Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai

Article

Stable CDE Autoencoders with Acuity Regularization for Offline Reinforcement Learning in Sepsis Treatment

Yue Gao

Keebo AI, Toronto, ON M5G, Canada; yue@keebo.ai

How To Cite: Gao, Y. Stable CDE Autoencoders with Acuity Regularization for Offline Reinforcement Learning in Sepsis Treatment. *Transactions on Artificial Intelligence* **2025**, *1*(1), 307–325. https://doi.org/10.53941/tai.2025.100021

Received: 13 October 2025
Revised: 13 November 2025
Accepted: 18 November 2025
Published: 12 December 2025

Abstract: Effective reinforcement learning (RL) for sepsis treatment depends on learning stable, clinically meaningful state representations from irregular ICU time series. While previous works have explored representation learning for this task, the critical challenge of training instability in sequential representations and its detrimental impact on policy performance has been overlooked. This work demonstrates that Controlled Differential Equations (CDE) state representation can achieve strong RL policies when two key factors are met: (1) ensuring training stability through early stopping or stabilization methods, and (2) enforcing acuity-aware representations by correlation regularization with clinical scores (SOFA, SAPS-II, OASIS). Experiments on the MIMIC-III sepsis cohort reveal that stable CDE autoencoder produces representations strongly correlated with acuity scores and enables RL policies with superior performance (WIS return > 0.9). In contrast, unstable CDE representation leads to degraded representations and policy failure (WIS return ~ 0). Visualizations of the latent space show that stable CDEs not only separate survivor and non-survivor trajectories but also reveal clear acuity score gradients, whereas unstable training fails to capture either pattern. These findings highlight practical guidelines for using CDEs to encode irregular medical time series in clinical RL, emphasizing the need for training stability in sequential representation learning.

Keywords: reinforcement learning; sepsis treatment; neural controlled differential equations; irregular time series; state representation; training stability

1. Introduction

Sequential decision-making is a cornerstone of modern healthcare, particularly in dynamic clinical scenarios, where timely diagnosis and adaptive treatment strategies are essential for patient survival [1,2]. A representative case is sepsis management, which requires timely diagnosis and appropriate treatment strategies. Reinforcement learning (RL) offers a promising framework for modeling sequential decision-making in clinical settings, where treatment strategies must adapt over time to a patient's evolving condition. In the context of sepsis management, RL has been employed to derive policies that optimize interventions including fluid resuscitation and vasopressor administration. For instance, the *AI Clinician* [3], demonstrated RL's potential by training on the Medical Information Mart for Intensive Care III (MIMIC-III) dataset to recommend treatment strategies that outperformed clinician baselines. Subsequent advances in deep and distributional RL further refined policy robustness and flexibility, underscoring RL's role in sepsis management [4–6]. These advancements underscore the potential of RL to generate effective sequential treatment decisions in critical care environments.

In this work, we use MIMIC-III [7–9] as experimental data, which containts clinical data from 40,000+ ICU patients, with vital signs, lab results, and treatments recorded at irregular time intervals. It provides a rich foundation for developing data-driven sepsis management tool. However, a key challenge lies in constructing informative state representations from this noisy, irregularly sampled data. Prior work has explored recurrent neural networks (RNNs) and autoencoders for this task, but their instability during training often leads to suboptimal representations and degraded policy performance [2,4,6].

Neural Controlled Differential Equations (Neural CDEs) are state-of-the-art models for irregular time series data due to their ability to model continuous-time dynamics and handle irregular sampling [4,10]. Unlike discrete-step

RNNs, CDEs use differential equations to propagate hidden states, capturing latent physiological trends more accurately. This makes them particularly suited for sepsis management, where patient states evolve smoothly between sparse observations. However, despite their potential, Neural CDEs remain underexplored in clinical RL due to severe training instability and lack of interpretability, which this work addresses explicitly. CDEs are prone to training instability such as gradient explosion or collapse when unregularized or trained for excessive epochs [11]. This instability stems fundamentally from numerical challenges in solving the underlying differential equations [12], particularly when backpropagating through adaptive ODE solvers [13].

A key challenge in applying Neural CDEs to clinical time series is distinguishing between numerical instability and model overfitting, which are two fundamentally different failure modes. Overfitting appears when the training loss continues to decrease while the validation loss starts going upward; In contrast, CDEs exhibit unique instability patterns where both losses increase simultaneously (Figure 1). This instability occurs when numerical solvers fail to handle stiffness in the learned vector fields [12]. For clinical time series modeling, this instability is particularly crucial, as trajectory smoothness and sudden acuity transitions must both be captured.

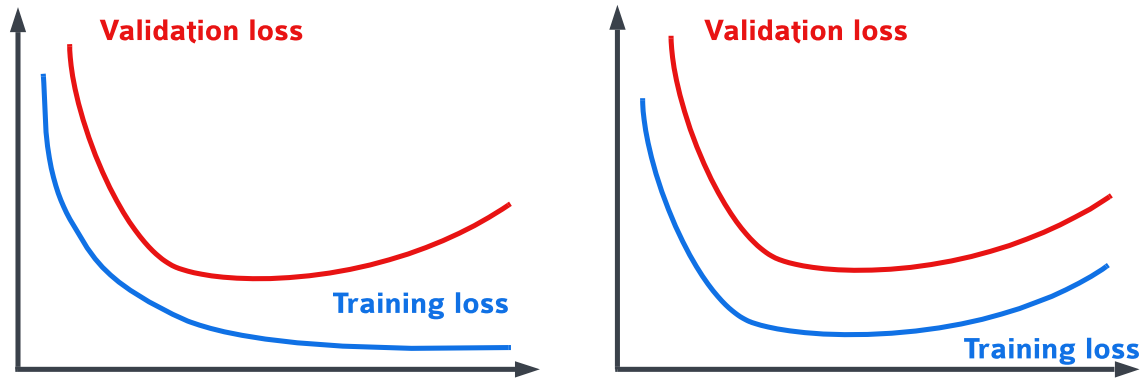

Figure 1. Distinguishing training failure modes: (**Left**) Classical overfitting; (**Right**) Numerical instability.

This limitation in CDE state representation was overlooked in prior studies. For instance, ref. [4] compared CDEs against other autoencoders for sepsis state representation and showed its superiority, but did not address their instability, leading to inconsistent results and an erroneous dismissal of clinical acuity score regularization.

In this work, we use the MIMIC-III database to train, evaluate, and compare CDE representations for sequential state encoding, with the goal of improving discrete Batch-Constrained Q-learning (dBCQ) models for offline sepsis treatment decision-making, with the following key contributions:

(1) Disentangle and mitigate numerical instability in Neural CDEs for clinical reinforcement learning. We systematically identify instability as a distinct failure mode from overfitting and propose stabilization strategies—including early stopping and solver regularization—that reliably improve representation robustness.
(2) Revisit the role of clinical acuity scores as effective regularizers for CDE state representations. Through stabilized training regimes, we show that SOFA/SAPS-II/OASIS scores provide meaningful physiological constraints, resolving prior inconsistencies regarding their effectiveness in clinical RL.
(3) Provide a unified evaluation of CDE stabilization techniques for clinical time-series modeling. We benchmark gradient clipping, implicit solvers, and stiffness regularization under a consistent framework, offering practical guidance for selecting methods that optimize downstream dBCQ policy performance.
(4) Ensure reproducibility and adoption in clinical ML workflows. We release all code, training configurations, and evaluation protocols for transparent and reproducible experimentation (Code available at https://github.com/GAOYUEtianc/RL_mimic_CDE_stable).

To our knowledge, this is the first work that systematically disentangles numerical instability from representation overfitting in Neural CDEs applied to clinical RL. This distinction provides new insight into why prior RL policies trained on irregular medical time series often failed unpredictably. Our findings revisit and clarify inconsistencies reported in prior works such as [4], which concluded that acuity-based regularization fails to improve CDE representations. We show this claim likely arose from unstable training regimes that obscure representation quality. By addressing CDE instability, we show that acuity regularization improves both state representation quality and RL performance.

An earlier preprint of this work is available on arXiv [14].

2. Background

2.1. MIMIC-III Data Irregular Properties

The MIMIC-III repository contains ICU patient data including physiological signals, laboratory results, medications, and interventions as multivariate time series [7]. However, these measurements were recorded at irregular intervals based on clinical needs, resulting in uneven sampling patterns and substantial missing data [8, 15]. The dataset contains inherent measurement noise and informative missing patterns, where missing observations could provide clinical insights. This creates a low signal-to-noise ratio that makes direct trajectory modeling challenging [16, 17]. Different variables are recorded at varying frequencies, causing covariate shift and temporal aliasing that reduce model generalization [18]. Simple imputation techniques like forward-filling or mean substitution fail to capture complex temporal dependencies and often introduce bias [17].

To mitigate these challenges, researchers have developed several approaches. Time-aware recurrent models use masking and time interval embeddings to explicitly handle missingness and irregularity [16]. Continuous-time methods such as ODE-RNNs and latent ODEs learn differential equations that naturally adapt to varying time intervals [15]. Among these, Neural Controlled Differential Equations offer a SOTA approach to encode MIMIC-III's high signal-to-noise trajectories. They produce well-representated states suited for downstream offline reinforcement learning in sepsis management [4, 15, 19].

2.2. Neural Controlled Differential Equations (Neural CDEs) as State Encoders

Neural Controlled Differential Equations are a class of continuous-time models that generalize recurrent neural networks by defining hidden state dynamics through differential equations [19]. We initialize the hidden state as $h(t_0) = Wo(t_0) + b$, using a learnable linear map $g : \mathbb{R}^d \to \mathbb{R}^h$ implemented as a single-layer neural network. Then for the continuous-time irregular observations $o_t \in \mathcal{O}$, CDE encodes a hidden state $h(t) \to \mathcal{H}$ in differential form via

$$\partial h(t) = f_\theta(h(t))\partial o_t \tag{1}$$

where f_θ is a neural network parameterizing the system dynamics, and the differential ∂o_t accounts for irregular sampling intervals [15, 20]. The CDE acts as an encoder $\psi : \mathcal{O} \to \mathcal{H}$, mapping time-based observations to a latent space. A decoder $\phi : \mathcal{H} \to \mathcal{O}$ reconstructing subsequent observations

$$\hat{o}_t = \phi(h(t)) \tag{2}$$

is trained to minimize the loss $\mathcal{L}_{\text{MSE}}(o_t, \hat{o}_t) = ||o_t - \hat{o}_t||^2$. This autoencoding framework ensures $h(t)$ retains clinically relevant information while discarding noise [21]. In this work, we use 4-th order Runge-Kutta (RK4) as the numerical solver for the CDE as a baseline [22].

2.3. Instability in Neural CDE Training

During training, Neural CDEs can exhibit numerical instability due to the sensitivity of ODE solvers to the learned vector field f_θ. If f_θ has large Lipschitz constant or long time sequences, the backpropagating gradients can explode, leading to erratic updates and divergence [12, 23]. Hence, picking a proper stabilization method that can balance the smoothness and sensitivity to rare high-acuity transitions is crucial.

Effective methodologies to stabilize Neural CDEs training include implicit solvers that handle stiffness by solving linearized equations at each step [13, 20, 24, 25], gradient clipping to prevent large updates [26], and regularization techniques smoothing the learned vector field by penalizing high local error and stiffness [27]. Building on these insights, we employ early stopping and stabilization techniques to ensure Neural CDEs training produces high-quality state representations that generalize well and enable effective downstream RL performance.

2.4. Reinforcement Learning with CDE State Representations

Sequential sepsis treatment can be modeled as a Partially Observable Markov Decision Process (POMDP):

- State : $s_t = h(t)$ (CDE encoded history of observations $o_{0:t}$).
- Action : $a_t \in \{1, ..., 25\}$ represents discrete combinations of intravenous fluids and vasopressor doses [3]
- Reward :

$$r_t = \begin{cases} +1 & \text{if patient survives at trajectory end} \\ -1 & \text{if patient dies at trajectory end} \\ 0 & \text{otherwise (at intermediate steps)} \end{cases} \tag{3}$$

- Policy: $\pi(a_t|s_t)$ maps states to probability distribution of actions, aiming to optimize cumulative reward via offline RL.

2.5. Clinical Acuity Scores as Priors

We leverage three established clinical acuity scores (SOFA, SAPS-II, and OASIS) as semi-supervision to regularize the CDE latent space. We define the correlation loss:

$$\mathcal{L}_{\text{corr}}(\hat{s}_t) = -(\rho_{\text{SOFA}}(\hat{s}_t) + \rho_{\text{SAPS-II}}(\hat{s}_t) + \rho_{\text{OASIS}}(\hat{s}_t)) \tag{4}$$

The regularized total loss function is defined to be:

$$\mathcal{L}_{\text{total}}(o_t, \hat{o}_t) = \mathcal{L}_{\text{MSE}}(o_t, \hat{o}_t) + \lambda \cdot \mathcal{L}_{\text{corr}}(\hat{s}_t) \tag{5}$$

where $\rho(\hat{s}_t)$ denotes the Pearson correlation between the latent state representation $\hat{s}_t$ and acuity score. For simplicity, we denote the losses at epoch i as $\mathcal{L}_{\text{MSE}}(i)$, $\mathcal{L}_{\text{corr}}(i)$, $\mathcal{L}_{\text{total}}(i)$.

3. Methodology

We present a framework as shown in Figure 2 for sepsis treatment policy learning using stabilized CDE state representations, evaluated through offline RL and clinical interpretability metrics.

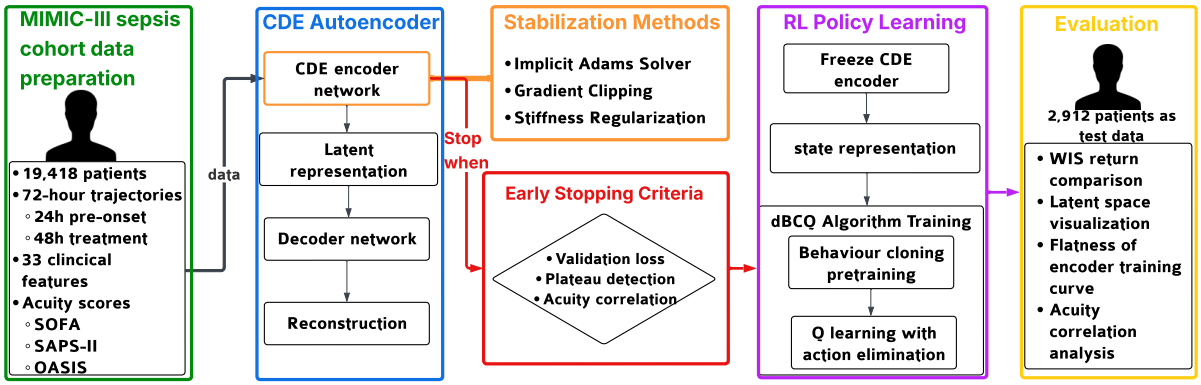

Figure 2. Overall architecture of the proposed framework.

3.1. Overall Architecture

Our framework builds on prior work in sepsis treatment RL [4–6] but addresses a critical gap: the training instability of Neural CDEs for state representation. While existing approaches use CDEs to encode patient history into hidden states and decode future observations, they overlook how training instability affects both representation quality and downstream policy performance. We identify two key instability symptoms:

- Unpredictable fluctuations in observation prediction loss
- Weakened correlation between learned states and clinical acuity scores

To address these, we introduce:

- Training stabilization via early stopping and specialized techniques
- Acuity-aware regularization to maintain clinical relevance

The stabilized representations then feed into Batch Constrained Q-learning, where we evaluate their impact on final policy performance through both quantitative metrics and clinical interpretability measures.

3.2. Data Preparation

Our study utilizes the MIMIC-III v1.4 critical care database, processed according to the established sepsis cohort definition from prior work [3]. After processing, the dataset comprises 19,418 adult sepsis patients, with each patient trajectory spanning a clinically relevant 72-h window around sepsis onset : capturing 24 h preceding identification through 48 h of subsequent treatment. These trajectories reflect the real-world challenges of ICU care, exhibiting irregular sampling intervals and heterogeneous measurements across 33 time-varying physiological features. The action space follows prior work in discretizing clinical interventions into 25 distinct combinations of intravenous fluids and vasopressor doses, binned by percentile ranges to maintain clinically

meaningful groupings while enabling reinforcement learning. Patient outcomes define trajectory termination, with mortality recorded for the 9.2% of patients who died within 48 h of their final observation, while survivors comprise the remaining 90.8%. To ensure faithful evaluation while preserving outcome distributions, we split the cohort into training (70%, n = 13,593), validation (15%, n = 2913), and test sets (15%, n = 2912), maintaining identical 9.2% mortality rates across splits. Clinical acuity scores (SOFA, SAPS-II, OASIS) for illness severity are calculated at each timestep using validated implementations that transform the 33 raw features into standardized risk metrics.

The original extracted data contains 48 variables including demograpics, elixhauser status, laboratory results, vital signs, fluids and vasopressors, and fluid balance. There are missing and irregulaly sampled data in original MIMIC-III dataset, following the method by [3], 19,418 adult sepsis patients were selected from the MIMIC-III database by applying the criteria:

- Patients are aged 18 or older;
- An acute increase in the Sequential Organ Failure Assessment (SOFA) score of 2 or more;
- Exclude admissions where treatment was withdrawn or mortality was not documented.

Then the data was filled using a time-limited approach based on clinically relevant periods for each variable, then further imputed using nearest neighbour algorithm. We then select 33 features (Table 1) that are most relevant to sepsis treatment.

Table 1. List of 33 time-varying continuous physiological features used for state representation training.

Feature 1–8	Feature 9–16	Feature 17–24	Feature 25–33
Glasgow Coma Scale	Potassium	White Blood Cells	PaO_2/FiO_2
Heart Rate	Sodium	Platelets	Bicarbonate (HCO_3)
Systolic BP	Chloride	PTT	SpO_2
Diastolic BP	Glucose	PT	BUN
Mean BP	INR	Arterial pH	Creatinine
Respiratory Rate	Magnesium	Lactate	SGOT
Body Temp (°C)	Calcium	PaO_2	SGPT
FiO_2	Hemoglobin	$PaCO_2$	Bilirubin
			Base Excess

For reinforcement learning in this task, actions are defined as combinations of intravenous (IV) fluids and vasopressors, discretized into 25 clinically meaningful bins based on percentile ranges. As shown in Table 2, these bins span from no administration to higher dose quartiles, forming a 5×5 action space, comprising 25 distinct treatment actions. Vasopressor doses are converted to noradrenaline-equivalents (mcg/kg/min), and IV fluids are normalized for tonicity before discretization.

Table 2. Discretized dosage bins for vasopressors and intravenous (IV) fluids used to define the 25-action reinforcement learning space.

Action Number	0	1	2	3	4
Vasopressors	0.00	$(0.00, 0.08]$	$(0.08, 0.22]$	$(0.22, 0.45]$	>0.45
IV fluids	0.00	$(0.00, 50.00]$	$(50.00, 180.00]$	$(180.00, 530.00]$	>530.00

4. Acuity Score Alignment

Patient acuity scores quantify illness severity and play a central role in clinical decision-making. To constrain the learning of state representations, we extract three acuity scores from the full patient observations from each 4 hour time step (in MIMIC-III dataset):

- SOFA (Sequential Organ Failure Assessment) [28]: Assesses dysfunction across respiratory, coagulation, liver, cardiovascular, CNS, and renal systems. Scores range from 0 (normal) to 24 (severe failure).
- SAPS-II (Simplified Acute Physiology Score) [29]: Predicts ICU mortality using 17 physiological measurements. Scores range from 0 to 163, with higher values indicating greater mortality risk.
- OASIS (Oxford Acute Severity of Illness Score) [30]: Estimates mortality risk from 10 clinical variables. Scores range from 10 to 83, where higher scores correspond to worse prognosis.

These scores regularize the CDE latent space through Pearson correlation between the learned state representations and their corresponding acuity scores. Equation (6) defines the acuity correlation regularized loss, where we choose the hyperparameter λ to be the same for all three acuity scores, yet these could be chosen independently of one another and yield a loss function :

$$\begin{aligned} &\mathcal{L}_{\text{total}}(o_t, \hat{o}_t) = \mathcal{L}_{\text{MSE}}(o_t, \hat{o}_t) \\ &- (\lambda_1 \cdot \rho_{\text{SOFA}}(\hat{s}_t) + \lambda_2 \cdot \rho_{\text{OASIS}}(\hat{s}_t) + \lambda_3 \cdot \rho_{\text{SAPSII}}(\hat{s}_t)) \end{aligned} \tag{6}$$

4.1. CDE Autoencoder Training with Rigorous Stopping Criteria

The CDE autoencoder is trained to learn clinically meaningful state representations through a carefully designed optimization process that addresses the inherent instability of continuous-time neural networks.
Continuous-Time Encoding Process Our CDE autoencoder learns continuous-time state representations through a neural controlled differential equation (Equation (1)), implemented as a 3-layer network with ReLU activations and layer normalization. The encoder outputs a hidden state $h(t) \in \mathbb{R}^d$ where d is the representation dimension.
Early Stopping Strategy To prevent training instability while preserving signal capture, we propose a multi-criteria early stopping method that selects the optimal checkpoint epoch e^* when all conditions are first met:

(1) Near-optimal validation loss:

$$\mathcal{L}_{\text{val}}(e^*) \leq \min_{e \leq e^*} \mathcal{L}_{\text{val}}(e) + \epsilon_1 \tag{7}$$

(2) Stable training plateau: For the last p epochs, the total loss variation remains within ϵ_2 fraction of the minimum loss:

$$\begin{aligned} &\left| \max_{e^*-p \leq i \leq e^*} \mathcal{L}_{\text{total}}(i) - \min_{e^*-p \leq i \leq e^*} \mathcal{L}_{\text{total}}(i) \right| \leq \\ &\epsilon_2 \cdot \min_{e^*-p \leq i \leq e^*} \mathcal{L}_{\text{total}}(i) \end{aligned} \tag{8}$$

(3) Clinically meaningful representations over training: Mean acuity score correlation on train set exceeds a threshold.

$$\rho(e^*) \geq \rho_{\text{threshold}} \tag{9}$$

We train CDE autoencoders across multiple random seeds through comprehensive hyperparameter tuning (learning rates, hidden sizes, ϵ_1, ϵ_2, p, $\rho_{\text{threshold}}$), then select the best performing configuration evaluated via the ultimate RL policy measured by WIS return (Weighted Importance Sampling (WIS) is a technique used in offline reinforcement learning to estimate the expected return of a target policy using data collected from a different behavior policy. By normalizing importance weights across trajectories, WIS reduces variance compared to ordinary importance sampling, albeit introducing some bias. This trade-off often results in more stable and reliable policy evaluations [31]). This acuity-aligned objective enforces clinically meaningful structure in the latent representations, which is later shown to improve both interpretability and reinforcement learning outcomes.

4.2. Stabilization Methods for CDE Training

To improve training stability and facilitate model selection, we apply stabilization techniques to the CDE encoder, specifically on the vector field f_θ that governs hidden state dynamics. We implement three stabilization techniques:

(1) Gradient Clipping: Constrains extreme gradient updates to prevent sudden spikes during backpropagation
(2) Implicit Adams Solver: Uses an implicit numerical integration scheme to handle stiffness in the continuous-time dynamics
(3) Stiffness Regularization: Directly penalizes high curvature in the learned vector field

Appendix A.2 introduces those methodologies in detail.
Evaluation Protocol: We conduct comprehensive hyperparameter tuning for hyperparameters in Table 3 across

- Multiple random seeds : 25, 53, 1234, 2020
- Learning rates : 1×10^{-4}, 2×10^{-4}, 5×10^{-4}
- Hidden dimensions : 4, 8, 16, 32, 64, 128.

Each configuration undergoes 100 epochs training until reaching early stopping criteria in Section 4.1, and is evaluated using two stability metrics:

(1) Plateau Length: Number of consecutive epochs where the training loss remains on a plateau:

$$|\mathcal{L}_{\text{total}}(i) - \min \mathcal{L}_{\text{total}}| \leq \epsilon_2 \cdot \min \mathcal{L}_{\text{total}} \tag{10}$$

(2) Mean Absolute Slope on Plateau:

$$S_1 = \frac{1}{T} \sum_{i}^{T} |\Delta \mathcal{L}_{\text{total}}(i)| \tag{11}$$

where $\Delta \mathcal{L}_{\text{total}}(i) = \mathcal{L}_{\text{total}}(i+1) - \mathcal{L}_{\text{total}}(i)$ and T is the plateau length.

Table 3. Stabilization methods and hyperparameters to be tuned.

Method	Hyperparameter	Search Range
Gradient clipping	max norm : τ	$\{0.1, 0.5, 1.0, 1.5\}$
Implicit Adams solver	step size : Δt	$\{1/8, 1/4, 1/2\}$
Stiffness regularization	λ_{reg}	$\{0.005, 0.01, 0.015\}$

The final model evaluation considers both these stability metrics and the downstream RL policy performance (WIS return).

4.3. dBCQ Policy Learning and Evaluation

We employ the trained CDE encoder to transform raw patient trajectories into continuous state representations for offline reinforcement learning. Our policy learning approach builds on discrete Batch-Constrained Q-learning (dBCQ) [4], which addresses key challenges in offline RL. The process is as follows:

- Behavior Cloning Pre-training: We first train a policy to replicate the action distribution in the dataset, providing a conservative starting point.
- Constrained Q-Learning: During policy optimization, the Q-function only considers actions that the behavior policy would likely take with probability $\geq \tau_{\text{BC}}$.
- Policy Evaluation: We compute WIS returns of the trained Q-policy on the validation dataset to assess policy performance.

5. Experiments and Empirical Results

5.1. Effect of CDE Early Stopping on Final RL Policy Result

To determine optimal training configurations, we conducted extensive hyperparameter tuning to identify optimal training configurations. A grid search was performed over the hidden state sizes $\{4, 8, 16, 32, 64, 128\}$, learning rates $\{1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$, acuity correlation coefficients $\lambda \in \{0, 0.5, 1, 1.5\}$, and early stopping criteria parameters $\epsilon_1 \in \{0.05, 0.1, 0.15, 0.2\}$, $p \in \{20, 30, 40, 50\}$, $\epsilon_2 \in \{0.02, 0.03, 0.04, 0.05\}$, and $\rho_{\text{threshold}} \in \{0.6, 0.65, 0.7, 0.75\}$. The best configuration was selected based on downstream RL policy performance, measured by the weighted importance sampling (WIS) return.

Table 4 summarizes the best hyperparameters for the early stopping criteria in Section 4.1. These values were chosen through grid search to achieve the most stable training dynamics and highest validation performance. Using these optimal parameters, the CDE autoencoder was trained for 100 epochs while recording the total loss $\mathcal{L}_{\text{total}}$ (defined in Equation (5)) at each step.

Figure 3 illustrates the training dynamics, showing the total loss, MSE loss, and correlation loss across epochs (mean $\pm$ one standard deviation over multiple random seeds).

Two representative checkpoints are compared to highlight the importance of early stopping:

(1) *Optimal and stable checkpoint:* Meets all stopping criteria in Section 4.1, achieving low and consistent losses with strong acuity correlation.

(2) *Overtrained and unstable checkpoint:* Continues training beyond the plateau, showing diverging losses and reduced generalization.

To assess how early stopping affects reinforcement learning, we froze the CDE encoder parameters at each checkpoint and trained two dBCQ policies under identical conditions (policy learning rate $= 1 \times 10^{-5}$, action elimination threshold $\tau_{\text{BC}} = 0.3$, and training epochs $= 2 \times 10^5$). The resulting WIS returns on the validation trajectories are shown in Figure 4. The policy initialized from the *optimal and stable* representation achieved

consistently high returns (final WIS $= 0.9195$). In contrast, the policy trained on the *overtrained and unstable* representation failed to learn meaningful value estimates, with WIS collapsing to 3.2×10^{-7}.

Table 4. Criteria for selecting optimal stopping epoch in CDE autoencoder training with hyperparameters to be tuned.

Criterion	Metric	Hyperparameter	Best Hyperparameter
Low Validation Loss	Final validation loss should reach a minimal value during training	ϵ_1	0.1
Plateau	Number of consecutive epochs with small relative loss change	p ϵ_2	30 0.02
Acuity Correlation	Mean Pearson correlation between learned features and acuity scores	$\rho_{\text{threshold}}$	0.7

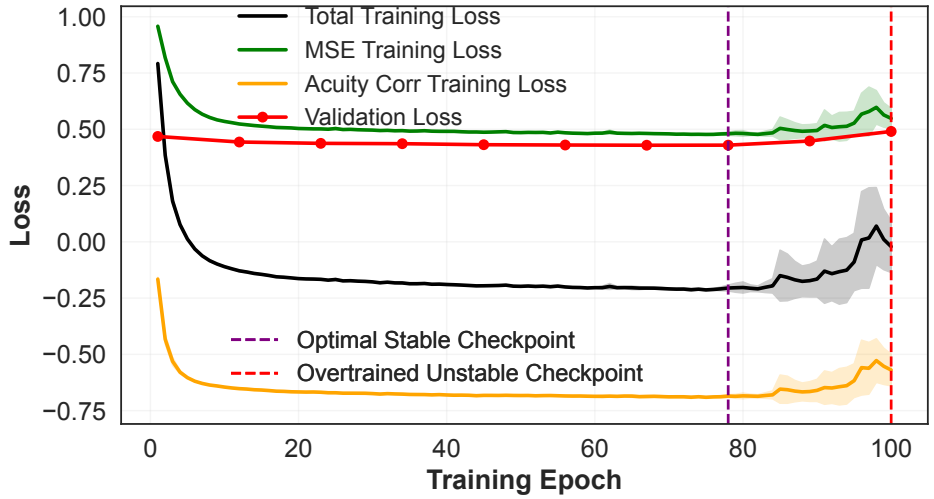

Figure 3. Training curves of CDE autoencoder losses (mean ± std across seeds) versus epoch. The *optimal and stable* checkpoint achieves smooth convergence, whereas the *overtrained and unstable* model shows divergence and instability.

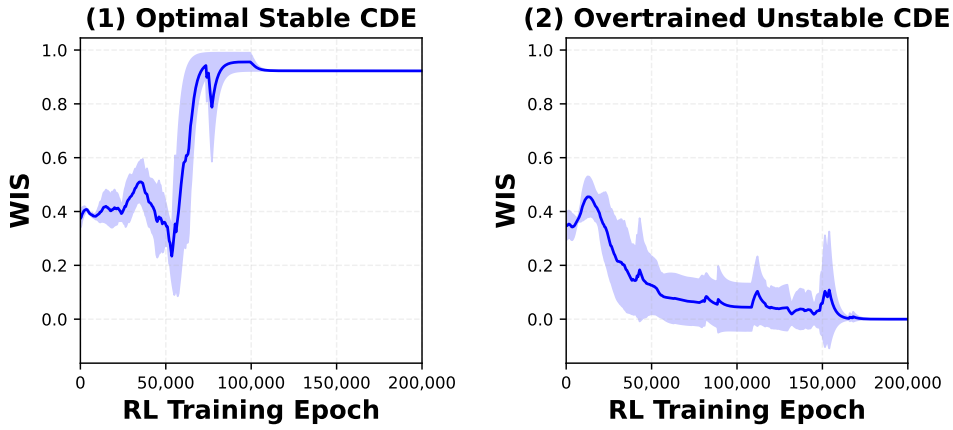

Figure 4. Evaluated WIS on validation set ($mean \pm std$ across random seeds) for dBCQ policies trained using *optimal and stable* and *overtrained and unstable* CDE state representations.

These results clearly demonstrate that selecting the correct stopping epoch is crucial for learning stable and generalizable representations from the CDE autoencoder. Our proposed multi-criteria early stopping strategy not only stabilizes training but also significantly enhances downstream RL policy performance. The best configuration achieved WIS > 0.9, outperforming prior works such as [4,32], which reported approximately 0.775. This confirms that training stability, rather than architecture alone, is a critical determinant of effective clinical RL models.

5.2. Clinical Alignment Through Early Stopping

To further understand how early stopping improves the quality and interpretability of learned representations, we analyze the clinical alignment of the CDE latent space and its relation to training stability. Specifically, we

compare latent spaces from the *optimal and stable* checkpoint and the *overtrained and unstable* checkpoint, both trained with the best hyperparameters identified in Section 5.1. For each model, latent representations from the validation trajectories (first and last observations) are projected into two dimensions using principal component analysis (PCA).

Figure 5 highlights three complementary views of how early stopping affects the learned representations:

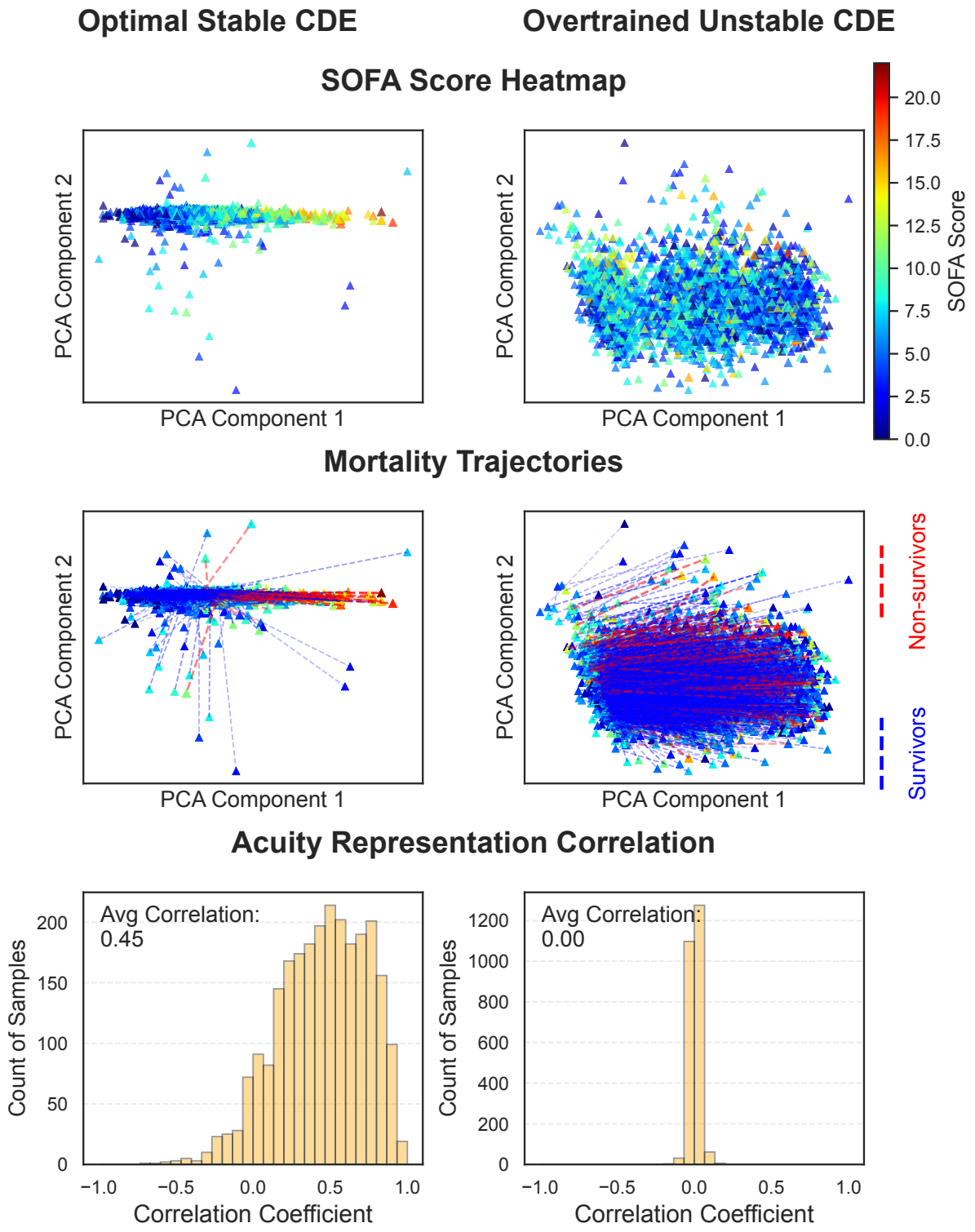

Figure 5. Comparative visualization of clinical alignment of *optimal and stable* and *overtrained and unstable* CDE representations on validation dataset trajectories

(1) SOFA Score Distribution: Each point is a validation observation colored by its SOFA score. In the *optimal and stable* model, patient states form a compact manifold with a smooth gradient of severity—low to high scores across the latent space. In contrast, the *overtrained and unstable* model produces sparse, unstructured embeddings lacking any discernible severity pattern.

(2) Mortality Trajectories: We connect each patient's first and last latent observations, with survivors in blue and non-survivors in red. Under the *optimal and stable* checkpoint, trajectories of non-survivors cluster together, reflecting prognostic separation. In the *overtrained* representation, survivor and non-survivor paths overlap chaotically, indicating a loss of clinical discriminability.

(3) Acuity Representation Correlation: Distributions of mean Pearson correlation between latent features and acuity scores show that the *optimal and stable* model maintains a right-skewed distribution (mean correlation ~ 0.45), whereas the *overtrained* model's distribution collapses around zero, losing severity-related information.

These visual findings demonstrate that early stopping not only prevents training instability but also preserves clinical structure in the learned representations. The clear separation of severity gradients and outcome trajectories suggests that stable training promotes latent spaces that align naturally with clinical acuity measures.

Motivated by these observations, we next investigate the quantitative relationship between the reconstruction loss $\mathcal{L}_{\text{MSE}}$ and the acuity correlation loss $\mathcal{L}_{\text{corr}}$. If these two losses evolve together, it would imply that improving reconstruction fidelity inherently enhances the model's ability to capture acuity-related structure in the latent space.

Figure 6 visualizes their relationship across training epochs and random seeds. A strong positive trend is evident—the smaller the reconstruction loss, the higher the alignment with acuity scores. Table 5 confirms this quantitatively, showing a high Pearson correlation ($r = 0.958 \pm 0.013$, $p < 10^{-45}$) and mutual information of 1.53 ± 0.18 nats. This co-evolution indicates that both metrics capture the same underlying latent quality, suggesting that state representations and clinical acuity are intrinsically linked.

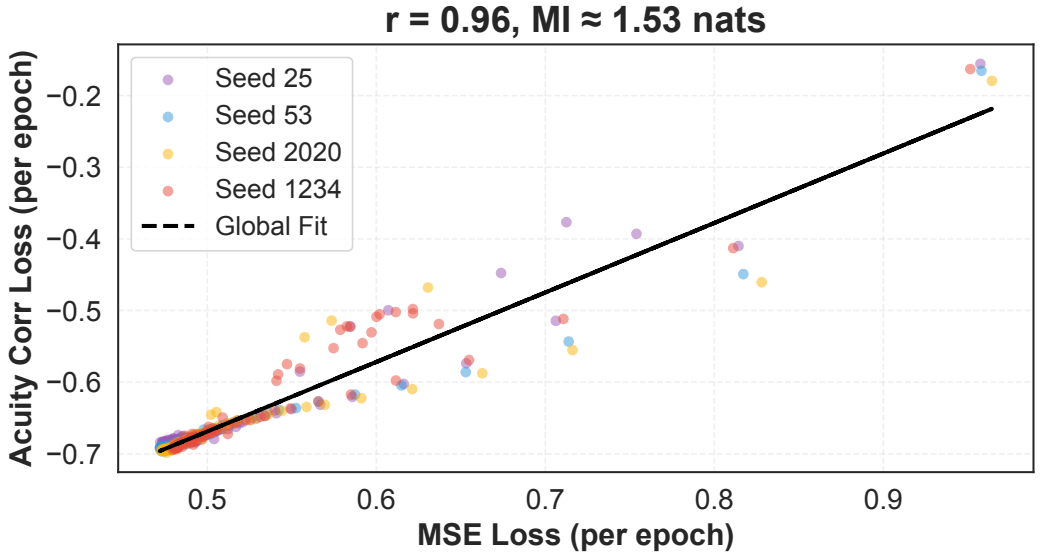

Figure 6. Relationship between reconstruction loss ($\mathcal{L}_{\text{MSE}}$) and acuity correlation loss ($\mathcal{L}_{\text{corr}}$) during CDE autoencoder training. Each point represents an epoch across random seeds. The fitted regression line shows strong positive correlation ($r = 0.958$, $p < 10^{-45}$).

Table 5. Correlation between $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{corr}}$ during CDE autoencoder training (averaged across random seeds).

Pearson corr. Coefficient r	Significant p-Value	Mutual Information
0.9578 ± 0.0127	$<10^{-45}$	1.5296 ± 0.1847 nats

Table 6 further quantifies training stability by reporting the Pearson correlation between the training loss $\mathcal{L}_{\text{total}}$ and the validation loss $\mathcal{L}_{\text{val}}$. We compute the Pearson correlation for each random seed across (i) the entire training run and (ii) the plateau region identified by our stopping criteria, then report the mean and standard deviation across seeds. The correlation across all epochs is moderate ($r \approx 0.62$), reflecting epochs where training dynamics are more variable (including transient instability and divergent phases). Crucially, on the plateau the correlation increases to $r \approx 0.93$, indicating nearly synchronous behavior of training and validation losses during the stable phase. This near-perfect plateau correlation demonstrates that epochs satisfying the plateau criterion provide reliable validation signals for model selection (i.e., stopping within the plateau is likely to yield models that generalize), whereas selecting checkpoints outside this region risks instability-induced degradation despite apparent training improvements.

Table 6. Correlation between training loss $\mathcal{L}_{\text{total}}$ and validation loss $\mathcal{L}_{\text{val}}$ over CDE autoencoder training epochs, averaged across random seeds, compared on plateau V.S. whole training epochs.

	All Epochs	**Plateau Epochs**
Pearson Corr.	0.6241 ± 0.1934	0.9284 ± 0.0295
p-value	<0.05	<0.05

From these findings, we propose that the observed alignment between reconstruction and acuity losses reflects a causal relationship: training signals that improve reconstruction fidelity also strengthen clinical interpretability in the learned latent space. Formally, we hypothesize that incorporating acuity-based regularization directly into the training objective will further enhance state representation quality and downstream reinforcement learning performance.

Hypothesis 1. *Let $\hat{s}_t^{(\lambda)}$ denote the latent state representation learned by a CDE autoencoder trained with acuity regularization weight λ. For a given early-stopped, numerically stable training regime,*

$$\mathit{RepQual}(\hat{s}_t^{(\lambda=1)}) > \mathit{RepQual}(\hat{s}_t^{(\lambda=0)}), \quad \textit{and} \tag{12}$$

$$\mathit{WIS}(\hat{s}_t^{(\lambda=1)}) > \mathit{WIS}(\hat{s}_t^{(\lambda=0)}), \tag{13}$$

where $\mathit{RepQual}(\cdot)$ measures the quality of the learned representation (e.g., via correlation with acuity scores or clinical separability), and $\mathit{WIS}(\cdot)$ denotes the downstream policy's weighted importance sampling return.

We test Hypothesis 1 through an ablation study comparing dBCQ policies trained on stable CDE representations with and without acuity regularization ($\lambda = 1$ vs. $\lambda = 0$), under identical stable training and early-stopping configurations.

As shown in Figure 7 and Table 7, incorporating acuity alignment yields significantly higher RL performance (WIS $= 0.9195$ vs. 0.6381). This directly contradicts prior findings by Killian et al. [4], who reported that acuity regularization was ineffective. When instability is eliminated through early stopping, its benefit becomes both clear and substantial.

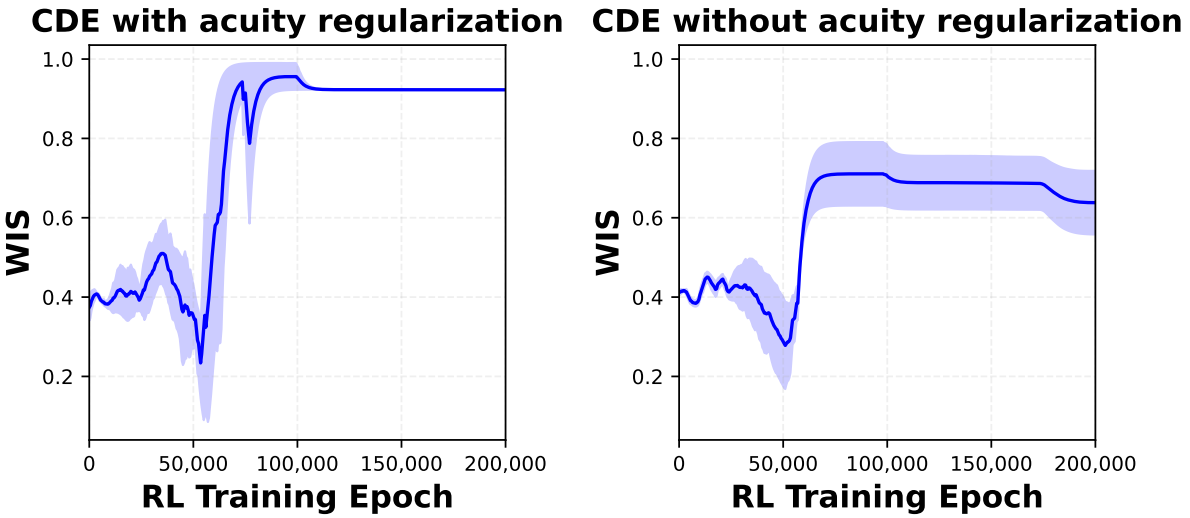

Figure 7. Ablation on acuity regularization: comparison of WIS returns for dBCQ policies trained on CDE representations with ($\lambda = 1$) and without ($\lambda = 0$) acuity alignment. Early stopping ensures stable convergence in both cases.

Table 7. Comparison with prior work [4] on acuity regularization findings. Our study revisits their conclusion regarding acuity regularization and shows clear improvement in both representation and policy performance.

Method	**WIS Return**	**Finding on Acuity Regularization**
Prior work [4]	0.775	No benefit from acuity regularization
Ours (Stable CDE, $\lambda = 1$)	0.920	Acuity regularization significantly improves both representation quality and RL policy performance.

Early stopping and stability: Unlike conventional early stopping, which primarily prevents overfitting, our multi-criteria stopping rule is designed to prevent *numerical instability*—A phase in which both training and validation

losses simultaneously increase after extended optimization. By halting training at the onset of instability, we preserve a clinically coherent latent space that aligns strongly with patient acuity and supports stable, generalizable downstream RL performance.

Summary: Visual and quantitative analyses jointly show that early stopping maintains both numerical stability and clinical interpretability in CDE training. Stable representations exhibit clear alignment with clinical severity, and acuity regularization amplifies this property, resulting in substantial improvements in both interpretability and RL policy performance. This overturns prior assumptions that acuity-based regularization is ineffective, providing new evidence that clinical indicators are powerful tools for guiding representation learning in healthcare.

5.3. Stabilization Method Comparison

We evaluate three stabilization methods using our early stopping criteria, with all methods using $\lambda = 1$ for acuity correlation. Each model was trained for 100 epochs with optimal hyperparameters from Table 3. For each method, we compute flatness metrics mentioned in Section 4.2 on the loss curve, respectively plateau length and mean absolute slope on plateau (S_1). Each encoder is then frozen at its *optimal and stable* checkpoint, and a downstream dBCQ RL policy is trained under identical conditions: policy learning rate $= 1 \times 10^{-5}$, dBCQ action elimination threshold $\tau_{BC} = 0.3$, and training epochs $= 2 \times 10^5$. The mean evaluated WIS returns on the validation trajectories are reported below.

As shown in Table 8, the implicit Adams solver produces the longest plateau, lowest mean absolute slope, and achieves the highest mean RL return ($\overline{WIS} = 0.9206$). Both the implicit Adams solver and stiffness regularization yield performance comparable to the baseline, while gradient clipping weakens both training flatness and policy performance.

Table 8. Flatness metrics and downstream RL WIS return for each stabilization method (under best hyperparameters, across random seeds).

Method	Plateau Length	S_1	$\overline{\text{WIS}}$
Baseline	42 ± 8.6	0.0028 ± 0.0005	0.9195
Gradient clipping	45.8 ± 4.5	0.0031 ± 0.0021	0.6547
Implicit Adams solver	46.5 ± 5.7	0.0027 ± 0.0006	0.9206
Stiffness regularization	43 ± 9.4	0.0030 ± 0.0008	0.9189

Figure 8 further demonstrates that when training RL on CDE representations from the baseline and implicit Adams solver, policies converge rapidly to high WIS returns. The stiffness-regularized representation converges more slowly but ultimately reaches a comparable final return. In contrast, gradient clipping leads to a noisy loss trajectory and a suboptimal WIS outcome, suggesting that it may overly dampen the CDE dynamics.

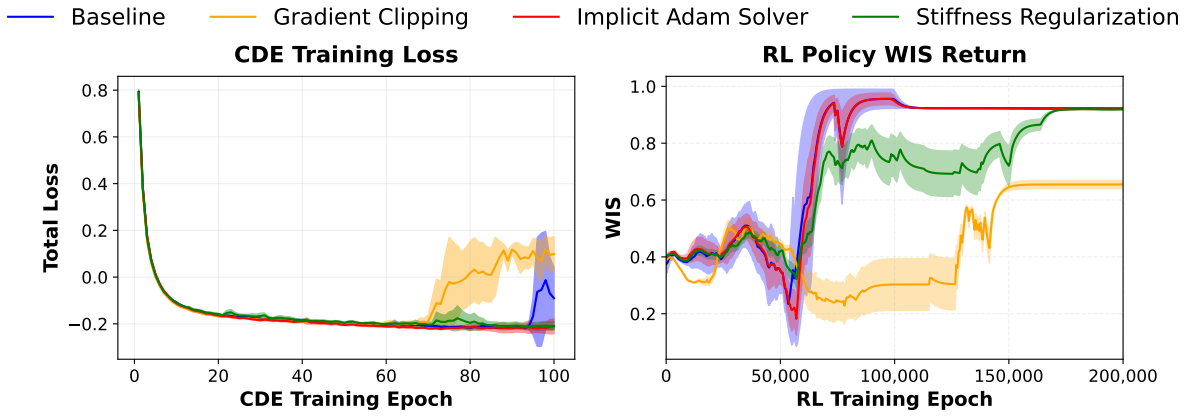

Figure 8. CDE autoencoder training loss ($mean \pm std$) and downstream RL WIS return ($mean \pm std$) on validation trajectories for each stabilization method.

Beyond these quantitative comparisons, we also visualize how each stabilization technique affects the clinical structure of the learned latent space. Figure 9 shows the SOFA, OASIS, and SAPS-II score distributions, as well as mortality trajectories, for the validation set under each stabilization method. All models follow the same early stopping criteria to ensure numerical stability.

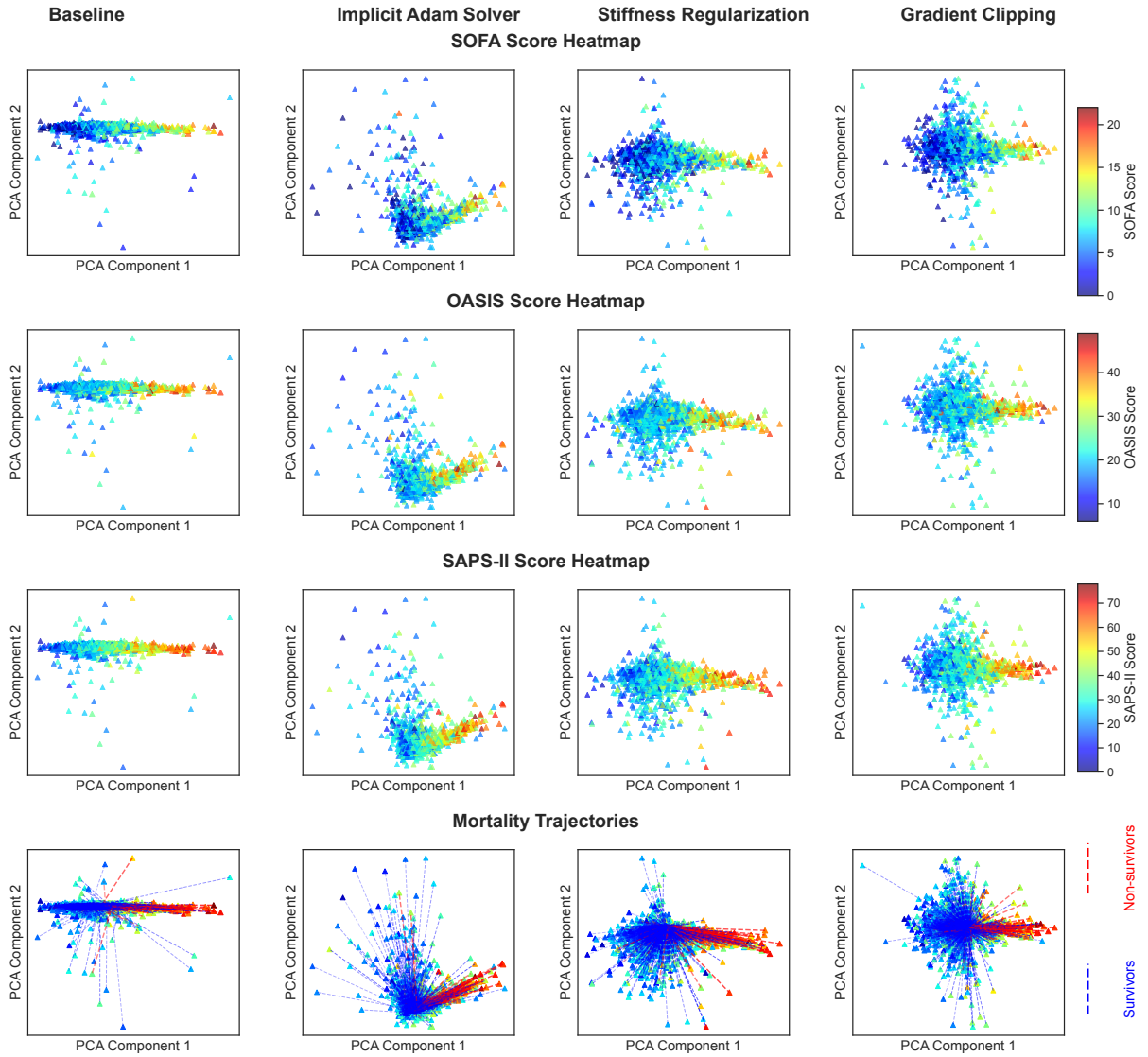

Figure 9. Clinical alignment visualization for each stabilization method. Shown are SOFA, OASIS, and SAPS-II score distributions and mortality trajectories of CDE autoencoders on the validation dataset. Under proper early stopping, all methods exhibit clear clustering of similar-acuity patients and visible separation between survivor and non-survivor trajectories. The implicit Adams solver produces the smoothest severity gradients and most coherent mortality boundaries, confirming its advantage in stability and interpretability.

6. Conclusions

This work provides a principled framework for stabilizing Neural CDE training in clinical reinforcement learning, addressing long-standing issues of instability and weak clinical interpretability in prior research. We demonstrate, through rigorous empirical evaluation, that careful stabilization and early stopping are not merely training heuristics but necessary conditions for obtaining robust, clinically meaningful state representations. Our proposed multi-criteria early stopping strategy and stabilization pipeline consistently improve policy performance, achieving a new state-of-the-art WIS return (>0.9 versus $\sim$0.775 in [4]) on the sepsis treatment task. These improvements are achieved without architectural changes, underscoring the effectiveness of our stability-focused training design.

Beyond performance, this study provides the first systematic evidence that *clinical acuity scores can actively enhance representation learning* when training is stabilized. Through quantitative correlation analysis and ablation studies, we show that the latent dynamics captured by Neural CDEs are intrinsically aligned with clinical indicators such as SOFA, SAPS-II, and OASIS. This finding overturns prior conclusions that acuity regularization offers little benefit, revealing instead that its effectiveness depends critically on stable optimization. The discovered strong correlation between reconstruction loss and acuity alignment highlights a shared latent structure, suggesting that clinically grounded supervision can guide representation learning in a more interpretable and physiologically consistent manner.

In summary, our work establishes both a practical and conceptual foundation for future research in clinical reinforcement learning: stabilization techniques and clinical indicator–aware regularization are key to bridging the gap between model optimization and clinical reasoning. We hope these insights inspire the design of next-generation RL systems that are not only high-performing but also clinically trustworthy and interpretable.

7. Future Work

While our study establishes a stable and clinically aligned framework for Neural CDE training, several important directions remain open for exploration.

(1) Understanding and mitigating WIS variance. Our current evaluation relies on Weighted Importance Sampling (WIS), which is known to exhibit a delicate bias–variance tradeoff. For the sepsis treatment task, we observed that the WIS variance is highly sensitive to the behavior policy threshold τ_{BC}: as τ_{BC} decreases, variance grows dramatically. Like prior studies, our work did not systematically analyze this effect. Future work should quantify how τ_{BC} influences the stability of policy evaluation and explore variance-reduction strategies such as doubly robust estimators and adaptive behavior-cloning thresholds.

(2) Expanding stabilization strategies. Our current stabilization design focuses on implicit solvers and stiffness regularization. Future extensions will examine additional techniques such as Gaussian noise injection [33], vector-field dropout [34], and spectral normalization for continuous-time networks. These methods may further suppress numerical oscillations and improve generalization, especially under data-scarce or high-frequency regimes.

(3) Broadening the scope of acuity alignment. A key insight from this work is that clinical acuity regularization not only enhances policy performance but also improves interpretability through latent–clinical alignment. In future research, we plan to extend this principle to other clinical time-series tasks such as septic shock progression [35], ICU stroke recovery prediction [36], and post-surgical complication forecasting [37]. We also aim to investigate whether similar alignment can be achieved using alternative physiological indicators or clinician-defined risk metrics, thereby validating the generalizability of acuity-informed representation learning.

(4) Integrating diverse offline RL algorithms. To evaluate learned state representations more comprehensively, future work will benchmark across state-of-the-art offline RL algorithms, including Fitted Q-Iteration (FQI) and Conservative Offline Model-Based Policy Optimization (COMBOP). Such comparisons will clarify how stability and clinical alignment interact with algorithmic design choices in offline RL.

Overall, our future research will continue to bridge the gap between stable representation learning and clinically grounded decision-making, advancing toward reinforcement learning systems that are both reliable and interpretable in real-world healthcare.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Patient consent was waived because this study utilized the MIMIC-III v1.4 database, which contains de-identified health data. The requirement for individual patient consent was waived during the original creation of the MIMIC-III database by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, as the research does not impact clinical care and all data are de-identified.

Data Availability Statement

The data used in this research were obtained from the Medical Information Mart for Intensive Care III (MIMIC-III) database, version 1.4, Available online: https://physionet.org/content/mimiciii/1.4/ (accessed on 4 September 2016) [7,8].

Acknowledgments

We acknowledge the contributions of the researchers, clinicians, and patients who made this dataset possible, and we confirm that our use of this data complies with PhysioNet's data use agreements. The author would like to thank Yue Xing (Purdue University) for guidance and assistance with the MIMIC-III data access process through PhysioNet.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used GPT to improve writing quality and correct grammatical errors. After using these tools, the author thoroughly reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A

Appendix A.1. Addtional Acuity Score Heatmaps

In Section 5.2, we have compared the SOFA score heatmap of *optimal and stable* and *overtrained and unstable* CDE representations. Here we provide the heatmaps of other acuity scores, including OASIS and SAPS-II. The heatmaps are generated by projecting the latent features of validation trajectories into a lower dimensional space using PCA, and coloring the points by their respective acuity scores.

The hyperparameters used for generating these heatmaps are the same as those used in Section 5.1, hidden size = 64, learning rate = 2×10^{-4}, $\lambda = 1$, $\epsilon_1 = 0.1$, $p = 30$, $\epsilon_2 = 0.02$, $\rho_{\text{threshold}} = 0.7$.

As shown in Figures A1 and A2, the *optimal and stable* CDE representation shows a clear clustering of patients with similar OASIS and SAPS-II scores, while the *overtrained and unstable* CDE representation exhibits a more scattered and random distribution, indicating a loss of clinical alignment. This further supports our finding that early stopping is crucial for maintaining acuity score alignment in CDE representations.

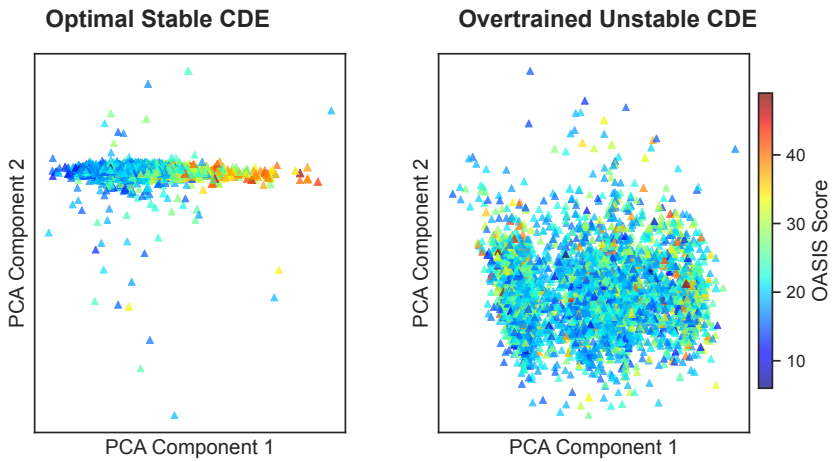

Figure A1. Heatmap of OASIS score distribution at *optimal and stable* and *overtrained and unstable* CDE representations on validation dataset.

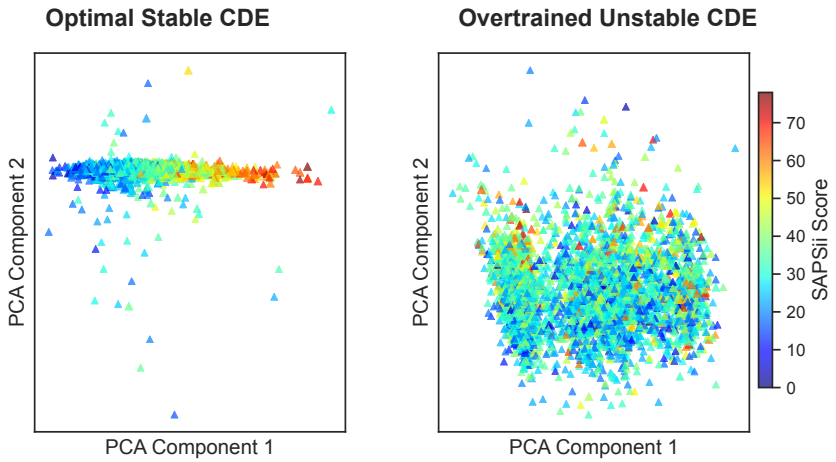

Figure A2. Heatmap of SAPS-II score distribution at *optimal and stable* and *overtrained and unstable* CDE representations on validation dataset.

Appendix A.2. Stabilization Methods

In this section, we provide a detailed description of the stabilization methods used in Section 5.3 to stabilize CDE autoencoder training. We evaluate three methods: implicit Adams solver, gradient clipping, and stiffness regularization. Each method is designed to improve the stability of the CDE autoencoder training process. Following the early stopping criteria, we evaluate the effectiveness of these methods by measuring the RL policy performance on the CDE representations, and also show the acuity alignment of each stabilization method on validation dataset in Figure 9.

Appendix A.3. Implicit Adams Solver

When training the Neural CDE latent state $h(t)$ evolving under observation o_t as in Equation (1), implicit Adams methods advance $h(t)$ using future values of the target function in the integration step. An s step Adams-Moulton update rule from step t_n to t_{n+1} is defined as follows [38]:

$$h(n+1) = h(n) + \Delta t \sum_{j=0}^{s} \beta_j \cdot f_\theta(h(n+1-j), o_{t_{n+1-j}}) \tag{14}$$

where $\Delta t = t_{n+1} - t_n$ and β_j are fixed coefficients. Note from Equation (14) that both sides include the term $f_\theta(h(n+1-j), o_{t_{n+1-j}})$, so a nonlinear equation needs to be solved at each step. Here we list orders 0, 1, 2 and 4 for Adams-Moulton:

$$h(n) = h(n-1) + \Delta t \cdot f_\theta(h(n), o_{t_n}); \tag{15}$$

$$\begin{aligned} h(n+1) = h(n) + \Delta t \cdot \Big(&\frac{1}{2} f_\theta(h(n+1), o_{t_{n+1}}) \\ &+ \frac{1}{2} f_\theta(h(n), o_{t_n})\Big); \end{aligned} \tag{16}$$

$$\begin{aligned} h(n+2) = h(n+1) + \Delta t \cdot \Big(&\frac{5}{12} f_\theta(h(n+2), o_{t_{n+2}}) \\ &+ \frac{8}{12} f_\theta(h(n+1), o_{t_{n+1}}) - \frac{1}{12} f_\theta(h(n), o_{t_n})\Big); \end{aligned} \tag{17}$$

$$\begin{aligned} h(n+4) = h(n+3) + \Delta t \cdot \Big(&\frac{251}{270} f_\theta(h(n+4), o_{t_{n+4}}) \\ &+ \frac{646}{720} f_\theta(h(n+3), o_{t_{n+3}}) - \frac{264}{720} f_\theta(h(n+2), o_{t_{n+2}}) \\ &+ \frac{106}{720} f_\theta(h(n+1), o_{t_{n+1}}) - \frac{19}{720} f_\theta(h(n), o_{t_n})\Big) \end{aligned} \tag{18}$$

In our experiments in Section 5.3, we used the implicit Adams-Moulton method of order 4, which is a 4th-order implicit method. We do a grid search for *step size* $\Delta t \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$, and the best performance is achieved with $\Delta t = \frac{1}{8}$, resulting in a mean WIS return of 0.9206 on the downstream RL policy.

As shown in Figure 8 and Table 8, the implicit solver not only makes the curvature of the loss function smoother, makes the plateau longer and more stable, but also achieves the best performance in downstream RL policy. This indicates that implicit solver method balances the trade-off between smoothness and extracting sharp transits in the latent space (e.g, sudden deteriorations in septic patients), which is crucial for ICU time-series data.

Appendix A.4. Gradient Clipping

Norm-based gradient clipping is a technique that limits the magnitude of the gradient vector during backpropagation. Given a parameter vector θ with gradient $g = \nabla_\theta L$ at a training step, the Euclidean norm $\|g\|_2$ is computed, then a threshold $\tau > 0$ is chosen, and the gradient is rescaled if its norm exceeds τ. Formally, the clipped gradient g_{clip} is defined by:

$$g_{\text{clip}} = \begin{cases} g, & \text{if } \|g\|_2 \le \tau \\ \tau \frac{g}{\|g\|_2} & \text{if } \|g\|_2 > \tau \end{cases} \tag{19}$$

Mathematically, gradient clipping can be seen as a form of regularization on the optimization trajectory. It does not alter the loss function or objective explicitly, but it modifies the optimization dynamics to avoid excessively

large parameter jumps.

When training Neural CDE model, as introduced in Equation (1), one backpropagates through the numerical solver to compute the gradients of the loss function with respect to the model parameters. If f_θ has large Lipschitz constant or the integration time is long, the backward gradients can suffer from gradient explosion. Previous works [26,39] explicitly cite gradient clipping as an effective method for controlling gradient explosion and oscillatory in continuous-time models.

In Section 4.2, we apply gradient clipping with a grid search on the threshold $\tau \in \{0.5, 1.0, 1.5\}$, and as a result, when $\tau = 1.0$, the CDE autoencoder achieves the best performance in downstream RL policy. With a mean WIS return of 0.6547, it is lower than the baseline CDE autoencoder without stabilization, which achieves a mean WIS return of 0.9195. An explanation for this could be that in irregular ICU time-series, patient trajectories are sparese and short, and gradient clipping uniformly dampens high-magnitude updates, however, some large gradients might be clinically meaningful. Also, in ICU trajectories, there are sudden deteriorations in septic patients, which causes large updates to θ, while gradient clipping will suppress that update, leading to a less informative representation.

Appendix A.5. Stiffness Regularization

This regularization technique leverages stiffness indicators to shape the training dynamics of Neural CDEs. Recall from Section 2.2 that the encoder is defined as :

$$\partial h(t) = f_\theta(h(t))\partial o_t \tag{20}$$

Assume that we solve it over time interval $[t_0, t_1]$, and reconstruct the downstream target $\hat{o}_{t_1} = \phi(h(t_1))$. Let $\mathcal{L}_{\text{total}}$ in Equation (5) be the reconstruction loss, $\{t_j\}_{j=1}^N$ be the solver time steps. Specifically, for each time step t_j, the stiffness score $\mathcal{S}_j$ is apprixmated via the real parts of the eigenvalues of the local Jacobian:

$$\mathcal{S}_j = \max\left\{|\Re(\lambda_i)| : \lambda_i \in \text{eig}\left(\frac{\partial f_\theta(h(t_j))}{\partial h(t)}\right)\right\} \tag{21}$$

where

- $\frac{\partial f_\theta(h(t_j))}{\partial h(t)} \in \mathbb{R}^{d\times d}$ denotes the Jacobian of the vector field w.r.t. the hidden state at time t_j.
- $\text{eig}(\cdot)$ denotes the set of eigenvalues.
- $\Re(\lambda_i)$ denotes the real part of λ_i.

Then the total loss becomes:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{total}} + \lambda_{\text{reg}} \cdot \underbrace{\sum_{j=1}^{N} S_j}_{\text{Stiffness Regularization}} \tag{22}$$

This objective encourages the encoder's neural vector field f_θ to not only fit the data but also generate dynamics that are less stiff, leading to smoother trajectories in the latent space. The hyperparameter λ_{reg} controls the strength of this regularization, which balances the tradeoff between fitting the data and minimizing the solver complexity. As mentioned in Table 3, we do a grid search on $\lambda_{\text{reg}} \in \{0.005, 0.01, 0.015\}$, and find that $\lambda_{\text{reg}} = 0.01$ achieves the best performance in downstream RL policy, with a mean WIS return of 0.9189. Figure 8 shows that the stiffness regularization stabilizes the CDE training and lead to a longer, smoother plateau, which is beneficial for picking the optimal stopping epoch. And its downstream RL policy reaches a comparable optimal WIS return.

In summary, stiffness regularzation is a promising method for stabilizing CDE autoencoder training for learning MIMIC-III data representation, as it encourages the model to learn smoother dynamics, which is crucial for capturing the underlying patterns in irregular ICU time-series data. It also helps to extend the plateau of the training loss, making it easier to find the optimal stopping epoch.

References

1. Shashikumar, S.P.; Josef, C.S.; Sharma, A.; et al. DeepAISE—An interpretable and recurrent neural survival model for early prediction of sepsis. *Artif. Intell. Med.* **2021**, *113*, 102036.
2. Solís-García, J.; Vega-Márquez, B.; Nepomuceno, J.A.; et al. Comparing artificial intelligence strategies for early sepsis detection in the ICU: An experimental study. *Appl. Intell.* **2023**, *53*, 30691–30705.
3. Komorowski, M.; Celi, L.; Badawi, O.; et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **2018**, *24*, 1716–1720.

4. Killian, T.W.; Zhang, H.; Subramanian, J.; et al. An Empirical Study of Representation Learning for Reinforcement Learning in Healthcare. In Proceedings of the Machine Learning for Health NeurIPS Workshop, Online, 11 December 2020; Volume 136, pp. 139–160.
5. Jayaraman, P.; Desman, J.; Sabounchi, M.; et al. A Primer on Reinforcement Learning in Medicine for Clinicians. *NPJ Digit. Med.* **2024**, *7*, 337.
6. Böck, M.; Malle, J.; Pasterk, D.; et al. Superhuman performance on sepsis MIMIC-III data by distributional reinforcement learning. *PLOS ONE* **2022**, *17*, e0275358.
7. Johnson, A.; Pollard, T.; Mark, R. MIMIC-III Clinical Database (Version 1.4). Available online: https://physionet.org/content/mimiciii/1.4/ (accessed on 4 September 2016).
8. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035.
9. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220.
10. Morrill, J.; Kidger, P.; Yang, L.; et al. Neural Controlled Differential Equations for Online Prediction Tasks. *arXiv* **2021**, arXiv:2106.11028.
11. Agarwala, S.; Dees, B.; Lowman, C. Geometric instability of out of distribution data across autoencoder architecture. *arXiv* **2022**, arXiv:2201.11902.
12. Shoosmith, J.N. Numerical Analysis. In *Encyclopedia of Physical Science and Technology*, 3rd ed.; Meyers, R.A., Ed.; Academic Press: New York, NY, USA, 2003; pp. 39–70.
13. Baker, J.; Xia, H.; Wang, Y.; et al. Proximal Implicit ODE Solvers for Accelerating Learning Neural ODEs. *arXiv* **2022**, arXiv:2204.08621.
14. Gao, Y. Stable CDE Autoencoders with Acuity Regularization for Offline Reinforcement Learning in Sepsis Treatment. *arXiv* **2025**, arXiv:2506.15019.
15. Rubanova, Y.; Chen, R.T.Q.; Duvenaud, D.K. Latent Ordinary Differential Equations for Irregularly-Sampled Time Series. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
16. Che, Z.; Purushotham, S.; Cho, K.; et al. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* **2018**, *8*, 6085.
17. Yoon, J.; Jordon, J.; van der Schaar, M. GAIN: Missing Data Imputation using Generative Adversarial Nets. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
18. Zhang, K.; Xue, Y.; Flores, G.; et al. Modelling EHR Timeseries by Restricting Feature Interaction. *arXiv* **2019**, arXiv:1911.06410.
19. Kidger, P.; Foster, J.; Li, X.; et al. Neural SDEs as Infinite-Dimensional GANs. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; Volume 139, pp. 5453–5463.
20. Chen, R.T.Q.; Rubanova, Y.; Bettencourt, J.; et al. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
21. Timothée, L.; Natalia, D.R.; Jean-Franois, G.; et al. State representation learning for control: An overview. *Neural Netw.* **2018**, *108*, 379–392.
22. Hairer, E.; Nørsett, S.; Wanner, G. *Solving Ordinary Differential Equations I Nonstiff Problems*, 2nd ed.; Springer: Berlin, Germany, 2000.
23. Kim, S.; Ji, W.; Deng, S.; et al. Stiff neural ordinary differential equations. *Chaos Interdiscip. J. Nonlinear Sci.* **2021**, *31*. https://doi.org/10.1063/5.0060697.
24. Gebregiorgis, S.; Gonfa, G. The comparison of runge-kutta and adams-bashforh-moulton methods for the first order ordinary differential equations. *Int. J. Curr. Res.* **2021**, *8*, 27356–27360.
25. Fronk, C.; Petzold, L. Training Stiff Neural Ordinary Differential Equations with Implicit Single-Step Methods. *arXiv* **2024**, arXiv:2410.05592.
26. Nguyen, H.H.N.; Nguyen, T.; Vo, H.; et al. Improving Neural Ordinary Differential Equations with Nesterov's Accelerated Gradient Method. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 7712–7726.
27. Pal, A.; Ma, Y.; Shah, V.; et al. Opening the Blackbox: Accelerating Neural Differential Equations by Regularizing Internal Solver Heuristics. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8325–8335.
28. Vincent, J.L.; Moreno, R.; Takala, J.; et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* **1996**, *22*, 707–710.
29. Gall, J.R.L.; Lemeshow, S.; Saulnier, F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* **1993**, *270*, 2957–2963.
30. Johnson, A.E.W.; Kramer, A.A.; Clifford, G.D. A New Severity of Illness Scale Using a Subset of Acute Physiology and

Chronic Health Evaluation Data Elements Shows Comparable Predictive Accuracy. *Crit. Care Med.* **2013**, *41*, 1711–1718.
31. Mahmood, A.R.; van Hasselt, H.; Sutton, R.S. Weighted Importance Sampling for Off-Policy Learning with Linear Function Approximation. In *Advances in Neural Information Processing Systems*; Ghahramani, Z., Welling, M., Cortes, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
32. Huang, Y.; Cao, R.; Rahmani, A. Reinforcement Learning For Sepsis Treatment: A Continuous Action Space Solution. In Proceedings of the 7th Machine Learning for Healthcare Conference, Durham, NC, USA, 5–6 August 2022; Volume 182, pp. 631–647.
33. Alain, G.; Bengio, Y. What Regularized Auto-Encoders Learn from the Data Generating Distribution. *J. Mach. Learn. Res.* **2014**, *15*, 3563–3593.
34. Le, L.; Patterson, A.; White, M. Supervised Autoencoders: Improving Generalization Performance with Unsupervised Regularizers. In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
35. Giannini, H.M.; Ginestra, J.C.; Chivers, C.; et al. A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice. *Crit. Care Med.* **2019**, *47*, 1485–1492.
36. Choo, Y.J.; Chang, M.C. Use of Machine Learning in Stroke Rehabilitation: A Narrative Review. *Brain NeuroRehabil.* **2022**, *15*, e26.
37. Hassan, A.M.; Rajesh, A.; Asaad, M.; et al. Artificial Intelligence and Machine Learning in Prediction of Surgical Complications: Current State, Applications, and Implications. *Am. Surg.* **2023**, *89*, 25–30.
38. Brouwer, E.D.; Krishnan, R.G. Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections. *arXiv* **2023**, arXiv:2303.01841.
39. Coelho, C.; Costa, M.F.P.; Ferrás, L. Enhancing continuous time series modelling with a latent ODE-LSTM approach. *Appl. Math. Comput.* **2024**, *475*, 128727.