Learning Representations from Incomplete EHR Data with Dual-Masked Autoencoding

Xiao Xiang XIAO0605@MIT.EDU
Massachusetts Institute of Technology, École Polytechnique Fédérale de Lausanne

David Restrepo DAVIDRES@MIT.EDU
Massachusetts Institute of Technology

Hyewon Jeong HYEWONJ@MIT.EDU
Massachusetts Institute of Technology

Yugang Jia YUGANG@MIT.EDU
Massachusetts Institute of Technology

Leo Anthony Celi LCELI@MIT.EDU
Massachusetts Institute of Technology, Harvard University, Beth Israel Deaconess Medical Center

Abstract

Electronic health records (EHR) arrive masked. Clinicians order measurements selectively, and any patient table thus contains only a subset of the values that characterize the underlying physiological state. Prior masked modeling approaches on EHR data either impute the table before learning, represent missingness through a dedicated placeholder signal, or optimize solely for imputation, which limits the representations they learn for downstream clinical tasks and carries every unobserved entry through the encoder. We introduce **AID-MAE**, an **A**ugmented-**I**ntrinsic **D**ual-Masked Autoencoder that learns directly from incomplete tables by combining the intrinsic mask the record already carries with an augmented mask that hides a subset of observed values for reconstruction during pretraining. Neither type of masked entry enters the encoder, so attention operates only over what was observed. AID-MAE achieves consistent improvements over strong baselines across multiple clinical tasks on two datasets. Across experiments, we discuss that recovering the missing entries is not a prerequisite for learning and show that the representations learned carry clinical structure without supervision.

1. Introduction

Clinical decisions are routinely made from incomplete records, and clinical signals are coupled by the underlying physiology that generates them. For example, a systemic infection triggers inflammation that lowers blood pressure (Jarczak et al., 2021), clinicians administer vasopressors to sustain perfusion (Evans et al., 2021), prolonged support can lead to tissue hypoxia and rising lactate, an early warning of organ failure (Jozwiak et al., 2022), and heart rate, oxygen saturation, and temperature shift alongside, reflecting the same underlying process (Mao et al., 2018). An ordered measurement therefore carries information about the others that were ordered, and about those that were not, so the observed values can place a patient without the record being complete. This motivates our core question: can we learn directly from this partial information, without completing it first?

Electronic Health Records (EHRs) contain irregularly-sampled time series with diverse, feature-specific missingness (Li et al., 2021). Existing methods approach this data in different ways to address the irregular sampling of clinical time series, including a set-based representation that encodes time series as triplets of time, variable, and value, accommodating missingness by construction (Horn et al., 2020; Tipirneni and Reddy, 2022; Oufattole et al., 2024). Other methods (Labach et al., 2023; Restrepo et al., 2025) transformed irregularly-sampled time series onto a regular grid (tabular format) with missing entries. The intrinsic missingness in the resulting EHR tables is pervasive, heterogeneous and contains information (Groenwold, 2020), and existing deep learning approaches exhibit difficulties in outperforming classical methods (Shwartz-Ziv and Armon, 2022), such as gradient-boosted decision trees (Chen and Guestrin, 2016) on many supervised tabular benchmarks (Grinsztajn et al., 2022).

Masked autoencoding (He et al., 2022) in the vision literature learns from a deliberately incomplete view, relying on the redundancy among pixels, where a hidden patch is recoverable from its neighbors. In EHR, a tabular format gives the model the same access to a value's neighbors, aligning features with time. Capturing this cross-temporal and cross-feature context (Futoma et al., 2017; Che et al., 2018; Labach et al., 2023) is essential for the models to understand the time series data. These models, however, still resolve the empty cells before encoding, whether by imputation or by an explicit missingness signal (Labach et al., 2023). Feeding the encoder only the observed entries with masked autoencoding could remove this step, which changes both what the model computes over and what it has to learn from.

A small observation strengthens our motivation to investigate further: pretraining AID-MAE on MICE-imputed (Buuren and Groothuis-Oudshoorn, 2011) PhysioNet Challenge 2012 data (Silva et al., 2012) degrades performance relative to pretraining on the incomplete table directly (Appendix A.2).

We hence make the following contributions towards learning representations directly from incomplete records:

- We introduce a dual-mask mechanism for EHR table pre-training. An intrinsic mask marks events that were not observed and an augmented mask hides a random subset of observed values. Reconstruction targets come from the augmented subset only, and neither kind of masked token enters the encoder. AID-MAE does not require imputation or a missingness-specific input, and its attention cost follows approximately the number of observed values rather than the strict size of the grid.

- We examine what the dual mask contributes beyond outperforming tree-based and self-supervised baselines on downstream tasks across datasets. Since unobserved features never enter the encoder, AID-MAE transfers to a dataset with a different feature set without modification, while maintaining its stronger performance. We also discuss how design choices tuned for imputation quality do not carry the same interpretation in general-purpose representation learning.

- We provide qualitative analyses of the learned representations. Patient-level embeddings separate ICU admission types without supervision, and feature-level embeddings

organize laboratory measurements into clusters consistent with clinical groupings, suggesting that the model captures physiologically meaningful structure beyond what the training objective explicitly requires.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work provides evidence for learning representations directly from what was observed, rather than imputing incomplete records into artificial completeness before modeling. Representations learned this way outperform both tree-based and prior self-supervised baselines on every task (Table 2), and transfer to a second EHR dataset despite substantial differences in feature availabilities and patient populations (Table 3), suggesting that learning from incomplete data produces representations robust to the feature heterogeneity that typically hinders multi-site deployment. We also find that design choices tuned for imputation quality do not carry over. A masking schedule that upweights rarely observed features helps imputation benchmarks but leaves downstream result flat (Table 4).

2. Related Work

2.1. Masked Autoencoding for Incomplete Tabular Data

Masked autoencoders (MAE) (He et al., 2022) have been adapted to tabular data (Majmundar et al., 2022; Du et al., 2024; Kim et al., 2025). ReMasker (Du et al., 2024) applies MAE for self-supervised imputations on data with simulated missingness. PMAE (Kim et al., 2025) further adds a proportional masking schedule that corrects the imbalance in mask sampling propensity, and improves imputation accuracies. Lab-MAE (Restrepo et al., 2025), extends this line to imputing lab values in EHRs. However, imputation accuracy remains the reported objective in all three, and none evaluates the learned representation on clinical downstream tasks. Xu et al. (2025) explored representation learning for wearables time series, where gaps arise from device dropout on a regular lattice rather than from ordering decisions over irregularly sampled clinical events.

2.2. Self-Supervised Learning for EHR Time Series

Recent self-supervised approaches for EHR time series explore diverse pretext tasks and data representations. STraTS (Tipirneni and Reddy, 2022) uses a forecasting pretext task with the set representation (Horn et al., 2020) to learn contextual embeddings under sparsity and irregular sampling. EBCL (Oufattole et al., 2024) instead focuses on temporally local information by contrasting representations before and after clinically significant events. Within the masked modeling approaches, Labrador (Bellamy et al., 2025) showed that masked modeling yields meaningful embeddings, while also highlighting the difficulty of surpassing tree-based methods with self-supervised approaches on clinical benchmarks (Chen and Guestrin, 2016). DuETT (Labach et al., 2023) alternates attention across time and events to capture dependencies in a matrix input, fills unobserved entries with zeros and then adds a missingness token to mark them, which competes for attention and computation. In contrast, AID-MAE operates directly on incomplete matrices, explicitly leveraging interactions among available features, efficiently learning contextual representations that are well suited for various downstream tasks.

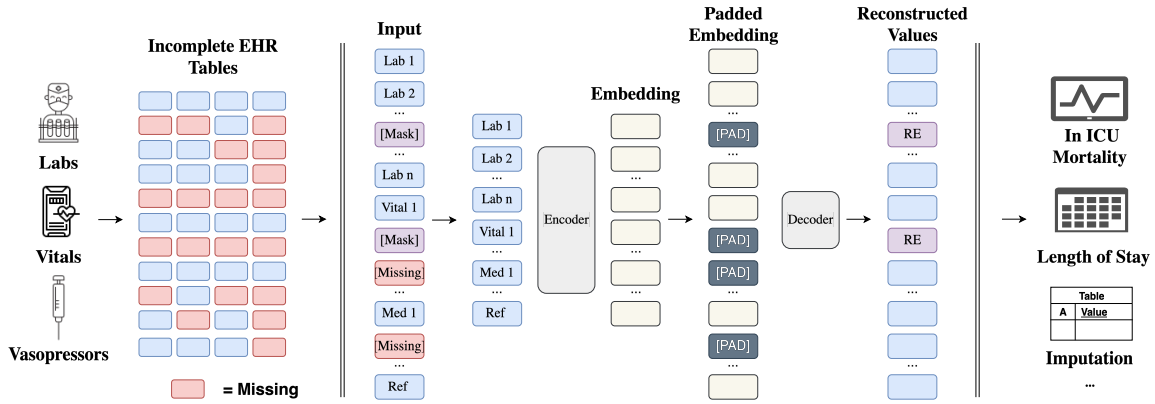

Figure 1: **AID-MAE (Augmented-Intrinsic Dual-Masked AutoEncoder)** A subset of observed tokens of the original data with inherent missingness `MISSING` is randomly masked (`[MASK]`). Each measurement (value and timestamp) is embedded with positional encodings. The encoder processes only unmasked tokens, and the decoder receives the encoded representations along with a learned padding token `[PAD]` in place of missing or masked entries. Training optimizes a dual loss of reconstructing unmasked values and predicting features under augmented masking, while intrinsically missing entries are excluded from the loss. Code: https://github.com/xiaoxiang0605/AID-MAE

3. AID-MAE: Augmented-Intrinsic Dual-Masked Autoencoder

EHR time series are recorded at irregular time points. Before training, we transform each patient's heterogeneous time-series onto a fixed, ordered grid of length L. The resulting token array $\mathbf{x} \triangleq (x_1, \ldots, x_L) \in \mathbb{R}^L$ is embedded and processed by our model with two distinct masks.

Intrinsic Missingness Mask: Given many slots contain no measurement, we represent this with a binary vector

$$\mathbf{m} \triangleq (m_1, \ldots, m_L) \in \{0,1\}^L, \quad m_i = \begin{cases} 1 & \text{if token } i \text{ is recorded,} \\ 0 & \text{if token } i \text{ is missing.} \end{cases}$$

On the index set $\mathcal{I} = \{1, \ldots, L\}$, we define recorded set R and missing set M as:

$$R := \{\, i \in \mathcal{I} : m_i = 1 \,\} \qquad M := \mathcal{I} \setminus R = \{\, i \in \mathcal{I} : m_i = 0 \,\}$$

Augmented Mask: Given the intrinsic missingness mask $\mathbf{m} \in \{0,1\}^L$, we sample an augmented mask $\mathbf{m}' \in \{0,1\}^L$ on the recorded tokens, where $m'_i = 1$ if token i is kept and $m'_i = 0$ if it is hidden by the augmented mask. We define:

$$R \setminus A \;=\; \{\, i \in \mathcal{I} : m_i = 1 \;\wedge\; m'_i = 1 \,\}, \quad A \;=\; \{\, i \in \mathcal{I} : m_i = 1 \;\wedge\; m'_i = 0 \,\}$$

where $R \setminus A$ contains tokens that are unmasked and A contains tokens hidden by augmented masks.

Building on the masked autoencoder paradigm, our architecture consists of an encoder-decoder Transformer with fixed sinusoidal positional encodings for each feature position (Vaswani et al., 2017; Du et al., 2024). Given per-token embeddings $\mathbf{z}_i \in \mathbb{R}^d$, only unmasked features $\{\, \mathbf{z}_i : i \in R \setminus A \,\}$ enter the encoder f, whose layers comprise multi-head self-attention, feed-forward blocks, residual connections, and layer normalization.

Notably, most prior work on masked autoencoding for incomplete tables include unmasked tokens by cropping each array in the batch to the minimum observed input array length (Du et al., 2024; Kim et al., 2025), which would remove significantly more measurements for longer sequences, often coming from high acuity patients (Agniel et al., 2018). In practice, this reduces, for example, a septic patient with hourly blood gases and a full daily laboratory panel to the same length as a stable patient with a single morning draw, discarding precisely the dense monitoring that makes critically ill stays most informative. We instead fix length at the batch maximum (Restrepo et al., 2025), denoted as ℓ_{keep} and right pad shorter rows with a dedicated token to this length. We then apply a binary attention mask $\Gamma \in \{0,1\}^{\ell_{\text{keep}} \times \ell_{\text{keep}}}$ so that padded tokens receive zero attention. This preserves equal loss per measurement, prevents pad interactions, and keeps compute at $O(\ell_{\text{keep}}^2)$ per batch.

Input Representation: Each feature (both numerical value and its associated numerical time) is linearly projected to a d-dimensional embedding. We combine the value embedding and time embedding for every measurement. Let $\mathbf{X} \in \mathbb{R}^{L \times d}$ denote the ordered token array of length L with embedding size d, we add positional information to form the initial embedded array before masking: $\mathbf{Z} := \mathbf{X} + \mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{L \times d}$ denotes the positional encoding. The token indices define a fixed order for features columns.

Reconstruction: Denote the encoder outputs by $\mathbf{h}_i \in \mathbb{R}^d$ and stack them as $\mathbf{H} = \big(\mathbf{h}_i\big)_{i \in R \setminus A} = f\big(\{\mathbf{z}_i : i \in R \setminus A\}\big) \in \mathbb{R}^{\ell_{\text{keep}} \times d}$, where $\mathbf{z}_i \in \mathbb{R}^d$ is the embedding of token i, ℓ_{keep} is the maximum length of the fed input arrays in a batch, and $\mathbf{h}_i$ denotes its encoded representation. We pad a shared learnable mask token $\mathbf{m}_{\text{token}} \in \mathbb{R}^d$ at all masked positions, i.e. at both intrinsic missing positions $M = \mathcal{I} \setminus R$ and positions with augmented mask $A \subseteq R$. We once again add positional encoding:

$$\mathbf{z}_i^{\text{dec}} = \begin{cases} \mathbf{h}_i, & i \in R \setminus A, \\ \mathbf{m}_{\text{token}}, & i \in A \cup M \end{cases} \qquad \mathbf{Z}_{\text{dec}} = \big(\mathbf{z}_1^{\text{dec}}, \ldots, \mathbf{z}_L^{\text{dec}}\big)^\top + \mathbf{P}$$

The decoder g predicts the original values at positions with both augmented masks and intrinsic masks, as in Figure 1.

Training Objective: The model is pretrained by optimizing dual reconstruction loss, which reconstructs features in both set of unmasked tokens $R \setminus A$ and augmented masks A, while ignoring intrinsic missing entries (Du et al., 2024). The reconstruction loss for a training example is defined as:

$$\mathcal{L} = \frac{1}{|R(\mathbf{m}) \setminus A(\mathbf{m},\mathbf{m}')|} \sum_{i \in R(\mathbf{m}) \setminus A(\mathbf{m},\mathbf{m}')} (\hat{x}_i - x_i)^2 + \frac{1}{|A(\mathbf{m},\mathbf{m}')|} \sum_{i \in A(\mathbf{m},\mathbf{m}')} (\hat{x}_i - x_i)^2$$

The separation of intrinsic and augmented masks is motivated by a key asymmetry. Augmented-masked entries have known ground-truth values and provide valid reconstruction targets, whereas intrinsically missing entries do not. Including intrinsically missing positions in the

encoder input or the reconstruction loss would require imputing targets that no clinician chose to measure, introducing potentially unreliable supervision into the training signal.

4. Experiments

4.1. Data and Tasks

Data We have selected and preprocessed the most frequently sampled 50 laboratory values, five vital signs (heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, and temperature), oxygen saturation, and five vasopressors (norepinephrine, epinephrine, vasopressin, dopamine, and phenylephrine) from MIMIC-IV ICU (Johnson et al., 2023). For the second dataset, PhysioNet Challenge 2012 (Silva et al., 2012), we included 23 lab values and 4 vital signs that are also included in our curated MIMIC-IV dataset. Each event contains an item ID, a timestamp, and a numerical value. We first cap each numeric feature at the 5th and 95th percentiles (winsorization) to reduce extreme outliers (Wilcox, 2011). We then apply a per-feature min–max normalization. If the values are not recorded, those values are represented as a missing entry. We provide a full list of included features in Table 10.

In particular, we applied signal-specific preprocessing for each type of events. For each input array:

- **Laboratory values** were selected with a daily frequency. We include a reference value for each lab, which corresponds to the most recent corresponding lab result prior to the day. This aims to align with real-life clinical setting and provide a longitudinal baseline.
- **Vital signs** were included with finer granularity of 1 hour. Each vital sign constitutes 24 data points in an input array.
- **Vasopressors** were carried forward for each of the recorded dosages from its start time till its end time. Vasopressor data are not available in the PhysioNet Challenge 2012 dataset, inducing a feature-level distribution shift that allows us to assess robustness and cross-dataset generalization.

We include all ICU patient stays, discarding input arrays with very few measurements or without any laboratory measurement. The final datasets comprise 92,938 stays spanning 412,365 patient-days for MIMIC-IV, and 11,987 stays with 23,682 patient-days in PhysioNet Challenge 2012. More detailed information on the data pre-processing can be found in Appendix B.2.

Task Definitions For downstream evaluation, we construct one sample per patient by restricting inputs to the first 24 hours of the ICU admission, following common benchmarks (Purushotham et al., 2018; Johnson and Mark, 2018; Hempel et al., 2023; Yeh et al., 2024). This design reflects clinical practice, where physicians perform early assessments within hours of admission and subsequently refine decisions as new measurements become available (Rivers et al., 2001).

We evaluate model performance on the following downstream clinical prediction tasks:

- **Mortality**: A binary indicator of death during the admission. On MIMIC-IV we use death before ICU discharge; on PhysioNet Challenge 2012 we use the in-hospital outcome provided with the dataset.
- **Length of Stay (LOS)**: A binary indicator denoting whether ICU length of stay is strictly less than 72 hours from admission.
- **Acute Kidney Injury (AKI)**: Defined using serum creatinine measurements according to the criteria proposed by Khwaja (2012). Patients without any creatinine measurement were excluded from AKI evaluation.

Mortality and LOS are evaluated on MIMIC-IV, while mortality and AKI are evaluated on the PhysioNet Challenge 2012 dataset. Task label distributions are reported in Table 1:

Table 1: Downstream tasks and cohort sizes for MIMIC-IV and PhysioNet Challenge 2012.

MIMIC-IV

Task	# Stays	Prevalence
Mortality	92,938	7.6%
LOS < 72h	92,938	66.1%

PhysioNet Challenge 2012

Task	# Stays	Prevalence
Mortality	11,981	14.3%
AKI	11,811	29.3%

4.2. Results

AID-MAE Outperforms All Baselines on Fine-Tuning Tasks Table 2 shows that our model outperforms a range of baselines consistently on all tasks evaluated. This includes XGBoost (Chen and Guestrin, 2016), as well as masked modeling method (Labach et al., 2023). We note that the supervised transformer is trained with the same architecture, with weights initialized at random and updated only through task-specific supervision. This demonstrates that pre-training is an essential component in the superior performance over purely supervised training. Notably, AID-MAE achieves these results with under one million trainable parameters compared to 5.2 million in DuETT, suggesting the gains are attributable to the architectural design rather than increased model capacity. We report fine-tuning details of our model and the baseline models in Appendices C and D.1, and the complete result in Table 6.

Linear Probing Shows Pretrained Embeddings Transfer Strongly Across Data Regimes Figure 2 shows that our model achieve superior results compared to the DuETT baseline (Labach et al., 2023) in all data availabilities. We additionally include logistic regression with median imputation on raw data as a reference. The performance gap is pronounced in low-data regimes, suggesting our pretrained embeddings encode informative inductive biases, especially when labeled samples are scarce. The full linear probing results are shown in Figure 5. We additionally note that linear probing asks how much useful information is encoded in the model's learned patient representations, without any further training.

Table 2: Fine-tuning results on MIMIC-IV and PhysioNet Challenge 2012. We report AUROC (mean ± SD over 5 seeds). Best scores per column are in **bold**. AID-MAE outperforms all baselines.

Model	MIMIC-IV		PhysioNet Challenge 2012	
	Mortality	LOS	Mortality	AKI
Logistic Regression	80.7 ± 0.0	68.8 ± 0.0	72.7 ± 0.0	74.8 ± 0.0
Supervised Transformer	83.9 ± 0.5	72.3 ± 0.5	76.7 ± 1.4	76.1 ± 0.9
XGBoost	86.7 ± 0.0	76.9 ± 0.0	76.9 ± 0.0	77.1 ± 0.0
DuETT	86.4 ± 0.2	77.2 ± 0.0	77.7 ± 0.4	76.7 ± 0.2
AID-MAE (ours)	**87.7 ± 0.1**	**77.6 ± 0.1**	**78.2 ± 0.3**	**77.3 ± 0.1**

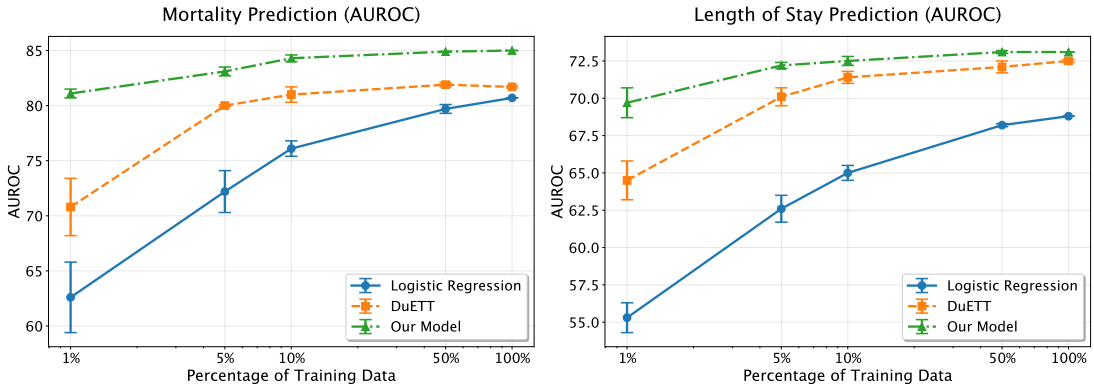

Figure 2: Linear probing results for mortality prediction and length of stay prediction tasks. We compare our model against Logistic Regression with raw data (median imputed) and DuETT with frozen encoder, across different training data percentages (1%, 5%, 10%, 50%, 100%). Error bars represent standard deviation across 5 seeds.

Learned Laboratory Feature Embeddings Exhibit Clinically Coherent Organization We analyze whether the learned encoder organizes laboratory measurements into coherent regions of representation space. For each laboratory feature, we extract its embedding from a pretrained AID-MAE model on MIMIC-IV and project a random subset of $N = 100{,}000$ embeddings into two dimensions using UMAP (McInnes et al., 2018) with a Euclidean metric.

As shown in Figure 3, the embedding space exhibits clear, clinically meaningful structure. Related laboratory tests form distinct and neighboring clusters, including Platelet Count and White Blood Cell count (both hematology measures), as well as Hemoglobin and Hematocrit, which are tightly coupled in clinical practice. Importantly, UMAP is label-agnostic and lab identities are not provided to the model, indicating that these groupings emerge from the learned geometry rather than supervision.

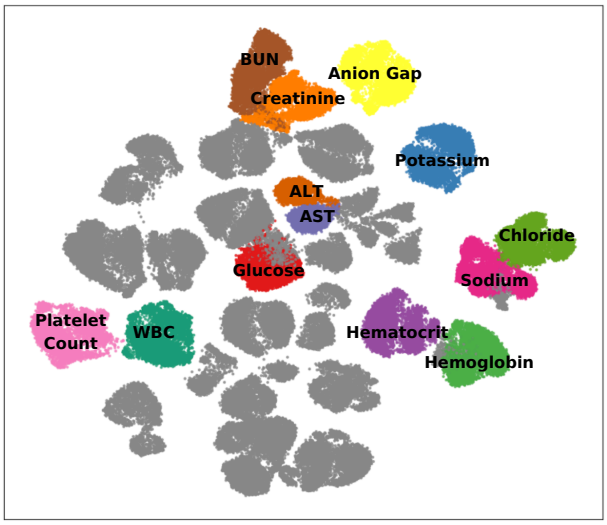

Figure 3: UMAP of feature embeddings. UMAP projection of $N = 100{,}000$ randomly sampled 64-D embeddings for 50 lab features. Colors denote 13 highlighted lab types. Each point corresponds to one measurement embedding. We denote two important patterns: Bottom-left: neighboring pair of pink (Platelet Count) and green islands (WBC); Bottom-right: neighboring pair of purple (Hemoglobin) and green islands (Hematocrit). The geometrical neighboring is consistent with their clinical coupling.

Learned Embeddings Stratify ICU Admissions Into Latent Subtypes Beyond linear probing, we explore the extent to which the learned embedding space naturally stratifies patients by physiological state. We assessed this by extracting the CLS embeddings, an input array level summary (Devlin et al., 2019), for medical intensive care unit (MICU) and the cardiothoracic vascular intensive care unit (CVICU) admissions. The CLS embeddings successfully differentiate stays admitted to the MICU and CVICU into distinct latent subgroups. This is followed by applying k-means clustering to the original embedding space.

Figure 4 reveals two well-separated clusters in the UMAP space (McInnes et al., 2018): Cluster 0 contains 97% CVICU admissions and Cluster 1 contains 80% MICU admissions. The high homogeneity of each cluster with respect to ICU type indicates that the learned representations capture clinically meaningful structure reflecting unit-specific physiology and care contexts.

We note that, to determine the optimal number of clusters, we conducted a silhouette analysis over k=2 to 10 clusters. The analysis identified k=2 as optimal, with a silhouette score of 0.42, which indicates moderate but meaningful separation between patient subtypes.

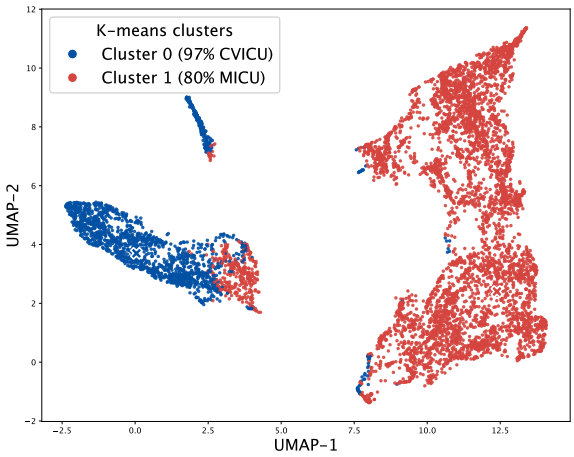

Figure 4: UMAP visualization of first-day CLS embeddings for initial MICU and CVICU admissions. Colors represent clusters from K-means ($k = 2$) applied in the embedding space.

5. Towards Generalizable EHR Representations

5.1. Cross-Dataset Considerations

Different EHR datasets exhibit significant distribution shifts in patient populations, institutional care practices, and measurement availability (Burger et al., 2024). A model pretrained on one dataset will encounter a different set of recorded variables, different missingness rates, and different clinical protocols on another. Understanding how pretrained representations behave under these shifts is important for assessing their practical utility.

One consideration is how the model handles features present during pretraining but absent at evaluation time. DuETT (Labach et al., 2023), for example, requires absent features to be carried through the encoder as explicit missing tokens to preserve positional compatibility, introducing unnecessary computation. AID-MAE sidesteps this by excluding unobserved features from the encoder, allowing it to operate on whatever subset of features is available without architectural modification.

Cross-Dataset Transfer: To evaluate this, we transfer MIMIC-IV (Johnson et al., 2023) pretrained weights to PhysioNet Challenge 2012 (Silva et al., 2012), a dataset with substantially fewer features, no vasopressor data, and a different patient population. As shown in Table 3, frozen AID-MAE representations outperform DuETT by 2.2 and 3.5 AUROC points on mortality and AKI under linear probing, suggesting that the pretrained representations capture clinical structure that generalizes beyond the pretraining dataset. Under fine-tuning, AID-MAE maintains consistent gains across both tasks.

Logit-Based Proportional Reweighting Driven by differences in ordering practice, missingness varies substantially within a single dataset, and across healthcare centers. Among features in EHRs, some variables are nearly always observed, while others are

Table 3: Transfer learning performance from MIMIC-IV pre-trained weights to PhysioNet Challenge 2012. We report AUROC for mortality and acute kidney injury (AKI) prediction under linear probing and full fine-tuning.

Training Regime	Method	Mortality	AKI
Linear Probing	Random Weights	49.8	52.7
	DuETT	70.9	66.2
	AID-MAE (ours)	**73.1**	**69.7**
Fine-Tuning	Random Weights	76.7 ± 1.4	76.1 ± 0.9
	DuETT	77.8 ± 0.3	76.4 ± 0.1
	AID-MAE (ours)	**78.6 ± 0.2**	**77.0 ± 0.2**

scarcely recorded (Li et al., 2021). In our data, for instance, serum sodium is missing around 7% of patient–day entries, whereas arterial oxygen saturation is absent in over 88% of cases. We thus were motivated to test a proportional augmented masking scheme explicitly designed to counterbalance the heterogeneous missingness within a single dataset and potential distribution shifts across datasets. A simple reweighting $r_j = \pm\, p_{\text{miss},j}$, given how much a feature is missing per batch, may collapse sampling to one regime (dense vs. sparse) (Cui et al., 2019; Li et al., 2019). We therefore map missing frequencies into weights by using a logit transform, a schedule that has the desired convex-concave curvature (Kim et al., 2025):

$$w_j = \begin{cases} a \log \dfrac{p_{\text{miss},j}}{1 - p_{\text{miss},j}} + b, & \text{if } p_{\text{miss},j} < 1, \\ 0, & \text{if } p_{\text{miss},j} = 1. \end{cases}$$

where the parameter a determines the resampling direction and offset b serves as a regularizer to avoid extreme weights when $p_{\text{miss},j}$ approaches zero. In light of these considerations, we examine whether this correction, developed for imputation, carries over to representation learning in the following ablation studies.

5.2. Ablation Studies on Masking and Input Design

Masking Ratio and Reweighting Table 4 reports results across masking ratios and reweighting hyperparameters. Downstream AUROC varies under one point over the whole grid, and neither sign of a is consistently better than random masking. We therefore use random masking at 25% ($a = 0$, $b = 0.25$) for all reported results.

The reweighting proposed by Kim et al. (2025) increases the masking probability of features with high missingness. This is similar to an inverse propensity score correction (Rosenbaum and Rubin, 1983). Rare but diagnostically relevant features receive an augmented mask with a higher probability and receive prediction signals, ensuring that their contribution to the gradient signal is not washed out by frequent variables (Seaman and White, 2013). This has been proved useful for imputation benchmarks with synthetic missingness (Kim et al., 2025).

Table 4: Ablation study on proportional masking hyperparameters. Performance is reported as mean ± standard deviation across 5 random seeds. Downstream performance is largely flat across the grid

a	b	Mortality		Length of Stay	
		AUROC	**AUPRC**	**AUROC**	**AUPRC**
Random masking ($a = 0$)					
0	0.125	87.4 ± 0.1	49.4 ± 0.2	76.8 ± 0.2	62.0 ± 0.1
0	0.25	87.7 ± 0.1	49.8 ± 0.1	77.6 ± 0.1	63.1 ± 0.1
0	0.5	87.3 ± 0.1	49.0 ± 0.1	77.2 ± 0.2	62.0 ± 0.3
0	0.75	87.5 ± 0.1	49.8 ± 0.2	76.8 ± 0.2	61.9 ± 0.1
Fixed $b = 0.25$					
-0.025	0.25	87.2 ± 0.1	49.7 ± 0.3	77.7 ± 0.0	63.4 ± 0.1
-0.0125	0.25	87.2 ± 0.0	48.9 ± 0.2	77.3 ± 0.1	62.8 ± 0.2
0.0125	0.25	87.5 ± 0.0	49.8 ± 0.0	77.4 ± 0.1	63.2 ± 0.1
0.025	0.25	87.7 ± 0.1	50.4 ± 0.1	77.3 ± 0.1	62.5 ± 0.1
Fixed $b = 0.5$					
-0.05	0.5	87.6 ± 0.0	49.9 ± 0.1	77.8 ± 0.0	63.2 ± 0.0
-0.025	0.5	87.0 ± 0.0	48.4 ± 0.2	77.4 ± 0.1	63.2 ± 0.2
-0.0125	0.5	87.8 ± 0.1	50.3 ± 0.1	77.7 ± 0.1	63.4 ± 0.2
0.0125	0.5	87.8 ± 0.0	50.6 ± 0.1	77.8 ± 0.1	63.4 ± 0.2
0.025	0.5	87.8 ± 0.1	50.6 ± 0.2	77.7 ± 0.1	63.3 ± 0.1
0.05	0.5	87.7 ± 0.2	50.7 ± 0.3	77.6 ± 0.1	63.1 ± 0.1

However, shifting to representation learning settings, where the goals involve classification tasks and beyond, the interpretation cannot be translated word for word from imputation contexts. For example, frequently observed variables may provide reliable supervision and broad physiological coverage. They matter for downstream tasks and serve as reference points for reconstructing other features (Heilbroner et al., 2025), so upweighting rare features could trade away signal as much as it recovers it. Neither direction dominates in our ablation, which would fit such an account.

Input Ablation Table 5 provides information that replacing hourly information of vital signs and vasopressors to daily decreases performance on downstream predictions. Additionally, the table shows insights that including a zero-imputation on vasopressor did not yield a better performance, suggesting the model handles these absent values effectively through the masking mechanism alone.

6. Discussion

We presented AID-MAE, a dual-masked autoencoder that explicitly combines an intrinsic missingness mask with augmented masking to learn contextual embeddings directly from incomplete EHR tables. This design reduces the need for prior imputation, or unneces-

Table 5: Input ablation study results (mean ± standard deviation across 5 random seeds).

Input type	Mortality		Length of Stay	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**
Without 24-hour information	85.4 ± 0.1	45.7 ± 0.2	75.6 ± 0.1	60.4 ± 0.1
Zero-filling	87.2 ± 0.1	49.3 ± 0.2	77.3 ± 0.1	62.7 ± 0.1
AID-MAE	87.7 ± 0.1	49.8 ± 0.1	77.6 ± 0.1	63.1 ± 0.1

sary computation, while enabling capturing temporal-feature dynamics inherent in clinical signals. Empirically, AID-MAE achieves consistent gains over strong baselines on both MIMIC-IV and PhysioNet Challenge 2012 across three clinical prediction tasks. Our results highlight that learning directly from incomplete EHRs data provides a scalable path toward tabular foundation models in healthcare.

Limitations Despite these strengths, AID-MAE has limitations. Informative missingness is currently modeled implicitly. Explicitly capturing missing-not-at-random mechanisms remains an important direction for future work, even though the learned representations are encouraged to be invariant to the missingness distribution (Du et al., 2024). We also see two complementary future directions: incorporating structured medical or causal knowledge to better guide contextual dependency learning, and extending AID-MAE to multimodal inputs, such as clinical text and medical images.

Acknowledgments

The authors thank Guillaume Obozinski and Nicolas Boumal for their valuable feedback. DR was supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie COFUND grant agreement No 101127936 (DeMythif.AI). LAC is supported by the National Institutes of Health (DS-I Africa U54 TW012043, Bridge2AI OT2 OD032701), the National Science Foundation (ITEST 2148451), the Boston–Korea Innovative Research Project (RS-2024-00403047), and the Korea Health Technology R&D Project (RS-2024-00439677) through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea.

References

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

David Bellamy, Bhawesh Kumar, Cindy Wang, and Andrew Beam. Labrador: Exploring the limits of masked language modeling for laboratory data. In Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang, editors, *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*,

pages 104–129. PMLR, 15–16 Dec 2025. URL https://proceedings.mlr.press/v259/bellamy25a.html.

Manuel Burger, Fedor Sergeev, Malte Londschien, Daphné Chopard, Hugo Yèche, Eike Gerdes, Polina Leshetkina, Alexander Morgenroth, Zeynep Babür, Jasmina Bogojeska, Martin Faltys, Rita Kuznetsova, and Gunnar Rätsch. Towards foundation models for critical care time series, 2024. URL https://arxiv.org/abs/2411.16346.

Stef Buuren and Catharina Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45, 12 2011. doi: 10.18637/jss.v045.i03.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KI9NqjLVDT.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine*, 49(11):e1063–e1143, 2021.

Joseph Futoma, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O'brien. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In *Machine learning for healthcare conference*, pages 243–254. PMLR, 2017.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

Rolf HH Groenwold. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and prognostic research*, 4(1):8, 2020.

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*, pages 16000–16009, 2022.

Samuel P Heilbroner, Curtis Carter, David M Vidmar, Erik T Mueller, Martin C Stumpe, and Riccardo Miotto. A self-supervised framework for laboratory data imputation in electronic health records. *Communications Medicine*, 5(1):251, 2025.

Lars Hempel, Sina Sadeghi, and Toralf Kirsten. Prediction of intensive care unit length of stay in the mimic-iv dataset. *Applied Sciences*, 13(12):6930, 2023.

Katharine Henry, David Hager, Peter Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7:299ra122, 08 2015. doi: 10.1126/scitranslmed.aab3719.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Dominik Jarczak, Stefan Kluge, and Axel Nierhaus. Sepsis—pathophysiology and therapeutic concepts. *Frontiers in medicine*, 8:628302, 2021.

Jacob C. Jentzer, Saraschandra Vallabhajosyula, Ashish K. Khanna, Lakhmir S. Chawla, Laurence W. Busse, and Kianoush B. Kashani. Management of refractory vasodilatory shock. *Chest*, 154(2):416–426, 2018. ISSN 0012-3692. doi: https://doi.org/10.1016/j.chest.2017.12.021. URL https://www.sciencedirect.com/science/article/pii/S0012369218300722.

Alistair EW Johnson and Roger G Mark. Real-time mortality prediction in the intensive care unit. In *AMIA Annual Symposium Proceedings*, volume 2017, page 994, 2018.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Mathieu Jozwiak, Guillaume Geri, Driss Laghlam, Kevin Boussion, Charles Dolladille, and Lee S Nguyen. Vasopressors and risk of acute mesenteric ischemia: a worldwide pharmacovigilance analysis and comprehensive literature review. *Frontiers in Medicine*, 9: 826446, 2022.

Arif Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184, 08 2012. ISSN 1660-2110. doi: 10.1159/000339789. URL https://doi.org/10.1159/000339789.

Jungkyu Kim, Kibok Lee, and Taeyoung Park. To predict or not to predict? proportionally masked autoencoders for tabular data imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17886–17894, 2025.

Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. Duett: dual event time transformer for electronic health records. In *Machine Learning for Healthcare Conference*, pages 403–422. PMLR, 2023.

Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, 188(1):250–257, 2019.

Jiang Li, Xiaowei S Yan, Durgesh Chaudhary, Venkatesh Avula, Satish Mudiganti, Hannah Husby, Shima Shahjouei, Ardavan Afshar, Walter F Stewart, Mohammed Yeasin, et al. Imputation of missing values for electronic health record laboratory data. *NPJ digital medicine*, 4(1):147, 2021.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Kushal Majmundar, Sachin Goyal, Praneeth Netrapalli, and Prateek Jain. Met: Masked encoding for tabular data. *arXiv preprint arXiv:2206.08564*, 2022.

Qingqing Mao, Melissa Jay, Jana L Hoffman, Jacob Calvert, Christopher Barton, David Shimabukuro, Lisa Shieh, Uli Chettipally, Grant Fletcher, Yaniv Kerem, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. *BMJ open*, 8(1):e017833, 2018.

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

Michael Moor, Bastian Rieck, Max Horn, Catherine R. Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the icu using machine learning: A systematic review. *medRxiv*, 2020. doi: 10.1101/2020.08.31.20185207. URL https://www.medrxiv.org/content/early/2020/09/02/2020.08.31.20185207.

Nassim Oufattole, Hyewon Jeong, Matthew B.A. McDermott, Aparna Balagopalan, Bryan Jangeesingh, Marzyeh Ghassemi, and Collin Stultz. Event-based contrastive learning for medical time series. In Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary Lipton, Rajesh Ranganath, and Iñigo Urteaga, editors, *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR, 16–17 Aug 2024. URL https://proceedings.mlr.press/v252/oufattole24a.html.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

David Restrepo, Chenwei Wu, Yueran Jia, Jaden K. Sun, Jack Gallifant, Catherine G. Bielick, Yugang Jia, and Leo A. Celi. Representation learning of lab values via masked autoencoder, 2025. URL https://arxiv.org/abs/2501.02648.

Emanuel Rivers, Bryant Nguyen, Suzanne Havstad, Julie Ressler, Alexandria Muzzin, Bernhard Knoblich, Edward Peterson, and Michael Tomlanovich. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England journal of medicine*, 345 (19):1368–1377, 2001.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

Satya Narayan Shukla and Benjamin M Marlin. A survey on principles, models and methods for learning from irregularly sampled time series. *arXiv preprint arXiv:2012.00168*, 2020.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pages 245–248. IEEE, 2012.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Patrick D. Tyler, Hao Du, Mengling Feng, Ran Bai, Zenglin Xu, Gary L. Horowitz, David J. Stone, and Leo Anthony Celi. Assessment of intensive care unit laboratory values that differ from reference ranges and association with patient mortality and length of stay. *JAMA Network Open*, 1(7):e184521–e184521, 11 2018. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2018.4521. URL https://doi.org/10.1001/jamanetworkopen.2018.4521.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

Yu-Chang Yeh, Yu-Ting Kuo, Kuang-Cheng Kuo, Yi-Wei Cheng, Ding-Shan Liu, Feipei Lai, Lu-Cheng Kuo, Tai-Ju Lee, Wing-Sum Chan, Ching-Tang Chiu, et al. Early prediction of mortality upon intensive care unit admission. *BMC Medical Informatics and Decision Making*, 24(1):394, 2024.

Table 6: Downstream task performance on the MIMIC-IV ICU dataset. We compare our proposed model against established baselines across two clinical tasks. Results are reported as mean ± standard deviation across 5 random seeds. Our model (random masking 25%) achieves superior performance on both tasks, demonstrating the effectiveness of AID-MAE. Best results are shown in **bold**. AUROC and AUPRC are reported as percentages with one decimal place.

	Mortality		Length of Stay	
Model	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Logistic Regression	80.7 ± 0.0	37.3 ± 0.0	68.8 ± 0.0	53.1 ± 0.0
XGBoost	86.7 ± 0.0	47.5 ± 0.0	76.9 ± 0.0	62.2 ± 0.0
Supervised Transformer	83.9 ± 0.5	42.2 ± 1.3	72.3 ± 0.5	56.7 ± 1.5
DuETT	86.4 ± 0.2	47.3 ± 0.2	77.2 ± 0.0	62.8 ± 0.2
AID-MAE (ours)	**87.7 ± 0.1**	**49.8 ± 0.1**	**77.6 ± 0.1**	**63.1 ± 0.1**

Table 7: Downstream task performance on the PhysioNet 2012 Challenge dataset. We compare our proposed model against established baselines across two clinical tasks. Results are reported as mean ± standard deviation across 5 random seeds. Our model (random masking 25%) achieves superior performance on both tasks, demonstrating the effectiveness of AID-MAE. Best results are shown in **bold**. AUROC and AUPRC are reported as percentages with one decimal place.

	Mortality		Acute Kidney Injury	
Model	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Logistic Regression	72.7 ± 0.0	31.3 ± 0.0	74.8 ± 0.0	58.6 ± 0.0
Supervised Transformer	76.7 ± 1.4	34.1 ± 3.5	76.1 ± 0.9	59.6 ± 1.3
XGBoost	76.9 ± 0.0	36.9 ± 0.0	77.1 ± 0.0	61.8 ± 0.0
DuETT	77.7 ± 0.4	38.8 ± 0.9	76.7 ± 0.2	61.8 ± 0.5
AID-MAE (ours)	**78.2 ± 0.3**	**39.3 ± 1.7**	**77.3 ± 0.1**	**62.5 ± 1.2**

Appendix A. Additional Result

A.1. Complete results

Tables 6 and 7 show that our model achieves the best performance across both prediction tasks compared to all baselines. Figure 5 further demonstrates that these gains hold consistently for linear probing.

A.2. Pretraining on Imputed Data

Table 8 shows that pretraining AID-MAE on MICE-imputed PhysioNet Challenge 2012 data (Buuren and Groothuis-Oudshoorn, 2011; Silva et al., 2012) gives lower downstream AUROC than pretraining on the incomplete table directly. The masked objective recon-

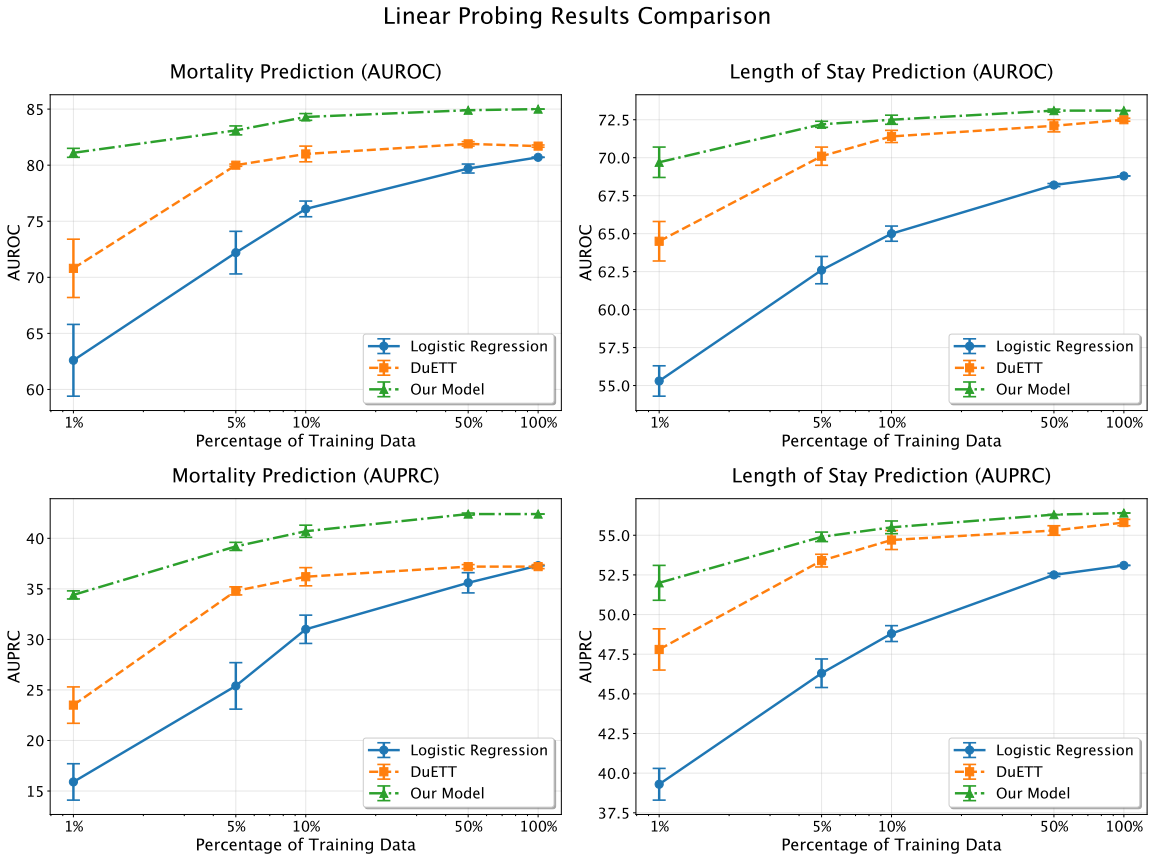

Figure 5: Linear probing results for mortality prediction and length of stay prediction tasks. We compare our model against Logistic Regression with median imputation and DuETT across different training data percentages (1%, 5%, 10%, 50%, 100%). Results are shown for both AUROC (top row) and AUPRC (bottom row) metrics. Our model consistently outperforms baseline methods across all data regimes and tasks, with particularly strong performance in low-data scenarios. Error bars represent standard deviation across 5 random seeds. The x-axis uses logarithmic scaling to better visualize performance across the range of training data percentages.

structs values, so imputing first makes the model recover statistical estimates rather than measurements a clinician took.

A.3. Single-Value Reconstruction Analysis

Imputation in EHR ranges from predicting one missing value, completing entire panels to filling the whole dataset. Here we first focus on single-value reconstruction to guide understanding on the pre-training quality. Single-value reconstruction tests whether the model can infer one measurement from the remaining context. As only one feature is

Table 8: Pretraining AID-MAE on MICE-imputed data (Buuren and Groothuis-Oudshoorn, 2011). Architecture, hyperparameters, and splits are held fixed. Mean ± SD over 5 seeds.

Setting	Mortality	AKI
With MICE	73.9 ± 1.2	75.8 ± 0.1
No Imputation	**78.2 ± 0.3**	**77.3 ± 0.1**

hidden at a time, the task is transparent and easy to interpret. This procedure relies purely on the self-supervised training signal of the masked autoencoder, without requiring extra task-specific heads.

We consider each input as a feature vector $\mathbf{x} \in \mathbb{R}^p$ together with a mask $\mathbf{m} \in \{0,1\}^p$, where $m_j = 1$ means feature j is observed and $m_j = 0$ means it is missing. The mask ensures that the model never uses values that are unavailable.

For evaluation, we hide one additional observed feature at a time. Let j be such a feature with $m_j = 1$. We set $m_j = 0$, feed the masked vector into the model, and obtain a reconstruction

$$\hat{\mathbf{x}} = f(\mathbf{x} \odot \mathbf{m}).$$

We therefore compute feature-wise evaluation metrics by comparing the predicted value $\hat{x}_j$ with the corresponding ground-truth observation x_j for each feature.

Our model has consistently outperformed XGBoost (Chen and Guestrin, 2016), one of the strongest baselines for single tabular data imputation, across standard error and goodness-of-fit criteria on MIMIC-IV. As shown in Table 9, our model achieves superior performance in 19 out of 20 most frequently taken clinical features across all metrics, demonstrating consistent advantages in normalized root mean squared error (NRMSE), normalized mean absolute error (NMAE), and coefficient of determination (R^2). This advantage is clinically meaningful, as accurate imputation of missing laboratory values is still critical for robust clinical prediction and unbiased model evaluation in many clinical scenarios.

Appendix B. Dataset Details

B.1. Dataset Details

The selected set of 61 features in MIMIC-IV in Table 10 provides complementary views on a patient's physiological state and disease progression.

For mortality prediction, vital signs, blood gas features, and vasopressor administration directly capture hemodynamic instability and organ dysfunction, while laboratory panels such as the Basic Metabolic Panel (BMP), Complete Blood Count (CBC), coagulation studies, Liver Function Tests (LFT), and cardiac enzyme assays reflect metabolic imbalance, infection response, and multi-organ failure. They are all central indicators of adverse ICU outcomes (Moor et al., 2020; Henry et al., 2015). For length-of-stay prediction, persistent abnormalities in electrolytes, hematological markers, and liver or kidney function are associated with slower recovery trajectories, while ongoing vasopressor dependence often signals prolonged ICU admission (Bellamy et al., 2025).

Feature	NRMSE		NMAE		R^2	
	Model	XGBoost	Model	XGBoost	Model	XGBoost
Glucose finger stick	**0.2030**	0.2033	**0.1516**	0.1523	**0.4806**	0.4787
Potassium (serum)	**0.1736**	0.2026	**0.1328**	0.1591	**0.5883**	0.4391
Sodium (serum)	**0.0676**	0.0846	**0.0416**	0.0626	**0.9362**	0.9000
Chloride (serum)	**0.0556**	0.0702	**0.0348**	0.0514	**0.9567**	0.9310
Hemoglobin	**0.0571**	0.0622	**0.0391**	0.0429	**0.9602**	0.9528
Hematocrit (serum)	**0.0636**	0.0662	**0.0429**	0.0452	**0.9495**	0.9452
HCO3 (serum)	**0.0621**	0.0837	**0.0390**	0.0628	**0.9467**	0.9032
Creatinine (serum)	**0.0812**	0.0915	**0.0458**	0.0522	**0.9108**	0.8868
Anion gap	**0.0859**	0.1126	**0.0544**	0.0857	**0.8982**	0.8250
BUN	**0.0862**	0.0943	**0.0550**	0.0610	**0.9077**	0.8897
Glucose (serum)	**0.1797**	0.1860	**0.1296**	0.1355	**0.5726**	0.5419
Magnesium	**0.1966**	0.2024	**0.1482**	0.1540	**0.4629**	0.4304
Phosphorous	**0.1508**	0.1583	**0.1150**	0.1217	**0.6778**	0.6449
Calcium non-ionized	**0.1549**	0.1644	**0.1164**	0.1243	**0.6769**	0.6360
Platelet Count	**0.1098**	0.1177	**0.0705**	0.0767	**0.8326**	0.8074
WBC	**0.1475**	0.1547	**0.1037**	0.1096	**0.7014**	0.6716
PH (Arterial)	**0.0588**	0.0616	**0.0316**	0.0341	**0.9526**	0.9479
Arterial O2 pressure	**0.1958**	0.2124	**0.1426**	0.1583	**0.4576**	0.3614
Arterial CO2 Pressure	**0.0433**	0.0500	**0.0235**	0.0276	**0.9717**	0.9622
Arterial Base Excess	**0.0304**	0.0305	0.0224	**0.0206**	**0.9859**	0.9858

Table 9: Performance comparison between our proposed model and XGBoost baseline on the top 20 frequently collected clinical features ranked by frequencies. Metrics include normalized root mean squared error (NRMSE), normalized mean absolute error (NMAE), and coefficient of determination (R^2). Normalization is based on the value range. Bold values indicate superior performance (lower NRMSE/NMAE or higher R^2). Our proposed model demonstrates consistent superiority across 19 out of 20 features across all metrics.

Additionally, in the broader context of developing clinical foundation models, it is essential to include as many heterogeneous features as possible rather than restricting analyses to a single subset, such as laboratory tests (Restrepo et al., 2025), hence our choice of feature selection. To remain compatible with MIMIC-IV, we retain only those PhysioNet 2012 Challenge features that have a direct counterpart in MIMIC-IV.

B.2. Preprocessing Details

Clinical time series are recorded at uneven times. A direct feed would give the model very long and very sparse sequences. Labach et al. (2023) show that discretising time series into fixed time bins with the last observed value yields strong performance across clinical prediction tasks. Their ablation study finds that carrying forward the most recent measurement, rather than using an average or interpolation, preserves sharp physiological signals and aligns with clinical intuition. Shukla and Marlin (2020) support this finding, highlighting that discretisation with simple carry-forward is both effective and widely used in practice.

Following this, we transform each ICU stay into a sequence of daily rows. Each row summarizes one calendar day of the patient's record. For laboratory values, we retain the

Table 10: Features Used in the Experiments

ID	Feature	ID	Feature
Vital Signs & Physiological Monitoring			
220045	Heart Rate	220050	Arterial Blood Pressure systolic
220051	Arterial Blood Pressure diastolic	220210	Respiratory Rate
220277	O_2 saturation (pulse oximetry)	223761 (223762)	Temperature (°F & °C)
Blood Gas Panel			
220224	Arterial pO_2	220227	Arterial O_2 saturation
220235	Arterial pCO_2	220274	Venous pH
223679	Venous TCO_2 (calc)	223830	Arterial pH
224828	Arterial Base Excess	225698	Arterial TCO_2 (calc)
226062	Venous pCO_2	226063	Venous pO_2
227073	Anion Gap	227443	Bicarbonate (HCO_3)
225668	Lactic Acid		
Basic Metabolic Panel (BMP)			
220602	Chloride	220615	Creatinine
220621	Glucose (serum)	220645	Sodium (serum)
225624	Blood Urea Nitrogen	225625	Calcium (serum)
225664	Glucose (finger-stick)	226534	Sodium (whole blood)
226537	Glucose (whole blood)	227442	Potassium (serum)
227464	Potassium (whole blood)	220635	Magnesium
225677	Phosphate		
Complete Blood Count (CBC)			
220228	Hemoglobin	220545	Hematocrit (serum)
226540	Hematocrit (calc)	220546	White Blood Cells
227457	Platelets		
CBC with Differential			
225639	Basophils (%)	225640	Eosinophils (%)
225641	Lymphocytes (%)	225642	Monocytes (%)
225643	Neutrophils (%)		
Coagulation Panel			
227465	Prothrombin Time	227466	Partial Thromboplastin Time
227467	INR	227468	Fibrinogen
Liver Function Test (LFT) Panel			
220587	Aspartate Aminotransferase (AST)	220644	Alanine Aminotransferase (ALT)
225612	Alkaline Phosphatase	225690	Total Bilirubin
Cardiac Enzymes Panel			
220632	Lactate Dehydrogenase	225634	Creatine Kinase (CK)
227445	CK-MB isoenzyme	227429	Troponin-T
227456	Albumin		
Vasopressor Medications			
221289	Epinephrine (mcg/min)	221906	Norepinephrine (mcg/min)
221662	Dopamine (mcg/min)	221749	Phenylephrine (mcg/min)
222315	Vasopressin (units/min)		

last recorded result of each test within that day. This design mirrors standard clinical workflows, where the most recent lab panel is typically used for decision making. Within each row, we also include an hourly representation of the five vital signs, oxygen saturation, and five vasopressors infusion rates. These are recorded at hourly resolution by selecting the last observed value up to that hour. We additionally include a reference value for each lab, which corresponds to the most recent previous corresponding lab result that is recorded prior to the day. This aims to align with real-life clinical setting and provide a longitudinal baseline

Each numeric entry is paired with a time stamp that encodes its recency. We express this as the number of hours before midnight, rounded to one decimal place. For example, a value recorded at 21:00 today is encoded as 3.0, while a reference lab value taken at 20:30 the day before is encoded as 27.5. For vasopressors that are carried forward across hours, we assign each event a time stamp corresponding to the midpoint of the hour in which it was forwarded. We first winsorize each numeric feature at the 5th and 95th percentiles to reduce extreme outliers (Wilcox, 2011), then apply a per-feature min–max normalization. If the values are not recorded, those values will be represented as a missing entry.

We intentionally replace dopamine with norepinephrine equivalent dose since simple arithmetic on the features in the resulted table can recover the dopamine rate. Let Epi, NE, and Phen be the infusion rates of epinephrine, norepinephrine, and phenylephrine in $\mu g\,kg^{-1}\,min^{-1}$, and let Dop and Vas be the rates of dopamine and vasopressin in the same units and in $U\,min^{-1}$ respectively. Following standard practice (Jentzer et al., 2018), we compute

$$\mathrm{NE}_{\mathrm{eq}} = \mathrm{NE} + \mathrm{Epi} + \frac{\mathrm{Dop}}{150} + \frac{\mathrm{Phen}}{10} + 2.5\,\mathrm{Vas}.$$

Lastly, rows that do not contain any laboratory measurements are discarded, as they convey thin physiological signal (Tyler et al., 2018). This filter removes approximately seven percent of input samples in MIMIC-IV and prevents the model from over-fitting patterns driven purely by sparsity.

We split our curated dataset into training set and test set based on their time stamps. Admission prior to the year 2179 constitute the training set, and after as the test set. We note that year 2179 is a de-identified number by MIMIC-IV (Johnson et al., 2023), effectively yielding a reproducible random split. We further split 20% of the training set as the validation set. Disjoint subject_ids are ensured across the train, validation and test set.

Appendix C. Architectural Details

This section presents the comprehensive architectural and training specifications for our masked autoencoder framework, covering pretraining, linear probing, and fine-tuning.

C.1. Pretraining Architecture and Configuration

Pretraining Architecture The masked autoencoder follows a Vision Transformer-inspired encoder-decoder architecture adapted for tabular data. The encoder consists of a configurable number of transformer blocks with the following specifications:

- **Embedding dimension**: $d_{embed} = 64$
- **Encoder depth**: $L_{enc} = 8$ transformer blocks
- **Number of attention heads**: $h = 8$ (proportional to embedding dimension)
- **MLP ratio**: $r_{mlp} = 4.0$ (hidden dimension $= 4 \times d_{embed}$)
- **Decoder embedding dimension**: $d_{dec} = 64$ (matches encoder)
- **Decoder depth**: $L_{dec} = 4$ transformer blocks
- **Decoder attention heads**: $h_{dec} = 4$

The model employs a specialized CombinedEmbed module that processes value-time pairs by projecting each component to the full embedding dimension and combining them additively, effectively reducing the sequence length by half while preserving temporal information.

Training configuration The pretraining uses AdamW (Loshchilov and Hutter, 2019) optimizer with a base learning rate of $lr_{base} = 1 \times 10^{-3}$ and weight decay of $\lambda = 0.05$. The learning rate follows a cosine annealing schedule (Loshchilov and Hutter, 2016) with 20-40 warmup epochs, decaying to a minimum learning rate of $lr_{min} = 1 \times 10^{-5}$.

Training Dynamics:

- **Batch size**: $B = 64$ samples per batch with gradient accumulation support
- **Maximum epochs**: $E_{max} = 400$
- **Loss function**: Mean squared error

C.2. Linear Probing Methodology

Linear probing evaluation extracts frozen representations from the pretrained encoder to assess learned feature quality. The methodology involves:

Feature Extraction: CLS token embeddings ($d_{embed} = 64$-dimensional) are extracted from the pretrained encoder without any fine-tuning of the encoder parameters.

Classifier Configuration: Scikit-learn's LogisticRegression with liblinear solver, L2 regularization ($C = 1.0$).

Evaluation Protocol:

- **Data fractions**: We use $f \in \{1\%, 5\%, 10\%, 50\%, 100\%\}$ of the available training data.
- **Seeds**: 5 independent runs (2020-2024) for statistical robustness
- **Metrics**: AUROC and AUPRC for binary classification tasks
- **Tasks**: Mortality and 72-hour length-of-stay prediction for MIMIC-IV, In-hospital mortality and Acute Kidney Injury for PhysioNet 2012 Challenge.

C.3. Fine-tuning Architecture and Training

The fine-tuning approach utilizes a task-specific classification head appended to the pre-trained encoder:

Classification Head: A configurable feedforward network with the following default configuration:

- **Input dimension**: $d_{in} = 64$ (matching encoder embedding dimension)
- **Hidden layers**: 2-layer MLP architecture
- **Hidden dimension**: $d_{hidden} = 32$ neurons per hidden layer
- **Dropout rate**: $p_{drop} = 0.1$
- **Output dimension**: $d_{out} = 1$ (binary classification with sigmoid activation)

Fine-tuning Training Configuration

- **Encoder learning rate**: $lr_{enc} = 1 \times 10^{-5}$
- **Classification head learning rate**: $lr_{cls} = 1 \times 10^{-3}$

Optimizer and Regularization:

- **Optimizer**: AdamW with weight decay of $\lambda = 1 \times 10^{-5}$
- **Batch size**: $B = 128$ samples per batch
- **Maximum epochs**: $E_{max} = 100$ with early stopping

Training Protocol:

- **Early stopping**: Patience of 10 epochs based on validation AUROC
- **Validation strategy**: Stratified train/validation/test splits of 64%, 16% and 20%

The training framework allows systematic evaluation of self-supervised pretraining effectiveness across different adaptation strategies, from linear probing (with no parameter updates) to full fine-tuning (end-to-end optimization).

Appendix D. Experiment Details

D.1. Baseline Details

D.1.1. Logistic regression

As a linear baseline, we trained logistic regression models with both ℓ_1 and ℓ_2 penalties. Prior to training, we applied median imputation to all continuous variables to address missing values, ensuring that imputation was performed once on the training set and applied consistently to the test set. Model selection was carried out using a grid search over the regularization strength $C \in \{0.1, 1.0, 10.0\}$ and penalty type $\{l_1, l_2\}$. Optimization used the `liblinear` solver, with a maximum of 200 iterations and a convergence tolerance of 10^{-3}. To reduce the impact of random variation, we repeated experiments with five random seeds (2020–2024) and report the mean and standard deviation of AUROC and AUPRC on the test set.

D.1.2. XGBoost

We additionally benchmarked against XGBoost (Chen and Guestrin, 2016), a widely used gradient boosted decision tree method that has been shown to perform competitively on tabular and clinical time series data. The model was tuned with a grid search over {learning rate $\in \{0.05, 0.1, 0.2\}$, max depth $\in \{4, 5, 6\}$, number of estimators $\in \{500, 1000\}$}, and hyperparameters were selected via validation AUROC. For evaluation, the best model was applied to the test set to compute AUROC and AUPRC.

D.1.3. MICE Imputation

We used scikit-learn's IterativeImputer with BayesianRidge as the default estimator, 10 imputation rounds, and a fixed random seed. Imputation was fit on the training set and applied to validation and test sets. The imputed data was then used to train AID-MAE with identical architecture and hyperparameters, isolating the effect of imputation on the training signal.

D.1.4. DuETT: Dual Event Time Transformer

DuETT (Labach et al., 2023) extends the Transformer architecture to clinical time series by explicitly modelling three fundamental dimensions of electronic health record (EHR) data. First, temporal dependencies are captured by attention over discretized time bins, enabling the model to learn disease trajectories despite irregular sampling. Second, event-type relationships are captured by attention across heterogeneous clinical variables, allowing the model to integrate information from diverse measurements such as vitals and laboratory values. Third, DuETT leverages the presence or absence of observations as a predictive signal, by jointly reconstructing masked event values and missingness indicators during self-supervised pre-training. This design exploits the fact that not measuring a variable in itself conveys clinical intent.

By alternating attention across time and event axes, and by integrating missingness as a training signal, DuETT learns robust patient representations that outperform state-of-the-art baselines in both full fine-tuning and limited-label regimes. One limit that we currently address is that DuETT does not directly attend different features across time, losing information on temporal interactions among features.

We follow the training protocol of DuETT, which consists of a self-supervised pre-training phase followed by supervised adaptation.

Pre-training Following (Labach et al., 2023), we conducted self-supervised pre-training for 300 epochs using the AdamW optimizer. The learning rate was scheduled with linear warmup followed by inverse square-root decay, and gradient clipping was applied at 1.0 to stabilize training. At each iteration, one event type and one time bin were masked and the model was trained to reconstruct both event values and their presence. Inputs were normalized to zero mean and unit variance, with outliers clipped at three median absolute deviations from the median, and time-bin aggregation used the last observed value. Dropout was applied in attention and feedforward layers as in the original implementation.

Fine-tuning All encoder and classification head parameters were updated jointly using a single AdamW optimizer, without differential learning rates. This choice follows the official

DuETT implementation. In practice, this design is stabilized by the combination of linear warmup and inverse square-root decay, rather than by assigning separate learning rates to encoder and head. This is feasible since DuETT's representation token is of comparatively high dimension (approximately 1.5k), allowing sufficient capacity in the head to adapt during supervised training. We fine-tuned for the same number of epochs as reported in (Labach et al., 2023) (30 epochs on MIMIC-IV data), ensuring comparable training budgets. To further enhance robustness, we adopted the procedure of the paper, averaging the weights of the five best-performing checkpoints (ranked by validation AUROC) to obtain the final model. This weight-averaging strategy was shown to improve stability over single-checkpoint selection. Importantly, to rule out undertraining as a confound, we additionally performed a systematic learning rate sweep over $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ for both fine-tuning and linear probing, and report the best results across seeds.

Linear probing To directly assess representation quality, we froze the encoder and trained only a linear classification head (a single fully connected layer without hidden units). This corresponds to logistic regression applied to the fixed patient representation produced by the DuETT encoder. Since the encoder output dimension in DuETT is large, linear probing provides a stringent test of the quality of pre-trained representations without relying on capacity in the task head.