretraining and labeled-patient supervision extend to the 2.1B scale. Increasing pretraining from 119K to 1.6M waveform hours reduces average ranking error by 3.3%–4.3% across all four labeled-patient counts (Figure 5a). Increasing P from 408 to 4,071 reduces average ranking error by 7.6%–8.2% across all five pretraining settings (Figure 5b). These gains demonstrate consistent improvements from scaling our training recipe at 2.1B.

The law predicts performance at 2.1B with an MAE of 0.9% across all 20 configurations (Figure 5c). It also predicts the performance of 20M, 60M, and 180M FMs after 36.3M, 17.0M, and 5.9M pretraining

waveform hours, respectively, with an overall MAE of 1.0% across the 12 configurations (Figure 5d). On these longer runs, the law reduces MAE by 18.1% relative to nearest-configuration prediction. These results demonstrate the scaling law's ability to generalize through accurate extrapolation to larger FMs and longer pretraining durations.

5 Compute-Optimal Training of Physiological FMs

Q1: How should model size and waveform data scale with compute? We express Equation 2 under the training-compute approximation $C \simeq 6qND$ (Kaplan et al., 2020; Hoffmann et al., 2022), where $q = 18{,}000$ waveform tokens per hour for `Aether`. Substituting $D = C/(6qN)$ gives

$$\widehat{\mathcal{R}}(C, N, P) = \left[2 + \frac{P^g}{AN^{-a} + B\left(C/(6qN)\right)^{-b}}\right]^{-1}. \tag{3}$$

At fixed P, minimizing ranking error is equivalent to minimizing $AN^{-a} + BD^{-b}$. The interior compute-optimal allocation is

$$N^\star(C) = \left(\frac{aA}{bB}\right)^{1/(a+b)} \left(\frac{C}{6q}\right)^{b/(a+b)} \propto C^{0.088}, \tag{4}$$

$$D^\star(C) = \left(\frac{bB}{aA}\right)^{1/(a+b)} \left(\frac{C}{6q}\right)^{a/(a+b)} \propto C^{0.912}. \tag{5}$$

Figure 6 shows this allocation across all four labeled-patient cohorts. For interior solutions, a $10\times$ increase in compute scales FM size by $1.2\times$ and pretraining hours by $8.2\times$. These predictions provide two ways to use additional compute. ❶ Scale FM size and pretraining waveform hours together, with waveform data growing faster than model size as compute increases. ❷ When waveform data are limited, increase FM size to improve data efficiency and clinical performance. Unlike web-scale language corpora, real-world clinical waveform datasets are difficult to expand continuously because of patient privacy, legal restrictions, and institutional access constraints (Johnson et al., 2016; Rieke et al., 2020). At fixed D and P, larger FMs use the available waveforms more efficiently and continue to improve clinical performance.

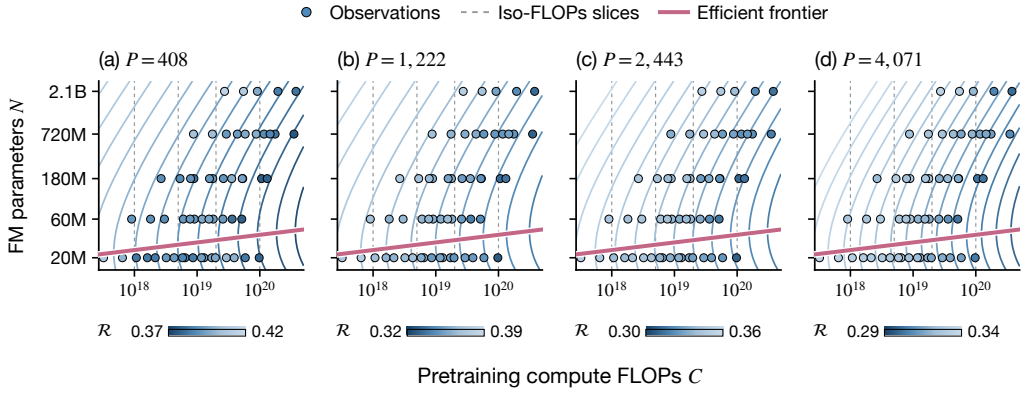

Figure 6: Iso-error contours under a FLOPs constraint, from Equation 3. (a–d) $P = 408$, 1,222, 2,443, and 4,071, respectively. Dashed lines mark iso-FLOPs slices; red curves show the efficient frontier over the observed FM-size range. Each panel shows all 57 configurations. Ranking errors and patient counts are eight-task means.

Q2: How much waveform data can larger FMs save? Larger FMs use pretraining waveform data more efficiently. At fixed P and target ranking error, Equation 2 gives $AN^{-a} + BD^{-b} = \text{constant}$. The iso-error contours in Figure 7 trace model–data tradeoffs across all four labeled-patient cohorts: increasing N reduces the waveform hours D needed for the same predicted performance. The saving varies with the starting N and D.

We draw the following implications: ❶ More waveform data increases the benefit of model scaling. For example, at $P = 4{,}071$, scaling from 20M to 720M yields a relative reduction in predicted

ranking error of 4.0% at 36.3M pretraining waveform hours, compared with 2.6% at 120K hours. ❷ Scaling pretraining data yields larger relative gains with more labeled patients. For a 720M FM, increasing D from 120K to 36.3M hours yields a relative reduction in predicted ranking error of 10.0% at $P = 4{,}071$, compared with 6.3% at $P = 408$. These complementary gains support scaling FM capacity together with pretraining waveform data.

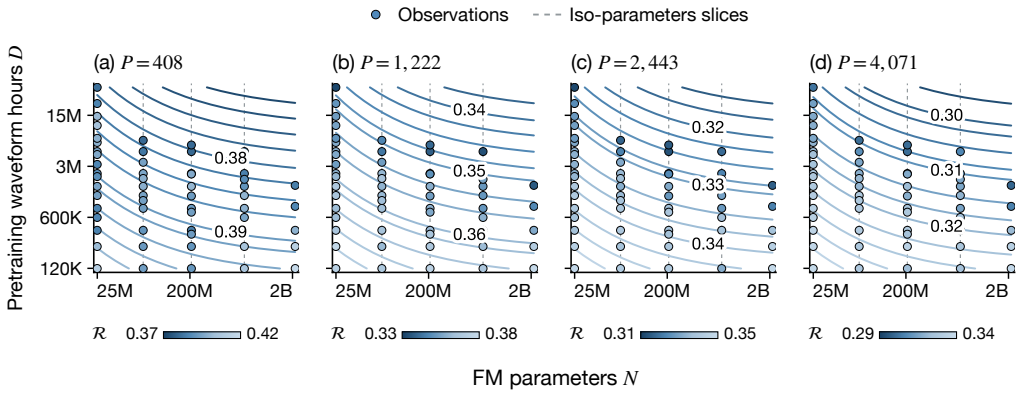

Figure 7: Iso-error contours for FM size and pretraining waveform hours. Each curve holds the predicted eight-task average ranking error fixed under Equation 2. (a–d) $P = 408$, 1,222, 2,443, and 4,071, respectively. Points show measurements, with color indicating observed ranking error. Dashed lines mark iso-parameters slices. Patient counts are averaged across the eight tasks. Moving right along a curve reduces waveform hours at fixed ranking error $\mathcal{R}$.

Q3: How can scaling pretraining reduce labeling requirements? Scaling pretraining reduces the labeled-patient count needed to attain a given clinical performance. For a target ranking error $\mathcal{R}_{\mathrm{t}}$, solving Equation 2 for P gives

$$P = \left[\left(AN^{-a} + BD^{-b}\right)\left(\mathcal{R}_{\mathrm{t}}^{-1} - 2\right)\right]^{1/g}. \tag{6}$$

Figure 8 translates pretraining resources into labeling requirements at two target errors across all five FM sizes. Increasing either N or D consistently reduces the required P.

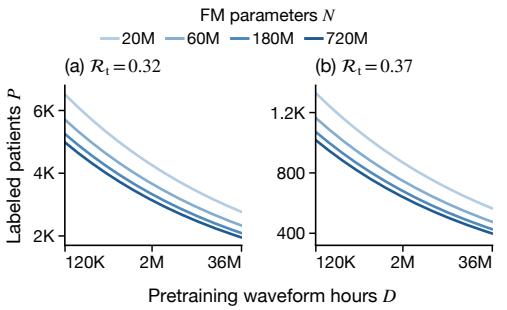

Figure 8: Predicted labeling requirements averaged across eight tasks. (a) $\mathcal{R}_{\mathrm{t}} = 0.32$. (b) $\mathcal{R}_{\mathrm{t}} = 0.37$.

We draw the following implications: ❶ Model scaling improves label efficiency. At $\mathcal{R}_{\mathrm{t}} = 0.32$ and 4.8M pretraining waveform hours, scaling from 20M to 720M reduces the predicted patient requirement by 27.2%, from approximately 3,733 to 2,717. ❷ Longer pretraining further reduces annotation requirements. For the 720M FM at the same target error, increasing D from 120K to 36.3M hours reduces the requirement by 61%, from approximately 4,986 to 1,949.

6 Conclusion

We establish a predictive scaling law for `Aether` through linear probing on eight clinical tasks, with generalization to larger FMs and longer pretraining runs. Our study yields three findings. (1) Compute-optimal training scales FM size and pretraining hours together, with hours growing faster. The fitted law allocates a $10\times$ compute increase to $1.2\times$ larger FMs and $8.2\times$ more pretraining hours. (2) Larger FMs use waveform data more efficiently, and more pretraining data amplifies model-scaling gains. Doubling a 25M FM from a 4.8M-hour reference reduces predicted waveform requirements by 51.8% at fixed performance and labeled-patient count. (3) Pretraining and clinical supervision reinforce each other: more labeled patients increase pretraining gains, while larger FMs and longer pretraining reduce labeling requirements. Extending 720M pretraining from 120K to 36.3M hours reduces the predicted patient requirement by 61% at a target ranking error of 0.32. These findings provide a quantitative recipe for using additional compute to reach target clinical performance with fewer waveform hours and labeled patients. Scaling our training recipe offers a promising and durable path toward physiological modeling and clinical prediction.

REPRODUCIBILITY STATEMENT

The experimental protocol specifies the entire development grid, acquisition rules, candidate heads, selection metric, optimization tolerances, and uncertainty estimand. We will release our code for reproduction upon acceptance.

ETHICS STATEMENT

This study evaluates retrospective waveform prediction. Clinical deployment requires separate assessment of calibration, alert utility, and safety. Patient identifiers support split, acquisition, and resampling bookkeeping; paper artifacts contain aggregate results. Authors confirm dataset access, licensing, and institutional compliance.

AI USE STATEMENT

AI coding and writing assistants supported implementation and manuscript revision. The authors take responsibility for the scientific claims, privacy, and final manuscript.

REFERENCES

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014. URL https://jmlr.org/papers/v15/agarwal14b.html.

M. A. Al-Masud and Nils Strodthoff. Pretraining strategies and scaling for ECG foundation models: A systematic study. *arXiv preprint arXiv:2605.12241*, 2026. doi: 10.48550/arXiv.2605.12241. URL https://arxiv.org/abs/2605.12241.

Ibrahim M. Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22300–22312, 2022. doi: 10.52202/068431-1620. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8c22e5e918198702765ecff4b20d0a90-Abstract-Conference.html.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Hubert Banville, Yohann Benchetrit, Stéphane d'Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*, 2025. doi: 10.48550/arXiv.2501.15322. URL https://arxiv.org/abs/2501.15322.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018. URL https://github.com/jax-ml/jax.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.14891.

Fangke Chen, Sirry Chen, Wei Chen, and Zhongyu Wei. SOTER: A generative time-series foundation model for wearable human physiological signals. *arXiv preprint arXiv:2609.16804*, 2026. URL https://arxiv.org/abs/2609.16804.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):e1009862, 2022. doi: 10.1371/journal.pcbi.1009862. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009862.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hR_SMu8cxCV.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Zongheng Guo, Tao Chen, and Manuela Ferrario. QualityFM: a multimodal physiological signal foundation model with self-distillation for signal quality challenges in critically ill patients. *arXiv preprint arXiv:2509.06516*, 2025. doi: 10.48550/arXiv.2509.06516. URL https://arxiv.org/abs/2509.06516.

Danny Dongyeop Han, Yonghyeon Gwon, Ahhyun Lucy Lee, Taeyang Lee, Seong Jin Lee, Jubin Choi, Sebin Lee, Jihyun Bang, Seungju Lee, David Keetae Park, Shinjae Yoo, Chun Kee Chung, and Jiook Cha. DIVER-1: Scaling intracranial EEG foundation models for transferable representations. *arXiv preprint arXiv:2512.19097*, 2026. doi: 10.48550/arXiv.2512.19097. URL https://arxiv.org/abs/2512.19097v3. Version 3.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021. doi: 10.48550/arXiv.2102.01293. URL https://arxiv.org/abs/2102.01293.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. doi: 10.48550/arXiv.1712.00409. URL https://arxiv.org/abs/1712.00409.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre.

An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. URL https://arxiv.org/abs/2404.06395.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4878278/.

Brennan C. Kahan, Fan Li, Andrew J. Copas, and Michael O. Harhay. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology*, 52(1):107–118, 2023. doi: 10.1093/ije/dyac131. URL https://academic.oup.com/ije/article/52/1/107/6644521.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.19466.

Gul Rukh Khattak, Konstantinos Patlatzoglou, Joseph Barker, Libor Pastika, Boroumand Zeidaabadi, Aidan R. Birdi, Jiayu Huo, Ahmed El-Medany, Hesham Aggour, Yixiu Liang, Antonio H. Ribeiro, Jeffrey Annis, Antonio Luiz Pinho Ribeiro, Junbo Ge, Daniel B. Kramer, Jonathan W. Waks, Evan Brittain, Nicholas Peters, Fu Siong Ng, and Arunashis Sau. Data distribution impacts the performance and generalisability of contrastive learning-based foundation models of electrocardiograms. *PLOS Digital Health*, 5(9):e0001623, 2026. doi: 10.1371/journal.pdig.0001623. URL https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0001623.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2021.653659/full.

Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022. doi: 10.1038/s41551-022-00914-1. URL https://doi.org/10.1038/s41551-022-00914-1.

Gang Li and Kefei Zhou. A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *Journal of the American Statistical*

Association, 103(482):705–713, 2008. doi: 10.1198/016214508000000364. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2832229/.

Jiawei Li, Fabio Bonassi, Ming Jin, Stefan Gustafsson, Johan Sundström, Thomas B. Schön, and Antônio H. Ribeiro. How do electrocardiogram models scale? *arXiv preprint arXiv:2605.17276*, 2026. doi: 10.48550/arXiv.2605.17276. URL https://arxiv.org/abs/2605.17276.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Yunfei Luo, Xi Chen, Yuliang Chen, Lanshuang Zhang, Md Mofijul Islam, Siwei Zhao, Peter Kotanko, Subhasis Dasgupta, Andrew Campbell, Rakesh Malhotra, and Tauhidur Rahman. Toward world modeling of physiological signals with chaos-theoretic balancing and latent dynamics. *arXiv preprint arXiv:2605.15465*, 2026a. URL https://arxiv.org/abs/2605.15465.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *ACM Transactions on Computing for Healthcare*, 2026b. doi: 10.1145/3803808. URL https://doi.org/10.1145/3803808.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024. doi: 10.48550/arXiv.2408.05178. URL https://arxiv.org/abs/2408.05178.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database Matched Subset. *PhysioNet*, 2020. doi: 10.13026/c2294b. URL https://physionet.org/content/mimic3wdb-matched/1.0/. Version 1.0.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376, 2023. doi: 10.52202/075280-2191. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024. doi: 10.48550/arXiv.2410.13638. URL https://arxiv.org/abs/2410.13638.

Nancy A. Obuchowski. Nonparametric analysis of clustered ROC curve data. *Biometrics*, 53(2): 567–578, 1997. doi: 10.2307/2533958. URL https://pubmed.ncbi.nlm.nih.gov/9192452/.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7871f51d332c8661cc729539d7987b21-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, 2020. doi: 10.1038/s41746-020-00323-1. URL https://www.nature.com/articles/s41746-020-00323-1.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=ryenvpEKDr.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. URL https://arxiv.org/abs/2002.05202.

Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore, IV, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32 (2):752–762, 2026. doi: 10.1038/s41591-025-04133-4. URL https://www.nature.com/articles/s41591-025-04133-4.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

John M. Williamson, Somnath Datta, and Glen A. Satten. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1):36–42, 2003. doi: 10.1111/1541-0420.00005. URL https://onlinelibrary.wiley.com/doi/10.1111/1541-0420.00005.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4904fad153f6434a7bcf04465d4be2cc-Abstract-Conference.html.

Weixuan Yuan, Zengrui Jin, Yichen Wang, Donglin Xie, Ziyi Ye, Chao Zhang, and Xuesong Chen. sleep2vec: Unified cross-modal alignment for heterogeneous nocturnal biosignals. *arXiv preprint arXiv:2602.13857*, 2026. doi: 10.48550/arXiv.2602.13857. URL https://arxiv.org/abs/2602.13857.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Zhai_Scaling_Vision_Transformers_CVPR_2022_paper.pdf.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1910.07467.

Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for EHR foundation models. *arXiv preprint arXiv:2505.22964*, 2025. doi: 10.48550/arXiv.2505.22964. URL https://arxiv.org/abs/2505.22964.

A CLINICAL DOWNSTREAM DATASETS

We construct eight clinical prediction tasks by linking physiological waveforms to clinical records in MIMIC-III. Each task uses one hour of waveforms to predict a future clinical outcome. Below, we describe the data sources, prediction windows, task definitions, and eligibility criteria. Table 2 reports cohort sizes for all three data splits.

A.1 DATA SOURCES AND PATIENT SPLITS

We use the MIMIC-III Waveform Database Matched Subset (Moody et al., 2020), which contains bedside recordings of ECG, arterial blood pressure, plethysmography, and respiration. Waveforms are sampled at 125 Hz, with available channels varying across recordings. MIMIC-III v1.4 (Johnson et al., 2016) provides vital signs, laboratory results, medications, procedures, ventilator settings, and outcomes. Eligible ICU stays involve adults with at least one matched waveform record. All downstream patients are held out from pretraining. We split patients into disjoint training, validation, and test cohorts before constructing windows, keeping all stays from each patient together.

A.2 PREDICTION WINDOWS

Each example has a prediction time T, a clinical history $[T - W, T)$, a forecast gap $[T, T + G)$, and an outcome horizon $[T + G, T + G + H)$. We use six hours of clinical history ($W = 6\,\text{h}$) to determine eligibility and prior clinical status. The FM receives only the one-hour waveform segment ending at T; the "Hours" columns in Table 2 sum these input durations. Prediction times follow a sliding schedule for surveillance tasks, daily checkpoints for ventilator readiness, and documented treatment events for fluid response. Table 4 summarizes these task-specific settings.

Table 4: Prediction windows and eligibility for the eight downstream tasks. W denotes clinical history, G the forecast gap, and H the outcome horizon. Each FM input contains one hour of waveforms ending at T. "Targets" gives the number of binary outcomes evaluated per task.

Task	Prediction time	W	G	H	Targets	Eligibility at T
Acute deterioration	Sliding	6 h	0	6 h	3	Eligible for at least one component
Circulatory deterioration	Sliding	6 h	0	6 h	1	No active vasopressors
Dynamic mortality	Sliding	6 h	6 h	24 h	1	All monitored patients
Respiratory escalation	Sliding	6 h	0	6 h	1	No invasive ventilation
Sepsis onset	Sliding	6 h	1 h	5 h	1	No prior Sepsis-3 onset
Fluid response	Infusion start	6 h	30 min	30 min	2	Documented rapid crystalloid course
KDIGO 2/3 AKI	Sliding	6 h	1 h	23 h	2	No stage 2/3 AKI or maintenance dialysis
Ventilator readiness	Daily checkpoint	6 h	0	24 h	1	Active invasive ventilation

A.3 TASK DEFINITIONS

Acute deterioration. We predict three binary outcomes within six hours: initiation of vasopressors, initiation of invasive mechanical ventilation, and a composite of either event. The first two targets include patients who are not receiving the corresponding treatment at T. The composite includes patients eligible for at least one component and is positive when an eligible component occurs.

Circulatory deterioration. We predict the first new vasopressor initiation within six hours among patients receiving no vasopressors at T.

Dynamic mortality. We predict in-hospital death during $[T + 6\,\text{h},\ T + 30\,\text{h})$. All monitored patients are eligible at T, with a six-hour gap between the waveform input and the outcome horizon.

Respiratory escalation. We predict new invasive ventilation within six hours among patients receiving no support, oxygen, high-flow nasal cannula, or non-invasive ventilation at T.

Sepsis onset. We predict the first Sepsis-3 onset during $[T + 1\,\mathrm{h},\ T + 6\,\mathrm{h})$. Onset requires suspected infection, identified by a culture order paired with antibiotic administration, and an acute increase of at least two points in the Sequential Organ Failure Assessment (SOFA) score. Patients with sepsis onset at or before T are ineligible.

Fluid response. The prediction time T is the start of a documented rapid crystalloid course; overlapping courses are merged into a single event. We compare median mean arterial pressure during minutes 30–60 after treatment begins with the pre-treatment reference. The two targets require increases of at least 15% and 5 mmHg, respectively. These endpoints measure the documented blood-pressure response following fluid administration.

KDIGO stage 2/3 acute kidney injury. We predict the first creatinine-defined stage 2 or 3 acute kidney injury during $[T + 1\,\mathrm{h},\ T + 24\,\mathrm{h})$, using Kidney Disease: Improving Global Outcomes (KDIGO) criteria. Eligible patients have no prevalent stage 2/3 injury or maintenance dialysis and at least one creatinine measurement in the outcome horizon. A second target uses urine-output criteria when hourly urine measurements, a reliable body weight, and catheter information are available.

Ventilator readiness. At each daily checkpoint during invasive ventilation, we predict liberation within 24 hours. The label is positive if ventilation ends during this period and the patient remains free of invasive ventilation for the following 48 hours.

A.4 Label availability and eligibility

We assign labels only when the outcome can be determined from the available records. For incident-event targets, patients with the event already present at T are ineligible. Surveillance examples with an event during the forecast gap are also excluded. A negative label requires sufficient follow-up to establish that the event did not occur within the outcome horizon. Unavailable labels are masked from the corresponding target's loss and metrics. Eligibility and label availability are determined separately for each target. Following Section 3.3, we compute patient-balanced AUROC for each target and average across targets to obtain the task score. Repeated or overlapping windows retain their patient identity; each patient's inputs receive equal total weight within a target.

A.5 Leakage controls

FM inputs contain only waveform samples recorded before T. Clinical history used to determine eligibility is restricted to information charted before T; subsequent records are used only to establish outcomes. Post-prediction interventions, final diagnoses and disposition, realized length of stay, and outcome labels are excluded from model inputs. Documented treatment before T informs eligibility: for example, active vasopressor treatment excludes a new-start target. Waveform normalization uses only the input window, and feature standardization for linear probing uses the selected training patients. For each example, we retain its data version, patient split, prediction time, input and outcome intervals, and available-channel mask to support reproduction.

B Expected Forms of the Scaling Law

Ranking risk. For patient-weighted positive and negative examples X^+ and X^-, the ranking error of a score s is (Agarwal, 2014):

$$\mathcal{R}(s) = \Pr\big[s(X^+) < s(X^-)\big] + \frac{1}{2}\Pr\big[s(X^+) = s(X^-)\big] = 1 - \mathrm{AUROC}(s). \tag{7}$$

A constant score gives $\mathcal{R} = 1/2$. Macro averaging applies this definition across targets and tasks, yielding the eight-task ranking error modeled in Section 4.2.

Resource-dependent contributions. Following the expected-risk decomposition of Hoffmann et al. (2022), let $\mathcal{R}_*$ be the best population ranking risk and $\mathcal{R}^*_N$ the best risk attainable by a size-N FM and a linear probe. For the pretrained FM representation $\widehat{\phi}_{N,D}$, define $\mathcal{R}^*_{\mathrm{LP}}(\widehat{\phi}_{N,D})$ as the best

population risk of a linear probe on that representation. With the learned score $\widehat{s}_{N,D,P}$, we obtain

$$\begin{aligned}\mathbb{E}\mathcal{R}(\widehat{s}_{N,D,P}) = \mathcal{R}_* &+ (\mathcal{R}_N^* - \mathcal{R}_*) \\ &+ \left[\mathbb{E}\mathcal{R}^*_{\mathrm{LP}}(\widehat{\phi}_{N,D}) - \mathcal{R}_N^*\right] \\ &+ \left[\mathbb{E}\mathcal{R}(\widehat{s}_{N,D,P}) - \mathbb{E}\mathcal{R}^*_{\mathrm{LP}}(\widehat{\phi}_{N,D})\right]. \end{aligned} \tag{8}$$

The three resource-dependent terms describe the contributions of FM capacity, pretraining, and labeled-patient supervision. For the theoretical data-scaling analysis, we adopt the single-pass view of Hoffmann et al. (2022): D measures pretraining data and the associated training progress. Frozen linear probing adds the supervision resource P to this description.

Expected forms of the scaling terms. To model the observed scaling regime, we express ranking error relative to its advantage over chance through $\mathcal{R}/(1-2\mathcal{R})$ for $0 < \mathcal{R} < 1/2$. Motivated by the capacity and training contributions in Equation 8, we assume decreasing power-law terms AN^{-a} and BD^{-b}, with positive amplitudes and exponents. We model the response to labeled-patient supervision with a common factor P^{-g}. These assumptions give

$$\frac{\widehat{\mathcal{R}}}{1-2\widehat{\mathcal{R}}} = (AN^{-a} + BD^{-b})P^{-g}. \tag{9}$$

Solving directly for predicted ranking error yields

$$\widehat{\mathcal{R}}(N,D,P) = \left[2 + \frac{P^g}{AN^{-a} + BD^{-b}}\right]^{-1}. \tag{10}$$

The resulting response lies between zero and the chance error of $1/2$ and decreases with all three resources. It also preserves the allocation ordering implied by the power-law terms, connecting pretraining and supervision in a common predictive model.

Fitting the parametric law. We fit A, B, a, b, g by nonlinear least squares on K complete eight-task resource configurations[2]:

$$\min_{A,B,a,b,g>0} \sum_{i=1}^{K} \left(\widehat{\mathcal{R}}(N_i, D_i, P_i) - \mathcal{R}_i\right)^2. \tag{11}$$

Each outcome is the equal-task mean of patient-level macro AUROC converted to ranking error. The fit uses exact FM parameter counts, pretraining waveform hours, and the corresponding mean absolute labeled-patient counts. We use 30 deterministic initializations and retain the solution with the smallest objective. Full-precision coefficients generate every prediction and contour; displayed coefficients are rounded to three decimal places. For resource-level prediction, we omit all configurations at one value of N, D, or P and refit on the remaining configurations. There are four model-size levels, six pretraining-hour levels, and four patient-cohort levels. We average prediction MAE across levels within each resource axis and then equally across the three axes. The 2.1B and longer-duration predictions use the same fitted law, connecting this procedure to Section 4.3.

Resource gains and data efficiency. The fitted law makes proportional resource gains explicit:

$$-\frac{\partial \log \widehat{\mathcal{R}}}{\partial \log N} = (1-2\widehat{\mathcal{R}})\frac{aAN^{-a}}{AN^{-a} + BD^{-b}}, \tag{12}$$

$$-\frac{\partial \log \widehat{\mathcal{R}}}{\partial \log D} = (1-2\widehat{\mathcal{R}})\frac{bBD^{-b}}{AN^{-a} + BD^{-b}}, \tag{13}$$

$$-\frac{\partial \log \widehat{\mathcal{R}}}{\partial \log P} = (1-2\widehat{\mathcal{R}})g. \tag{14}$$

[2]We use $K = 140$ resource configurations in our experiments.

All three quantities are positive. The labeled-patient response is largest throughout the fitting grid, supporting its role as the leading downstream scaling factor. At fixed P and target error, changing FM size from N to N' gives

$$D' = \left[D^{-b} + \frac{A}{B} \left(N^{-a} - N'^{-a} \right) \right]^{-1/b}. \tag{15}$$

Thus a larger FM reaches the same predicted performance using fewer pretraining waveform hours. The example in Section 4.2 uses $N = 25$M, $D = 4.8$M hours, and $N' = 50$M, giving $D' = 2.3$M hours with the full-precision fit.

Compute-optimal allocation. For fixed P, Equation 10 increases with $AN^{-a} + BD^{-b}$. Under $C \simeq 6qND$, its minimum therefore solves

$$\min_N \left[AN^{-a} + B \left(\frac{C}{6qN} \right)^{-b} \right]. \tag{16}$$

Differentiating with respect to $\log N$ balances the two marginal contributions, $aAN^{-a} = bBD^{-b}$, and gives Equations 4–5. At each compute value, the green frontier in Figure 6 uses $D = C/(6qN)$ and minimizes error over the observed FM-size range. The minimum and maximum FM sizes bound the allocation, giving a frontier with interior and boundary solutions.