Learning Representations from Incomplete EHR Data with Dual-Masked Autoencoding

Xiao Xiang[1,2] **David Restrepo**[1] **Hyewon Jeong**[1] **Yugang Jia**[1] **Leo Anthony Celi**[1,3,4*]
[1]Massachusetts Institute of Technology
[2]École Polytechnique Fédérale de Lausanne (EPFL)
[3]Harvard University
[4]Beth Israel Deaconess Medical Center

Abstract

Learning from electronic health records (EHRs) time series is challenging due to irregular sampling, heterogeneous missingness, and the resulting sparsity of observations. Prior self-supervised methods either impute before learning, represent missingness through a dedicated input signal, or optimize solely for imputation, reducing their capacity to efficiently learn representations that support clinical downstream tasks. We propose the Augmented-Intrinsic Dual-Masked Autoencoder (AID-MAE), which learns directly from incomplete time series by applying an intrinsic missing mask to represent naturally missing values and an augmented mask that hides a subset of observed values for reconstruction during training. AID-MAE processes only the unmasked subset of tokens and consistently outperforms strong baselines, including XGBoost and DuETT, across multiple clinical tasks on two datasets. In addition, the learned embeddings naturally stratify patient cohorts in the representation space.

1 Introduction

Electronic Health Records (EHRs) contain irregularly-sampled time series that exhibit missingness with diverse, feature-specific patterns [Li et al., 2021]. Existing methods introduced different data representations 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]. Nevertheless, a tabular structure offers key advantages: it aligns features with time, enabling learning contextual embeddings that capture temporal–feature dependencies [Huang et al., 2020] and, ultimately, effectively reflecting underlying patient states.

In this work, we introduce the Augmented-Intrinsic Dual-Masked AutoEncoder (AID-MAE), a self-supervised framework trained directly on incomplete EHR tables. Our contributions are:

- We present, to the best of our knowledge, one of the first dual-mask mechanisms for EHRs representation learning, which applies two complementary masking strategies: one that represents the intrinsic missingness already present in EHR tables (i.e., naturally unobserved values), and

*Correspondence to: lceli@mit.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

another that introduces augmented stochastic masking during training. By combining these, AID-MAE jointly uses real and augmented missingness, enabling effective representation learning without explicit imputation or additional missingness-specific input features.
- We show that AID-MAE outperforms both tree-based and state-of-the-art self-supervised baselines across multiple downstream prediction tasks, and achieves notable gains in low-label linear probing. Through experiments on a separate dataset, we further analyze how AID-MAE is effective to transfer and generalize within the masked modeling family, and we highlight a masking strategy that is specific to the dual-masking design.
- We offer qualitative analyses illustrative of the learned representations, showing the learned patient embeddings are stratified by physiological state. Additionally, the contextual feature embeddings are organized in the embedding space consistent to clinical knowledge.

2 Related Work

Masked Autoencoding for Incomplete Tabular Data: Masked autoencoders (MAE) [He et al., 2022] have been adapted to incomplete tabular data [Majmundar et al., 2022, Du et al., 2024, Kim et al., 2025]. ReMasker [Du et al., 2024] adapts MAE for self-supervised imputations on data with simulated missingness. PMAE [Kim et al., 2025] further applies proportional masking to address the imbalance in mask sampling propensity, which leads to improved reconstruction accuracies. Lab-MAE [Restrepo et al., 2025], building on this line of work, focuses on reconstructing lab values in EHRs. However, all works share final objective for imputation quality and do not yield general-purpose embeddings. Xu et al. [2025] introduced a pre-trained model on incomplete wearable time series for representation learning. Nevertheless, this remains in regularly-sampled time series data that does not readily translate to EHR Tables.

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, but failed to outperform tree-based methods [Chen and Guestrin, 2016]. DuETT [Labach et al., 2023] alternates attention across time and events to capture dependencies in a matrix input, but relies on zero imputation and adds a missingness token that 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.

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:

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

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:

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

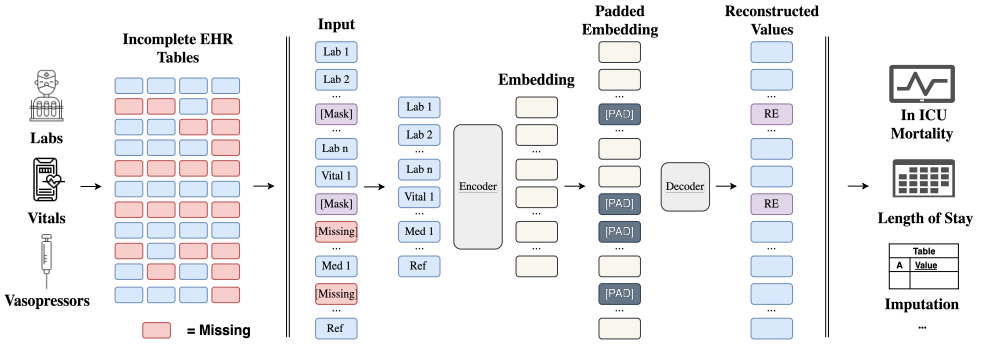


Figure 1: **AID-MAE (Augmented-Intrinsic Dual-Masked AutoEncoder)** A subset of observed tokens of the original data with inherent missingness `MISSING` is randomly masked (`[MASK]`). Each measurement (value and timestamp) is embedded with positional encodings. The encoder processes only unmasked tokens, 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.

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.

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, allowing the model to memorize feature positions in the input.

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 augmented masks, as in Figure 1.

Training Objective: The model is pretrained by optimizing dual reconstruction loss $\mathcal{L}$, 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]:

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

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),

Table 1: 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**

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 include all ICU patient stays, discarding only input arrays without laboratory measurements. 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. A complete list of features and preprocessing details is provided in Table 4 and Appendix B.

Task Definitions: For downstream evaluation, we construct one sample per patient by restricting inputs to the first 24 h of the ICU admission following common benchmarks [Purushotham et al., 2018, Johnson and Mark, 2018, Hempel et al., 2023, Yeh et al., 2024]. We evaluate the model's capacities through several downstream clinical tasks: in-ICU mortality and length-of-stay prediction for MIMIC-IV, and in-ICU mortality and Acute Kidney Injury (AKI) for PhysioNet Challenge 2012. Specifically, in-ICU Mortality is a binary indicator of death before ICU discharge during the admission. Length of Stay (LOS) is a binary indicator that ICU length of stay is strictly less than 72 hours from admission. The definition of Acute Kidney Injury (AKI) follows Khwaja [2012] using creatinine. Task distributions are provided in Table 6.

4.2 Results:

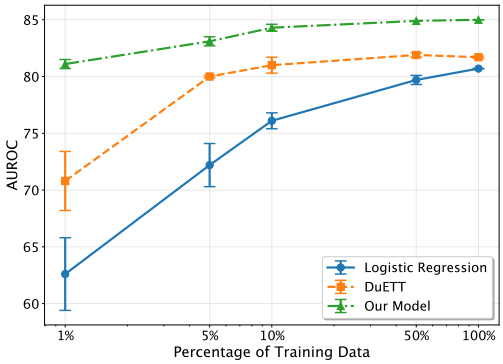
AID-MAE Outperforms All Baselines on Fine-Tuning Tasks: Table 1 shows that our model outperforms a range of baselines consistently on all tasks evaluated. This includes XGboost [Chen and Guestrin, 2016], as well as state-of-the-art masked modeling method [Labach et al., 2023]. We note that the supervised 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. Additionally, when pre-training on the curated MIMIC-IV dataset, AID-MAE has less than a million trainable parameters compared to 5.2 million in DuETT, ablating gains due to model size. We report fine-tuning details of our model and the baseline models in Appendices C and D.2, and the complete result in Table 2.

Linear Probing Shows Pretrained Embeddings Transfer Strongly Across Data Regimes: Figure 2(a) 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 3.

4.3 Towards EHR Foundation Models

Different EHR datasets exhibit significant distribution shifts [Burger et al., 2024]. We envision an effective pre-trained EHR model should transfer across datasets and avoid unnecessary computations:

Transfer Learning: Our results in Fig. 2(b) show that AID-MAE transfers well from MIMIC-IV to PhysioNet Challenge 2012, indicating both strong generalizability and practical deployment potentials. Prior masked models, such as DuETT, rely on a less flexible design: features that are absent in the target dataset must still be carried through the model as explicit missing tokens to



Linear Probing (Transfer Learning)		
Method	Mortality	AKI
Random Weights	49.8	52.7
DuETT	70.9	66.2
AID-MAE (ours)	**73.1**	**69.7**

Fine-Tuning (Transfer Learning)		
Method	Mortality	AKI
DuETT	77.8 ± 0.3	76.4 ± 0.1
AID-MAE (ours)	$\mathbf{78.6 \pm 0.2}$	$\mathbf{77.0 \pm 0.2}$

Figure 2: (a) Linear probing performance across label fractions on MIMIC-IV. (b) Transfer learning performance from MIMIC-IV pre-trained weights to PhysioNet 2012. We report AUROC for mortality and acute kidney injury (AKI) prediction under linear probing (top) and full fine-tuning (bottom).

preserve compatibility. This requirement introduces unnecessary computation in the encoder that complicates both transfer learning and scaling that requires merging multiple healthcare datasets.

Proportional Reweighting: Driven by the differences in patient populations, institutional care practices, and measurement frequency, missingness distributions vary significantly in different healthcare centers. We tested 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 thus map missing frequencies into weights by using a logit transform, a schedule that has the desired convex–concave curvature [Kim et al., 2025]:

$$w_j = a\, \log \frac{p_{\text{miss},j}}{1 - p_{\text{miss},j}} + b \;\text{ if } p_{\text{miss},j} < 1, \quad 0 \;\text{ if } p_{\text{miss},j} = 1.$$

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.

Our empirical results in Table 7 shows improvement in fine-tuning performances on MIMIC-IV. We provide further discussion regarding the design and the empirical results in the Appendix E.

5 Conclusion

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 unnecessary 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.

Despite its strengths, our model has several limitations. First, it currently captures informative missingness only implicitly. A potential direction for future work is to model missing-not-at-random patterns in EHR data explicitly, even though the model already encourages learning representations 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 and Disclosure of Funding

The authors thank Guillaume Obozinski and Nicolas Boumal for their valuable feedback. 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

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.

Anthony Bisulco, Rahul Ramesh, Randall Balestriero, and Pratik Chaudhari. From linearity to non-linearity: How masked autoencoders capture spatial correlations. *arXiv preprint arXiv:2508.15404*, 2025.

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.

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.

Ayush Dalmia and Suzanna Sia. Clustering with UMAP: why and how connectivity matters. *CoRR*, abs/2108.05525, 2021. URL https://arxiv.org/abs/2108.05525.

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.

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.

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.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings, 2020. URL https://arxiv.org/abs/2012.06678.

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.

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.

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.

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 Seaman and Ian White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22:278–295, 05 2011. doi: 10.1177/0962280210395740.

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.

Appendix

A Additional Result

A.1 Complete results

Table 2 and 3 show that our model achieves the best performance across both prediction tasks compared to all baselines. Figure 3 further demonstrates that these gains hold consistently for linear probing.

A.2 ICU admission Subtyping

We assessed this by extracting the CLS embeddings Devlin et al. [2019] when pre-training MIMIC-IV, which is the latent representation of the special [CLS] token summarizing the entire input array, derived from the first 24 hours of ICU stay. The CLS embeddings successfully differentiate patients admitted to the medical intensive care unit (MICU) and the cardiothoracic vascular intensive care unit (CVICU) into distinct latent subgroups. This is accomplished by applying k-means clustering ($k = 2$) to the original embedding space.

Figure 5 has revealed two well-separated clusters in the UMAP space [Dalmia and Sia, 2021], where each cluster is predominantly composed of patients from one unit. The high homogeneity of

Table 2: Downstream task performance on 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.

	In-ICU 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
Our model	**87.7 ± 0.1**	**49.8 ± 0.1**	**77.6 ± 0.1**	**63.1 ± 0.1**

Table 3: Downstream task performance on 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	77.0 ± 0.0	61.8 ± 0.0
DuETT	77.7 ± 0.4	38.8 ± 0.9	76.7 ± 0.2	61.8 ± 0.5
SMART (New Baseline)	77.8 ± 0.6	38.5 ± 1.2	76.9 ± 0.5	60.8 ± 0.5
AID-MAE (ours)	**78.2 ± 0.3**	**39.3 ± 1.7**	**77.3 ± 0.1**	**62.5 ± 1.2**

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 across k=2 to k=10 clusters. Figure 4 in the Appendix shows k=2 as the optimal, with a silhouette score of 0.42, indicating moderate but meaningful separation between patient subtypes.

A.3 Feature Embedding Analysis

To visualize whether the learned feature encoder organizes laboratory measurements into coherent regions of representation space, for each laboratory feature, we extracted a 64-dimensional embedding when training MIMIC-IV for every valid measurement instance.

For visualization, we drew a uniform random sample of $N = 100{,}000$ points from the test data. We computed a two-dimensional UMAP Dalmia and Sia [2021] embedding directly from the raw 64-dimensional vectors using the Euclidean metric. This visualization is enhanced by a large min_dist to render a well-separated islands when such structure is present.

In Figure 6, each dot is a single measurement's 64-D embedding projected to two dimensions, with a predefined subset of clinically important lab types colored, and all remaining types are shown in gray (Other Labs). We denote two interpretable patterns: (i) in the bottom left, the pink and green islands correspond to Platelet Count and White Blood Cell (WBC) counts, respectively both hematology cell-count measures drawn from the CBC panel, and they appear as neighbors; (ii) in the bottom-right, Hemoglobin and Hematocrit reside as a tight pair of neighbors in the embedding space, consistent with their strong correlation in clinical practice. UMAP is label-agnostic [Dalmia and Sia, 2021], and we do not reveal lab identities to the algorithm. Thus, the emergence of these islands from the geometry of the 64-D embeddings is strong evidence that the encoder captures medically grounded

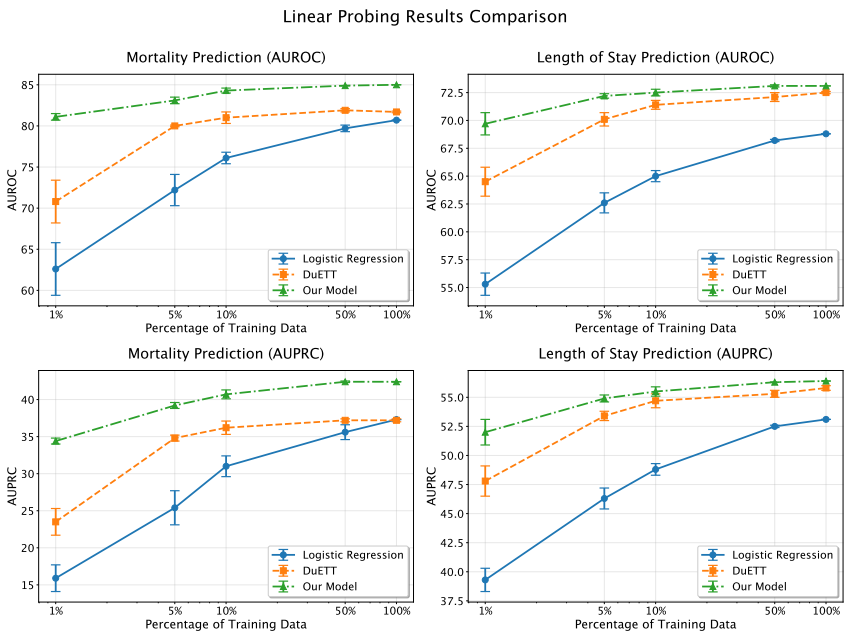


Figure 3: Linear probing results for mortality prediction and length of stay prediction tasks. We compare our model against Logistic Regression with median imputation and DuETT across different training data percentages (1%, 5%, 10%, 50%, 100%). Results are shown for both AUROC (top row) and AUPRC (bottom 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.

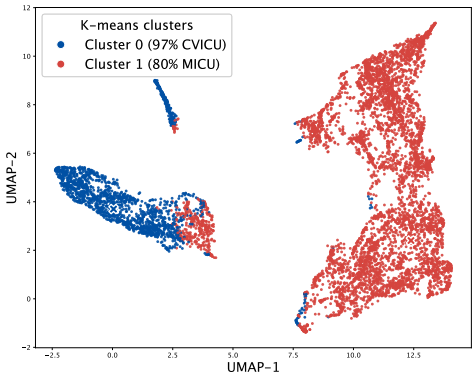


Figure 4: UMAP visualization of first-day CLS embeddings for initial MICU and CVICU admissions. Colors represent clusters from K-means ($k = 2$) applied in the embedding space.

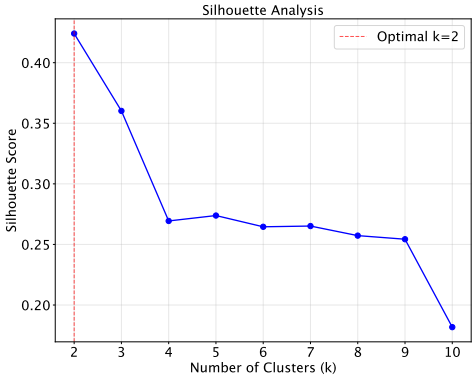


Figure 5: Silhouette analysis over $k \in \{2, \ldots, 10\}$. The mean silhouette score peaks at $k = 2$ ($s = 0.42$) and declines thereafter, indicating two moderately well-separated clusters.

structure. Similarly, clinically sensible groupings are visible for other lab families, strengthening this conclusion.

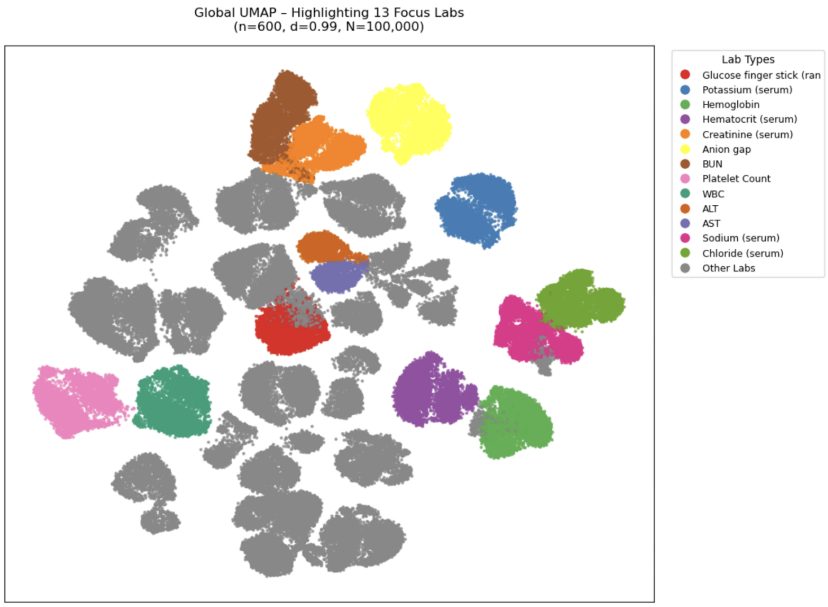


Figure 6: **UMAP of feature embeddings.** UMAP projection of $N = 100{,}000$ randomly sampled 64-D embeddings for 50 lab features. Colors denote 13 highlighted lab types. Each point corresponds to one measurement embedding. We denote two important patterns: **Bottom-left:** neighboring pair of pink (Platelet 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.

B Dataset Details

B.1 Dataset Details

The selected set of 61 features in MIMIC-IV in Table 4 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].

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.

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)		

Table 4: Features Used in The Experiments (by Clinical Panel)

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 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 according 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 are will be represented as a missing entry.

We intentionally replace dopamine with norepinephrine equivalent dose. Let Epi, NE, and Phen be the infusion rates of epinephrine, norepinephrine, and phenylephrine in $\mu\mathrm{g\,kg^{-1}\,min^{-1}}$, and let Dop and Vas be the rates of dopamine and vasopressin in the same units and in $\mathrm{U\,min^{-1}}$ respectively. Following standard practice [Jentzer et al., 2018], we compute

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

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 an 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.

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**: In-hospital 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 pretrained 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).

D Experiment Details

D.1 Task Distribution

Table 6 presents the downstream prediction tasks evaluated on our ICU cohort of 92,938 stays. The mortality prediction task exhibits class imbalance with only 7.6% prevalence, while the length of stay prediction (LOS<72h) shows a more balanced distribution at 66.1% prevalence.

Table 5: Downstream tasks and cohort sizes for MIMIC-IV. Counts aggregate train, validation, test splits. Prevalence is the overall fraction of positive labels.

ICU cohort (first 24 h of ICU admission)			
Task	**# Stays**	**# Events**	**# Prevalence**
Mortality	92,938	7,046	7.6%
LOS< 72h	92,938	61,476	66.1%

Table 6: Downstream tasks and cohort sizes for PhysioNet 2012 Challenge. Counts aggregate train, validation, and test splits. Prevalence reflects the fraction of positive labels. A small fraction of AKI labels could not be retrieved due to data availability.

ICU cohort (first 24 h of ICU admission)			
Task	**# Patients**	**# Events**	**Prevalence**
Mortality	11,981	1,709	14.3%
AKI	11,811	3,456	29.3%

D.2 Baseline Details

D.2.1 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.2.2 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.2.3 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.

E Further Motivation: Non-uniform Augmented Mask

Unlike datasets with simulated missingness, which most mask autoencoding methods for tabular data are built on [Du et al., 2024][Kim et al., 2025], EHR time series are collected opportunistically as a by-product of care. The probability that a value is recorded depends on latent clinical state (e.g., concern about instability) and workflow (e.g., staff availability, device uptime). Consequently, missingness is rarely Missing Completely At Random (MCAR) or even Missing At Random (MAR). In most clinical setting, it is Missing Not At Random (MNAR), which means the chance of observing a value depends on the possibly unobserved value itself or on the patient's condition. Overlooking inherent missingness pattern in EHR can bias estimation and degrade predictive performance.

In our input data, for example, serum sodium is missing in roughly 7–8% of entries per patient-day, whereas arterial oxygen saturation is absent in more than 88% of cases. Figure 7 illustrates the wide dispersion of missingness and highlights that each channel follows its own observation regime.

Moreover, when we train on combined datasets that exhibit substantial distribution shifts, these heterogeneous observation regimes may further distort learning. Features that are frequently measured in one dataset but rarely observed in another may be overweighted or underweighted. Uniform

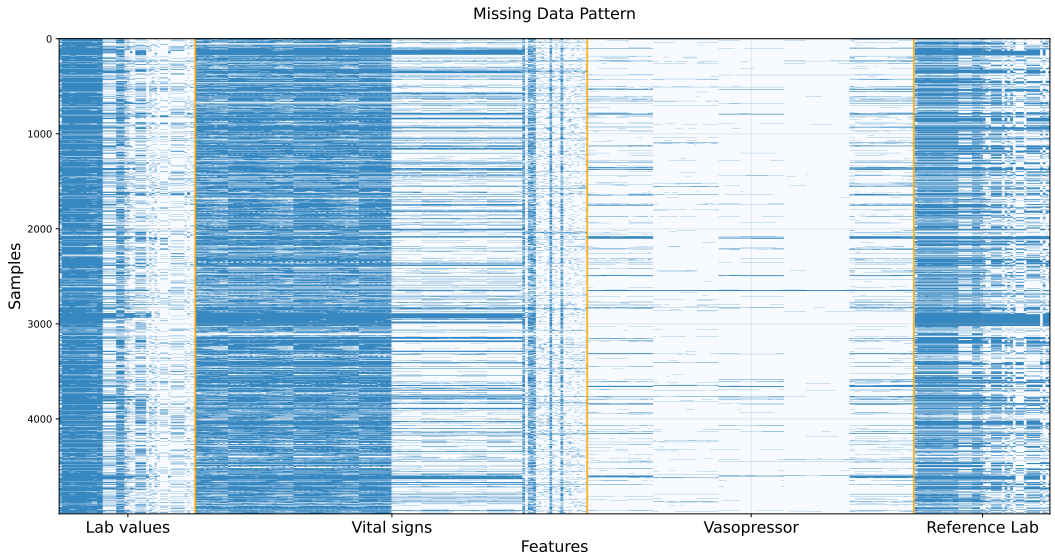


Figure 7: Missingness pattern across features in the MIMIC-IV ICU dataset. Each row corresponds to a patient sample and each column to a clinical feature, grouped by lab values, vital signs, vasopressors, and reference labs. Blue indicates observed measurements while white denotes missing entries, highlighting systematic sparsity and irregular sampling common in ICU data.

masking risks degrading performance, underscoring the need for non-uniform augmented masks that account for feature-specific missingness.

To address this, we advance to a scheme that estimates feature-specific missing rate within each batch and maps them into augmented masking weights, balancing the contributions of dense and sparse features during training [Kim et al., 2025].

Feature-Wise Missing Rate. Given a batch $\mathcal{B}$ with missingness indicator matrix $\mathbf{M} \in \{0,1\}^{|\mathcal{B}| \times \ell_{\text{keep}}}$, we compute the empirical missing rate for feature j:

$$p_{\text{miss},j} \;=\; 1 - p_{\text{obs},j} \;=\; \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathbf{M}.$$

Logit-Based Reweighting. We use the feature-wise masking weight w_j defined in Section 4.3, where the parameter a determines the resampling direction: if $a > 0$, features with low observation probability $p_{\text{obs},j}$ receive larger weights, corresponding to an upsampling of sparse features; conversely, $a < 0$, corresponds to undersampling. The offset b serves as a regularizer to avoid extreme weights when $p_{\text{obs},j}$ approaches zero or one.

Previously, the logit-based 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 more augmented mask and get predicted, ensuring that their contribution to the gradient signal is not washed out by frequent variables [Rosenbaum and Rubin, 1983, Seaman and White, 2011]. This has been proved useful for imputation benchmarks with synthetic missingness [Kim et al., 2025].

However, shifting to representation learning settings, where the goals involve classification tasks and beyond, the interpretation cannot not be translated word for word from imputation contexts. For example, frequently observed variables may provide reliable supervision and broad physiological coverage, making them both clinically important and useful reference points for reconstructing other features [Bisulco et al., 2025]. We thus reserve our choice to consider both the case when $a < 0$ and $a > 0$ and leave a more detailed discussion of this point to future work.

a	b	Mortality		Length of Stay	
		ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
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

Table 7: Ablation Study: Effect of Proportional Masking Hyperparameters

Input Type	Mortality		Length of Stay	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
Without 24 hours 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

Table 8: Input Ablation Study Results

F Complete Ablation Results

We present comprehensive ablations in Tables 7 and 8. First, varying the proportional masking hyperparameters (a, b), including the random-masking case a=0 and fixed $b \in 0.25, 0.5$, shows stable or improved AUROC/AUPRC for both mortality and length-of-stay prediction, indicating robustness to our masking schemes. Second, input ablations compare two input variants; zero-filling means we zero-impute non-recorded vasopressors, and without 24 hours information means we use daily vital signs and vasopressor information, with AID-MAE yielding the strongest overall metrics.