Trajectory Flow Matching with Applications to Clinical Time Series Modeling

Xi Zhang[1,2 *] **Yuan Pu**[3 *] **Yuki Kawamura**[4] **Andrew Loza**[3]
Yoshua Bengio[2,5,6] **Dennis L. Shung**[3 †] **Alexander Tong**[2,5 †]

[1]McGill University, [2]Mila - Quebec AI Institute,
[3]Yale School of Medicine
[4]School of Clinical Medicine, University of Cambridge,
[5]Université de Montréal, [6]CIFAR Fellow

Abstract

Modeling stochastic and irregularly sampled time series is a challenging problem found in a wide range of applications, especially in medicine. Neural stochastic differential equations (Neural SDEs) are an attractive modeling technique for this problem, which parameterize the drift and diffusion terms of an SDE with neural networks. However, current algorithms for training Neural SDEs require backpropagation through the SDE dynamics, greatly limiting their scalability and stability. To address this, we propose **Trajectory Flow Matching** (TFM), which trains a Neural SDE in a *simulation-free* manner, bypassing backpropagation through the dynamics. TFM leverages the flow matching technique from generative modeling to model time series. In this work we first establish necessary conditions for TFM to learn time series data. Next, we present a reparameterization trick which improves training stability. Finally, we adapt TFM to the clinical time series setting, demonstrating improved performance on three clinical time series datasets both in terms of absolute performance and uncertainty prediction, a crucial parameter in this setting.

1 Introduction

Real world problems often involve systems that evolve continuously over time, yet these systems are usually noisy and irregularly sampled. In addition, real-world time series often relate to other covariates, leading to complex patterns such as intersecting trajectories. For instance, in the context of clinical trajectories in healthcare, patients' vital sign evolution can follow drastically different, crossing paths even if the initial measurements are similar, due to the influence of the covariates such as medication intervention and underlying health conditions. These covariates can be time-varying or static, and often sparse.

Differential equation-based dynamical models are proficient at learning continuous variables without imputations [Chen et al., 2018, Rubanova et al., 2019, Kidger et al., 2021b]. Nevertheless, systems governed by ordinary differential equations (ODEs) or stochastic differential equations (SDEs) are unable to accommodate intersecting trajectories, and thus requires modifications such as augmentation or modelling higher-order derivatives [Dupont et al., 2019]. While ODEs model deterministic systems, SDEs contain a diffusion term and can better represent the inherent uncertainty and fluctuations present in many real world systems. However, fitting stochastic equations to real life

*Joint first authorship
†Joint senior authorship. Correspondence to `alexander.tong@mila.quebec`
Code available at: `https://github.com/nZhangx/TrajectoryFlowMatching`

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

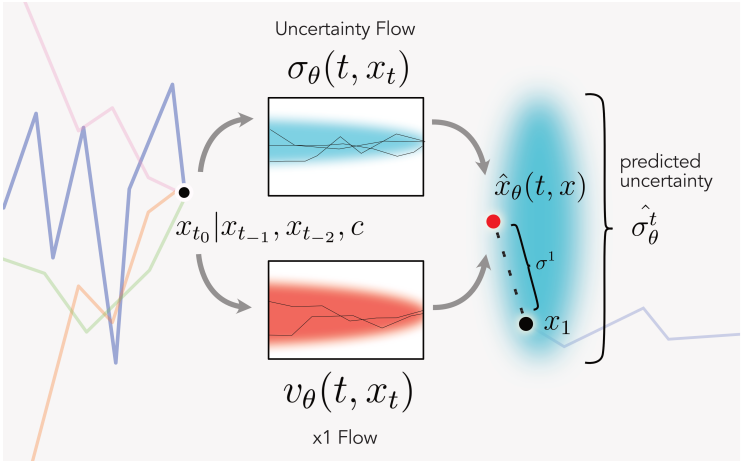

Figure 1: Trajectory Flow Matching trains both an estimator of the next timepoint ($\hat{x}_\theta(t, x)$) and an estimation of the uncertainty ($\sigma_\theta(t, x_t)$). Using the conditional flow matching framework, these can be used to predict the instantaneous velocity $v_\theta(t, x_t)$ and future observations. Both flows are conditioned on past data $x_{[t-h,t-1]}$ and conditional variables c.

data is challenging because they have thus far required time-consuming backpropagation through an SDE integration.

In the domain of generative models, diffusion models [Ho et al., 2020, Nichol and Dhariwal, 2021, Song et al., 2021] and more recently flow matching models [Lipman et al., 2023, Albergo et al., 2023, Li et al., 2020] have had enormous success by training dynamical models in a *simulation-free* framework. The simulation-free framework facilitates the training of much larger models with significantly improved speed and stability. In this work we generalize simulation-free training for fitting stochastic differential equations to time-series data, to learn population trajectories while preserving individual characteristics with conditionals. We present this method as **Trajectory Flow Matching**. We demonstrate that our method outperforms current state of the art time series modelling architecture including RNN, ODE based and flow matching methods. We empirically demonstrate the utility of our method in clinical applications where hemodynamic trajectories are critical for ongoing dynamic monitoring and care. We applied our method to the following longitudinal electronic health record datasets from multiple clinical settings: medical intensive care unit (MICU) data of patients with sepsis, Emergency Department (ED) data of patients with acute gastrointestinal bleeding, and MICU data of patients with acute gastrointestinal bleeding.

Our main contributions are:

- We prove the conditions under which continuous time dynamics can be trained simulation-free using matching techniques.
- We extend the approach to irregularly sampled trajectories with a *time predictive loss* and to estimate uncertainty using an *uncertainty prediction loss*.
- We empirically demonstrate that our approach reduces the error by 15-83% when applied to the real world clinical data modelling.

2 Preliminaries

2.1 Notation

We consider the setting of a distribution of trajectories over $\mathbb{R}^d$ denoted $\mathcal{X} := \{x^1, x^2, \ldots, x^n\}$ where each x^i consists of a set of trajectories of length T i.e. $x^i := \{x_1^i, x_2^i, \ldots, x_T^i\}$ with associated times $t^i := \{t_1^i, t_2^i, \ldots, t_T^i\}$. Let $x^i_{[t-h,t-1]}$ denote a vector of the last h observed time points. We denote a (Lipschitz smooth) time dependent vector field conditioned on arbitrary conditions $c \in \mathbb{R}^e$ $v(t, x_t, x_{[t-h,t-1]}, c) \to \frac{dx}{dt} : ([0, 1], \mathbb{R}^d, \mathbb{R}^{h\times d}, \mathbb{R}^e) \to \mathbb{R}^d$ with flow $\phi_t(v)$ which induces the time-dependent density $p_t = \phi_t(v)_\#(p_0)$ for any density $p_0 : \mathbb{R}^d \to \mathbb{R}_+$ with $\int_{\mathbb{R}^d} p_0 = 1$. We also consider the coupling $\pi(x_0, x_1)$ which operates on the product space of marginal distributions p_0, p_1.

2.2 Neural Stochastic Differential Equations

A stochastic differential equation (SDE) can be expressed in terms of a smooth drift $f : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ and diffusion $g : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d^2}$ in the Ito sense as:

$$dx_t = f\, dt + g\, dW_t$$

where $W_t : [0, T] \to \mathbb{R}^d$ is the d-dimensional Wiener process. A density $p_0(x_0)$ evolved according to an SDE induces a collection of marginal distributions $p_t(x_t)$ viewed as a function $p : [0, T] \times \mathbb{R}^d \to \mathbb{R}_+$. In a *Neural* SDE [Li et al., 2020, Kidger et al., 2021a,b] the drift and diffusion terms are parameterized with neural networks $f_\theta(t, x_t)$ and $g_\theta(t, x_t)$.

$$dx_t = f_\theta(t, x_t)dt + g_\theta(t, x_t)dW_t \tag{1}$$

where the goal is to select θ to enforce $x_T \sim X_{true}$ for some distributional notion of similarity such as the Wasserstein distance [Kidger et al., 2021b] or Kullback-Leibler divergence [Li et al., 2020]. However, these objectives are *simulation-based*, requiring a backpropagation through an SDE solver, which suffers from severe speed and stability issues. While some issues such as memory and numerical truncation can be ameliorated using the adjoint state method and advanced numerical solvers [Kidger et al., 2021b], optimization of Neural SDEs is still a significant issue.

We note that in the special case of zero-diffusion (i.e. $g_\theta(t, x_t) = 0$) this reduces to a neural *ordinary* differential equation (Neural ODE) [Chen et al., 2018], which is easier to optimize than SDEs, but still presents challenges to scalability.

2.3 Matching algorithms

Matching algorithms are a *simulation-free* class of training algorithms which are able to bypass backpropagation through the solver during training by constructing the marginal distribution as a mixture of tractable conditional probability paths.

The marginal density p_t induced by eq. 1 evolves according to the *Fokker-Plank* equation (FPE):

$$\partial_t p_t = -\nabla \cdot (p_t f_t) + \frac{g^2}{2} \Delta p_t \tag{2}$$

where $\Delta p_t = \nabla \cdot (\nabla p_t)$ denotes the *Laplacian* of p_t and gradients are taken with respect to x_t.

Matching algorithms first construct a factorization of p_t into conditional densities $p_t(x_t|z)$ such that $p_t = \mathbb{E}_{q(z)}\left[p_t(x_t|z)\right]$ and where $p_t(x_t|z)$ is generated by an SDE $dx_t = v_t(x_t|z)dt + \sigma_t(x_t|z)dW_t$. Given this construction it can be shown that the minimizer of

$$\mathcal{L}_{\text{match}}(\theta) := \mathbb{E}_{t, q(z), p_t(x|z)}\left[\left\|f_\theta(t, x_t) - v_t(x_t|z)\right\|^2 + \lambda_t^2 \left\|g_\theta(t, x_t) - \sigma_t(x_t|z)\right\|^2\right] \tag{3}$$

satisfies the FPE of the marginal p_t. This is especially useful in the generative modeling setting where q_0 is samplable noise (e.g. $\mathcal{N}(0, 1)$) and q_1 is the data distribution. Then we can define $z := (x_0, x_1)$ as a tuple of noise and data with $q(z) := q_0(x_0) \otimes q_1(x_1)$. This makes eq. 3 optimize a model which will draw new samples according to the data distribution $q_1(x_1)$ using

$$x_0 \sim q_0; \quad x_1 = \int_0^1 f_\theta(t, x_t)dt + g_\theta(t, x_t)dW_t \tag{4}$$

with the integration computed numerically using any off-the-shelf SDE solver. While this is guaranteed to preserve the distribution over time, it is not guaranteed to preserve the *coupling* of q_0 and q_1 (if given).

Paired bridge matching In generative modeling random pairings [Liu et al., 2023c, Albergo and Vanden-Eijnden, 2023, Albergo et al., 2023] or optimal transport [Tong et al., 2024, Pooladian et al., 2023] pairings are constructed for the conditional distribution $q(z)$. However, in some problems we would like to match pairs of points as is the case in image-to-image translation [Isola et al., 2017, Liu et al., 2023a, Somnath et al., 2023]. In this case, training data comes as pairs (x_0, x_1). In this case we set $q(z) := q(x_0, x_1)$ to be samples from these known pairs, and optimize eq. 3. While empirically, these models perform well, there are no guarantees that the coupling will be preserved outside of the special case when data comes from the (entropic) optimal transport coupling $\pi^*_\varepsilon(q_0, q_1)$ and defined as:

$$\pi^*_\varepsilon(q_0, q_1) = \underset{\pi \in U(q_0, q_1)}{\arg\min} \int d(x_0, x_1)^2\, d\pi(x_0, x_1) + \varepsilon\, \mathrm{KL}(\pi \| q_0 \otimes q_1), \tag{5}$$

Algorithm 1 General Trajectory Flow Matching

Input: Trajectories $\mathcal{X}$, noise σ, initial network v_θ.
while Training **do**
  $z \sim \mathcal{U}(\mathcal{X}), \quad k \sim \mathcal{U}\{1, T-1\}, \quad t \sim \mathcal{U}(0,1)$
  $\mu_t \leftarrow t\boldsymbol{x}_k + (1-t)\boldsymbol{x}_{k+1}$
  $x_t \sim \mathcal{N}(\mu_t, \sigma^2 t(1-t)I)$
  $\mathcal{L}_{\text{TFM}}(\theta) \leftarrow \left\| v_\theta(k+t, x_t) - \frac{x_{k+1}-x_t}{1-t} \right\|^2$
  $\mathcal{L}_{\sigma_t}(\theta) \leftarrow \|\sigma_\theta(k+t, x_t) - \mathcal{L}_{\text{TFM}}\|^2$
  $\theta \leftarrow \text{Update}(\theta, \nabla_\theta \mathcal{L}_{\text{TFM}}(\theta), \nabla_\theta \mathcal{L}_{\sigma_t}(\theta))$
return v_θ, σ_θ

where $U(q_0, q_1)$ is the set of admissible transport plans (i.e. joint distributions over x_0 and x_1 whose marginals are equal to q_0 and q_1) as shown in [Shi et al., 2023] for some regularization parameter $\varepsilon \in \mathbb{R}_{\geq 0}$.

3 Trajectory Flow Matching

We now describe our simulation-free method to learn SDEs on time-series data using *trajectory flow matching* as summarized in Alg. 1. In the case of time series we need to ensure that trajectory couplings are preserved. We first set out a general algorithm for flow matching on vector fields in §3.1 then present a numerical reparameterization which we find stabilizes training in §3.2, a next observation prediction for irregularly sampled time series in §3.3, and finally present how to learn the noise in §3.4.

3.1 Preserving Couplings

In this section, we assume access to fully observed and evenly spaced trajectories $\mathcal{X} = (x^1, x^2, \ldots, x^n)$ with $x^i := (x_1^i, x_2^i, \ldots, x_T^i)$ for clarity and notational simplicty. We note that our method is easily extensible to the more general setting of irregularly sampled trajectories. In this simplified case we let

$$z := (x_1, x_2, \ldots, x_T) \tag{6}$$
$$q(z) := \mathcal{U}(\mathcal{X}) \tag{7}$$
$$p_t(x|z) := \mathcal{N}((\lceil t \rceil - t)x_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)x_{\lceil t \rceil}, \sigma^2(\lceil t \rceil - t)(t - \lfloor t \rfloor)\mathbf{I}) \tag{8}$$
$$u_t(x|z) := \frac{x_{\lceil t \rceil} - x_t}{\lceil t \rceil - t} \tag{9}$$

where $\mathcal{U}(\mathcal{X})$ is the uniform empirical distribution over $\mathcal{X}$, $\lceil \cdot \rceil$, $\lfloor \cdot \rfloor$ are the ceiling and floor functions, and $\mathcal{N}(\cdot, \cdot)$ is the multivariate normal distribution. This is a valid regression in the sense that a function minimized with Alg. 1 will return a stochastic process that will match the observed marginal distributions over time as shown in the following lemma.

Lemma 3.1. *The SDE $dx_t = u_t(x|z)dt + \sigma^2 dW_t$ where u_t is defined in eq. 9 generates $p_t(x|z)$ in eq. 8 with initial condition $p_0 := \delta_{x_1}$ where δ is the Dirac delta function.*

however, while useful, this is still insufficient for time series modeling, as it does not ensure coupling preservation. For intuition why this is an issue see figure 2.

In TFM we ensure that the couplings are preserved for history lengths $h > 0$. i.e. $\hat{\pi}(x_{T-h}, x_{T-h+1}, \ldots, x_T) = \pi(x_{T-h}, x_{T-h+1}, \ldots, x_T)$. We first establish a method to ensure that these couplings are preserved allowing us to use simulation-free flow matching training for the time-series modeling task. Specifically, as long as the model takes as input $(x_{T-h}, x_{T-h+1}, \ldots, x_T)$ in predicting the flow from $T \to T+1$, then there exists a function $f_\theta(X_{T-h:T})$ such that the coupling is preserved.

Proposition 3.2 (Coupling Preservation)**.** *Under mild regulatory criteria on $u_t(\cdot|z)$, p_t, and q, if*

$$\mathbb{E}_{t\sim\mathcal{U}(0,T), z\sim q(z), c\sim q(c|z), x_t\sim p_t(x_t|z)} \|u_t(x_t|z,c) - u_t(x_t|c)\|_2^2 = 0$$

and $z, q(z), p_t(x|z)$ and $u_t(x|z)$ are as defined in eqs. 6-9 then $\Pi(u)^\star = \Pi^\star(x_{1:T})$.

Where $\Pi(u)^\star$ represents the coupling of a model which attains minimal loss according to eq. 3 and $\Pi^\star(x_{1:T})$ is the coupling of the data distribution. Intuitively, as long as no two paths cross given conditionals c, then the coupling is preserved. In prior work $c = \emptyset$, and the coupling is only preserved in special cases such as eq. 5.

We next enumerate three assumptions under which the coupling is guaranteed to be preserved at the optima. We note that these are

- **(A1)** When $c = x_0$ and there exists $T : \mathcal{X} \to \mathcal{X}$ such that $T(x_0) = x_1$ iff $\Pi^\star(x_0, x_1)$. We note that this is equivalent to asserting the existence of a Monge map $T^\star$ for the coupling $\Pi^\star$.
- **(A2)** There exist no two trajectories x^i, x^j such that $x^i_t = x^j_t$ for h consecutive observations and $g = 0$.
- **(A3)** Trajectories are associated with unique conditional vectors c independent of t.

Even in cases when **(A1)-(A3)** may not hold exactly, TFM is a useful model and can often still learn useful models of the data. In some sense uniqueness up to some history length is enough as it shows TFM is as powerful as discrete-time autoregressive models. Proofs and further examples are available in §A.1.

3.2 Target prediction reparameterization

While flow matching generally predicts the flow, there is a target predicting equivalent namely given $v_\theta(t, x) := \frac{\hat{x}_\theta^{\lceil t \rceil}(t, x_t) - x_t}{\lceil t \rceil - t}$ and $u_t(x|z) := \frac{x^{\lceil t \rceil} - x_t}{\lceil t \rceil - t}$ which is equivalent to $x_1 - x_0$ when $x_t : tx_1 + (1 - t)x_0$ then it is easy to show that the target predicting loss is equivalent to a time-weighted flow-matching loss. Specifically let the target predicting loss be

$$\mathcal{L}_{\text{target}}(\theta) = \mathbb{E}_{t, q(z), p_t(x|z)} \|\hat{x}_\theta^{\lceil t \rceil}(t, x) - x^{\lceil t \rceil}\|^2 \tag{10}$$

then it is easy to show that

Proposition 3.3. *There exists a scaling function $c(t) : \mathbb{R}_+ \to \mathbb{R}$ such that $\mathcal{L}_{target}(\theta) = c(t)\mathcal{L}_{match}(\theta)$.*

3.3 Irregularly sampled trajectories

We next consider irregularly sampled time series of the form $x^i := \big((x^i_1, t^i_1), (x^i_2, t^i_2), \ldots, (x^i_T, t^i_T)\big)$ with $t^i_1 < t^i_2 < \cdots < t^i_T$ with t_{next} denoting the next timepoint observed after time t. In this case, when combined with the target predicting reparameterization in §3.2, we can predict the time till next observation. We therefore parameterize an auxiliary model $h_\theta(t, x_t) : [0, T] \times \mathbb{R}^d \to [0, T]$ which predicts the next observation time. This is useful numerically, but also, perhaps more importantly, is useful in a clinical setting, where the spacing between measurements can be as informative as the measurements themselves [Allam et al., 2021]. h_θ is trained to predict the time till the next observation with the *time predictive loss*:

$$\mathcal{L}_{\text{tp}}(\theta) = \sum_{t \in \mathcal{T}^i} \|h_\theta(t, x_t) - (t_{\text{next}} - t)\|_2^2 \tag{11}$$

where t_{next} is the time of the next measurement. This can be used in conjunction with the x_{next} predictor to calculate the flow at time t as

$$v_\theta(t, x_t) := \frac{\hat{x}^1_\theta(t, x_t) - x_t}{h_\theta(t, x_t) - t} \tag{12}$$

which can be used for inference on new trajectories.

3.4 Uncertainty prediction

Finally, we consider uncertainty prediction. till now we have defined conditional probability paths using a fixed noise parameter σ. However, this does not have to be fixed. Instead, we consider a *learned* $\sigma_\theta(t, x_t)$ which can be learned iteratively with the loss:

$$\mathcal{L}_{\text{uncertainty}}(\theta, x) = \sum_{t \in \mathcal{T}} \Big\| \sigma_\theta(t, x_t) - \|\hat{x}_\theta(t, x_t) - x_{\text{next}}\|_2^2 \Big\|_2^2 \tag{13}$$

which learns to predict the error in the estimate of x_t. This loss can be interpreted as training an epistemic uncertainty predictor which is similar to that proposed in direct epistemic uncertainty prediction (DEUP) [Lahlou et al., 2023].

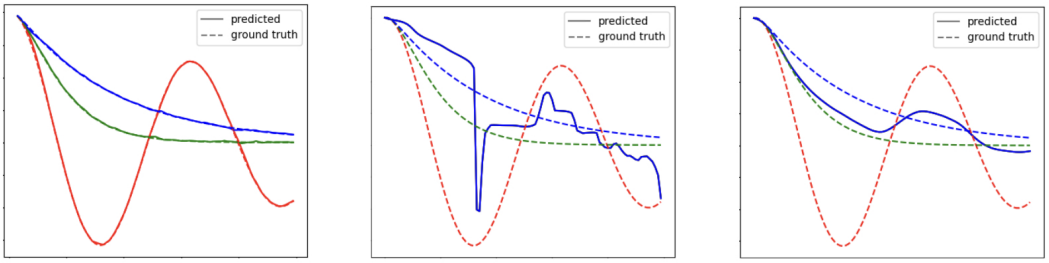

Figure 2: 1D harmonic oscillator overfitting experiment results. **Left:** TFM-ODE (ours) with memory = 3. **Middle:** TFM-ODE (ours) without memory. **Right:** Aligned FM [Liu et al., 2023a, Somnath et al., 2023].

4 Experimental Results

In this section we empirically evaluate the performance of the trajectory flow matching objective in terms of time series modeling error, but also uncertainty quantification. We also evaluate a variety of simulation-based and simulation-free methods including both stochastic and deterministic methods. Stochastic methods are in general more difficult to fit, but can be used to better model uncertainty and variance. Further experimental details can be found in §B. Experiments were run on a computing cluster with a heterogenous cluster of NVIDIA RTX8000, V100, A40, and A100 GPUs for approximately 24,000 GPU hours. Individual training runs require approximately one gpu day.

Baselines In addition to different ablations of trajectory flow matching, we also evaluate NeuralODE [Chen et al., 2018], NeuralSDE [Li et al., 2020, Kidger et al., 2021b, Kidger, 2022], Latent NeuralODE [Rubanova et al., 2019], and an aligned flow matching method (Aligned FM) [Liu et al., 2023a, Somnath et al., 2023] where the couplings are sampled according to the ground truth coupling during training.

Metrics We primarily make use of two metrics. The average mean-squared-error (Mean MSE) over left out time series to measure the time series modeling error defined as

$$\text{MSE}(\hat{x}, x) = \frac{1}{T-1} \sum_{t \in [2,T]} \|\hat{x}_t - x_t\|_2^2, \tag{14}$$

where $\hat{x}$ and x are the predicted and true trajectories respectively. We also use the *maximum mean discrepancy* with a radial basis function kernel (RBF MMD) which measures how well the distribution over next observation is modelled by comparing the predicted distribution to the distribution over next states in the ground truth trajectory. Specifically we compute:

$$\text{RBF-MMD}(\theta, \hat{x}, x) := \frac{1}{T-1} \sum_{t \in [2,T]} \text{MMD}(\hat{\Delta}_t, \Delta_t) \tag{15}$$

where $\hat{\Delta}_t = \hat{x}_t - x_{t-1}$, $\Delta_t = x_t - x_{t-1}$, and $\hat{x}_t := \int_{s=t-1}^{t} f_\theta(s, x_s)ds + g_\theta(t, x_s)dW_s$ is a set of samples from the model prediction at time t.

4.1 Exploring coupling preservation with 1D harmonic oscillators

We begin by evaluating how trajectory flow matching performs in a simple one dimensional setting of harmonic oscillators. We show that vanilla conditional flow and bridge matching [Liu et al., 2023c,b, Albergo and Vanden-Eijnden, 2023], specifically aligned approaches [Somnath et al., 2023, Liu et al., 2023a] are unable to preserve the coupling even in a simple one dimensional setting. However, augmented with our trajectory flow matching approach, and specifically using **(A2)**, which includes information on previous observations, the model is able to fit the harmonic oscillator dataset well.

The harmonic oscillator dataset consists of one-dimensional oscillatory trajectories from a damped harmonic oscillator, with each trajectory distinguished by a unique damping coefficient c. Specifically we sample trajectories x from:

$$x_i = x_{i-1} + v_{i-1}(t_i - t_{i-1}); \quad x_0 = 1 \tag{16}$$

where v is the velocity of the oscillator updated by

$$v_i = v_{i-1} + \left(-\frac{c}{m}v_{i-1} - \frac{k}{m}x_{i-1}\right)(t_i - t_{i-1}); \quad v_0 = 0 \tag{17}$$

with $t_i = 0.1 \cdot i$ for $i = 0, 1, 2, \ldots, 99$, spring constant $k = 1$, and mass $m = 1$.

As c increases, the trajectories evolve from underdamped scenarios with prolonged oscillations to critically and overdamped states where the oscillator quickly stabilizes. This leads to intersecting trajectories due to frequency and phase differences, despite their shared starting point. We perform overfitting experiments on three trajectories generated by varying c.

As shown in Figure 2, models without history information are unable to distinguish between the three crossing trajectories that share the same starting point, resulting in overlapping predictions. In contrast, TFM-ODE that incorporates three previous observations is able to fit the crossing trajectories with high accuracy, with the predicted trajectories almost completely overlapping the ground truth. This is because the dataset with satisfies **(A2)** with $h = 4$ (TFM-ODE), but not $h = 0$ (TFM-ODE no memory and Aligned FM).

4.2 Experiments on clinical datasets

Next we compared the performance of TFM and TFM-ODE with the current SDE and ODE baselines, respectively, for modeling real-world patient trajectories formed with heart rate and mean arterial blood pressure measurements within the first 24 hours of admission across three different datasets. These are clinical measurements that are taken most frequently and used to evaluate the hemodynamic status of patients, a key indicator of disease severity. Additionally, we evaluated our models against flow matching on these datasets, each with distinct characteristics, to assess their ability to generalize across different distributions. A full description of the datasets are available in Appendix B.2 with the publicly available datasets used under The PhysioNet Credentialed Health Data License Version 1.5.0 and the EHR dataset with local institutional IRB approval:

- **ICU Sepsis:** a subset of the eICU Collaborative Research Database v2.0 [Pollard et al., 2019] of patients admitted with sepsis as the primary diagnosis
- **ICU Cardiac Arrest:** a subset of the eICU Collaborative Research Database v2.0 [Pollard et al., 2019] of patients at risk for cardiac arrest
- **ICU GIB:** a subset of the Medical Information Mart for Intensive Care III [Johnson et al., 2016] of patients with gastrointestinal bleeding as the primary diagnosis
- **ED GIB:** patients presenting with signs and symptoms of acute gastrointestinal bleeding to the emergency department of a large tertiary care academic health system

4.2.1 Prediction accuracy and precision: TFM and TFM-ODE

TFM-ODE yields more accurate trajectory prediction Across the three datasets TFM-ODE outperformed the baseline models by 15% to 20%, as seen in table 1. We noticed that TFM has a similar performance as TFM-ODE. In one case TFM outperformed the non-stochastic TFM-ODE, as seen in the ICU GIB dataset. For ICU sepsis, the performance improvement from the baseline is the most significant, around 83%. This coincides with the ICU sepsis dataset having the most amount of measurement per trajectory. The improvement is seen in both TFM and TFM-ODE, possibly indicating they are able to learn better given more data, resulting in a more precise flow. Not formally measured, we noted that given the same time constraint, FM based models were significantly faster and often finished training before the time limit.

TFM yields better uncertainty prediction Though TFM-ODE had lower test MSE for two out of three times, TFM yielded better uncertainty prediction overall, as seen in table 2. Notably, TFM also had less variance in the uncertainty prediction than TFM-ODE. A plausible explanation in this case is a sacrifice in bias that subsequently decreases the variance for the stochastic implementation, reflecting the bias-variance trade off. Sampled graphs of TFM can be seen in figure 3. It is notable that the model is able to detect the measurement uncertainty at certain timepoints, matching the increase in amplitude of oscillation in patient trajectories.

4.2.2 Trajectory Variance Distribution Comparison

TFM trajectories accurately match the noise distribution in the data TFM is able to match the noise distribution in addition to the overall trajectory shape, which is useful in settings where

Table 1: Mean ± Std. deviation MSE ($\times 10^{-3}$) by models and datasets. Split into deterministic (top) and stochastic models (bottom). Top performing model for each setting and dataset in **bold**.

	ICU Sepsis	ICU Cardiac Arrest	ICU GIB	ED GIB
NeuralODE	4.776 ± 0.000	6.153 ± 0.000	3.170 ± 0.000	10.859 ± 0.000
FM baseline ODE	4.671 ± 0.791	10.207± 1.076	118.439 ± 17.947	11.923 ± 1.123
LatentODE RNN	61.806 ± 46.573	386.190 ± 558.140	422.886 ± 431.954	980.228±1032.393
TFM-ODE (ours)	**0.793 ± 0.017**	2.762 ± 0.021	2.673 ± 0.069	**8.245 ± 0.495**
NeuralSDE	4.747 ± 0.000	3.250 ± 0.024	3.186 ± 0.000	10.850 ± 0.043
TFM (ours)	0.796 ± 0.026	**2.755 ± 0.015**	**2.596 ± 0.079**	8.613 ± 0.260

Table 2: Uncertainty test MSE loss for TFM-ODE and TFM with two different ICU datasets.

	ICU sepsis	ICU Cardiac Arrest	ICU GIB
TFM-ODE	1.039 ± 0.1645	0.970 ± 0.1426	0.9843 ± 0.2233
TFM	**0.724 ± 0.0072**	**0.636 ± 0.0024**	**0.605 ± 0.0137**

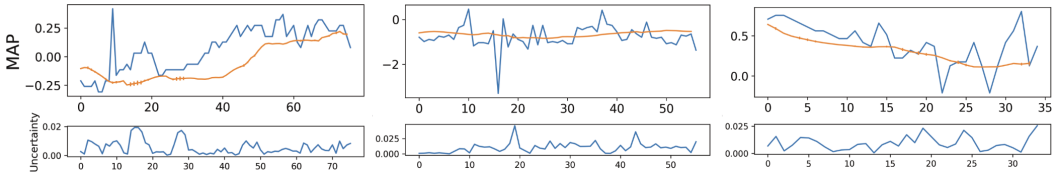

Figure 3: Three samples from predicted trajectory and uncertainty on ICU GIB test set. **Top**: Predicted (orange) and the ground truth (blue) mean arterial pressure (MAP). **Bottom**: The absolute value of the uncertainty predicted by TFM.

data has high stochasticity. We compared our models to NeuralODE and NeuralSDE in matching the variance in neighboring data points, seen in table 3. We verify that between the baseline NeuralSDE and NeuralODE, NeuralSDE has a lower MMD and is better able to match data points. We find in ICU GIB and ED GIB datasets, TFM outperforms both in matching the variance in data. Notably, the performance pattern is reversed for the MMD metrics and mean MSE metrics with respect to TFM and TFM-ODE where better MSE leads to worse MMD and vice versa. As such, this further confirms the bias-variance trade-off for both TFM and TFM-ODE implementation.

Table 3: Data variance MMD for by models and datasets. Split into deterministic models (top) and stochastic models (bottom). Top performing model for each setting and dataset in **bold**.

	ICU Sepsis	ICU Cardiac Arrest	ICU GIB	ED GIB
NeuralODE	1.988 ± 0.000	2.246 ± 0.000	2.090 ± 0.000	2.192 ± 0.000
TFM-ODE (ours)	**1.172 ± 0.017**	1.295 ± 0.006	1.087 ± 0.02	1.063 ± 0.031
NeuralSDE	1.212 ± 0.000	3.261 ± 0.020	1.332 ± 0.000	1.465 ± 0.122
TFM (ours)	1.199 ± 0.006	**0.993 ± 0.003**	**0.844 ± 0.013**	**0.717 ± 0.016**

4.2.3 Ablation Study

We performed ablation studies on TFM and TFM-ODE to attribute importance of various model components contributing to the performance, as seen in table 4. We examined three aspects of the model, two of which were part of our main contributions: uncertainty prediction and memory. We also ablate the model hidden dimension width to infer its potential in scaling effect.

TFM and TFM-ODE performance scales with model size In contrast to Neural DE based models, TFM and TFM-ODE exhibit scaling effect, in which the model performance becomes better with a larger hidden dimension. This has been observed in previous flow matching models in image generation [Tong et al., 2024]. This may pave the way for further improvements from larger models.

Table 4: Mean MSE ($\times 10^{-3}$) by ablated versions of TFM, TFM-ODE, and datasets.

	Uncertainty Prediction	Memory	Hidden Size	ICU Sepsis	ICU Cardiac Arrest	ICU GIB	ED GIB
TFM-ODE	✓	✓	256	**0.793 ± 0.017**	2.762 ± 0.017	2.673 ± 0.069	8.245 ± 0.495
		✓	256	1.170 ± 0.014	2.759 ± 0.015	3.097 ± 0.054	8.659 ± 0.429
			256	1.555 ± 0.122	3.242 ± 0.050	2.981 ± 0.161	**6.381 ± 0.451**
			64	1.936 ± 0.262	3.244± 0.025	4.003 ± 0.347	11.253± 4.597
TFM	✓	✓	256	0.796 ± 0.026	**2.596 ± 0.079**	**2.762 ± 0.021**	8.613 ± 0.260
		✓	256	0.816 ± 0.031	2.778 ± 0.021	2.754 ± 0.095	8.600 ± 0.389
			64	1.965 ± 0.289	3.271 ± 0.031	4.037 ± 0.314	7.549 ± 0.737

Uncertainty improves performance of trajectory prediction For TFM and TFM-ODE, the flow network used to learn the uncertainty σ_{x_t} is separate from the flow network learning x_t. The loss function of the network learning x_t is independent of uncertainty flow network. Therefore, it was unexpected that taking away the uncertainty prediction would result in increased MSE test loss for learning x_t. This implies further a process in the synergistic effects between x_t flow and σ_{x_t} flow.

Trajectory memory may improve performance in high frequency measurement settings We conditioned the model based on a sliding window of trajectory history to disentangle data points that otherwise look indistinguishable to FM models. This improved the interpolation performance in the ICU Sepsis and ICU GIB dataset. Notably, this modification did not improve performance the ED GIB dataset, which could be due to shorter trajectories for patients and lower measurement frequency in the defined time period. This may also be explained by the decreased severity of disease in the ED compared to the ICU. Adding memory as a condition may be more suitable for patients whose clinical trajectories have a higher frequency of measurements.

5 Related Work

Continuous-time neural network architectures have outperformed traditional RNN methods in modeling irregularly sampled clinical time series to optimize interpolation and extrapolation. Neural ODE with latent representations of trajectories [Rubanova et al., 2019] outperformed RNN-based approaches [Lipton et al., 2016, Che et al., 2018, Cao et al., 2018, Rajkomar et al., 2018] for interpolation while providing explicit uncertainty estimates about latent states. More recently, Neural SDEs appear to outperform LSTM [Hochreiter and Schmidhuber, 1997], Neural ODE [Chen et al., 2018, De Brouwer et al., 2019, Dupont et al., 2019, Lechner and Hasani, 2020], and attention-based [Shukla and Marlin, 2021, Lee et al., 2022] approaches in interpolation performance while natively handling uncertainty using drift and diffusion terms [Oh et al., 2024].

Discrete-time approaches offer an alternative to our continuous-time model model transformers utilize a discrete-time representation with a sequential processing [Gao et al., 2024, Nie et al., 2023, Woo et al., 2024, Ansari et al., 2024, Dong et al., 2024, Garza and Mergenthaler-Canseco, 2023, Das et al., 2024, Liu et al., 2024, Kuvshinova et al., 2024] models for traditional time series modeling. Adaptations to the baseline transformer includes structuring observations into text with finetuning [Zhang et al., 2023, Zhou et al., 2023], without finetuning [Xue and Salim, 2024, Gruver et al., 2023], or using autoregressive model vision transformers to model unevenly spaced time series data by converting time series into images [Li et al., 2023].

Continuous-time systems are of great interest for learning causal representations using assumptions by using observations to directly modify the system state [De Brouwer et al., 2022, Jia and Benson, 2019]. Variations include intervention modeling with separate ODEs for interventions and outcome processes [Gwak et al., 2020], using liquid time-constant networks [Hasani et al., 2021, Vorbach et al., 2021], or modeling treatment effects with either one [Bellot and van der Schaar, 2021] or multiple interventions [Seedat et al., 2022]. The importance of accounting for external interventions is a particular challenge in clinical data, where external interventions (change in environment due to treatment decisions or clinical context such as ED or ICU) are common in clinical data trajectories.

6 Conclusion

In this work we present Trajectory Flow Matching, a simulation-free training algorithm for neural differential equation models. We show when trajectory flow matching is valid theoretically, then demonstrate its usefulness empirically in a clinical setting. The ability to model the underlying

continuous physiologic processes during critical illness using irregular, sparsely sampled, and noisy data has the potential for broad impacts in care settings such as the emergency department or ICU. These models could be used to improve clinical decision making, inform monitoring strategies, and optimize resource allocation by identifying which patients are likely to deteriorate or recover. These use cases will require thorough prospective validation and calibration for specific clinical outcomes, for example using the likelihood of a patient crossing a specific heart rate or blood pressure threshold for decisions on level of care (ICU versus inpatient floors) or specific interventions such as transfusions. In these applications, it will be important to assess and control for bias that may be present due to which patient subpopulations are present in training data.

Limitations Limitations of the method includes the selective utility of integrating memory in clinical settings with high measurement frequency and no current capacity for estimating causal representations, though this will be an important future research direction. Potential harms include the following: erroneous predictions that either results in delayed care or overutilization of the health system. Accurate trajectory predictions have the potential to inform clinical decision-making regarding the appropriate level of care, leading to more timely and appropriate interventions.

Future work We hope to extend our method to cover other types of time series that have periodicity in the components, potentially incorporating Fourier transform [Li et al., 2021] and Physics-Inspired Neural Networks (PINN). Since interpretability is an important factor for clinical reliability, we are developing methods to further elucidate key components affecting the prediction. As well, we hope to incorporate functional flow matching for fully continuous setting [Kerrigan et al., 2024].

7 Broader Impact

Our work extends flow matching into the domain of time series modeling, demonstrating a specific instance of clinical time series prediction. In contrast to the large transformer-based models, our method has fewer in parameters and less training time needed. Notably, it scales well with parameters. As well, our parameterization on Stochastic Differential Equations (SDE) allow faster training time than traditional SDE integration.

Accurate timeseries modeling in healthcare has the potential for significant benefits, but also introduces risks. Benefits that could be derived from more accurate prediction of clinical courses include improved treatment decisions, resource allocation, as well as more informative discussions of prognosis with patients or family members. Risks may come from inaccuracies in predictions which could lead to harms by biasing decision making of clinical teams. In the general case of false negative prediction (prediction of trajectories with falsely favorable outcomes) this may lead to undertreatment and in the case of false positive prediction (prediction of trajectories with incorrect detrimental outcomes) or overtreating patients. These inaccuracies may also propagate biases in training data.

To move towards broad impact in the clinical domain, this work will require validation and bias estimates. Furthermore, models deployed in domains with high-stakes prediction require interpretability, which can help identify biases, miscalibration, discordance with domain knowledge, as well as build trust with teams using predictions from the model. At this time, flow-based methods have limited tools for interpretability, and we recognize this as a gap in need of future work.

Acknowledgements

The authors would like to thank Mathieu Blanchette for useful comments on early versions of this manuscript. We are also grateful to the anonymous reviewers for suggesting numerous improvements.

The authors acknowledge funding from the National Institutes of Health, UNIQUE, CIFAR, NSERC, Intel, and Samsung. The research was enabled in part by computational resources provided by the Digital Research Alliance of Canada (`https://alliancecan.ca`), Mila (`https://mila.quebec`), Yale School of Medicine and NVIDIA.

References

M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=li7qeBbCR1t`.

M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *CoRR*, abs/2303.08797, 2023. URL https://doi.org/10.48550/arXiv.2303.08797.

A. Allam, S. Feuerriegel, M. Rebhan, and M. Krauthammer. Analyzing patient trajectories with artificial intelligence. *J Med Internet Res*, 23(12):e29812, Dec 2021. ISSN 1438-8871. doi: 10.2196/29812. URL https://www.jmir.org/2021/12/e29812.

A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. Gordon Wilson, M. Bohlke-Schneider, and Y. Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

A. Bellot and M. van der Schaar. Policy analysis using synthetic controls in continuous-time. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 759–768. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/bellot21a.html.

W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li. Brits: Bidirectional recurrent imputation for time series. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf.

Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1), 2018. doi: 10.1038/s41598-018-24271-9.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

T. Chen and C. 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.

M. M. Churpek, T. C. Yuen, S. Y. Park, D. O. Meltzer, J. B. Hall, and D. P. Edelson. Derivation of a cardiac arrest prediction model using ward vital signs. *Critical Care Medicine*, 2012.

A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jn2iTJas6h.

E. De Brouwer, J. Simm, A. Arany, and Y. Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/455cb2657aaa59e32fad80cb0b65b9dc-Paper.pdf.

E. De Brouwer, J. Gonzalez, and S. Hyland. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4705–4722. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/de-brouwer22a.html.

J. Dong, H. Wu, Y. Wang, Y.-Z. Qiu, L. Zhang, J. Wang, and M. Long. Timesiam: A pre-training framework for siamese time-series modeling. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=wrTzLoqbCg.

E. Dupont, A. Doucet, and Y. W. Teh. Augmented neural odes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/21be9a4bd4f81549a9d1d241981cec3c-Paper.pdf.

S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik. Units: Building a unified time series model. *arXiv*, 2024. URL https://arxiv.org/pdf/2403.00131.pdf.

A. Garza and M. Mergenthaler-Canseco. Timegpt-1, 2023.

N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson. Large language models are zero-shot time series forecasters. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19622–19635. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3eb7ca52e8207697361b2c0fb3926511-Paper-Conference.pdf.

D. Gwak, G. Sim, M. Poli, S. Massaroli, J. Choo, and E. Choi. Neural Ordinary Differential Equations for Intervention Modeling. *arXiv e-prints*, art. arXiv:2010.08304, Oct. 2020. doi: 10.48550/arXiv.2010.08304.

R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu. Liquid time-constant networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7657–7666, May 2021. doi: 10.1609/aaai.v35i9.16936. URL https://ojs.aaai.org/index.php/AAAI/article/view/16936.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

J. Jia and A. R. Benson. Neural jump stochastic differential equations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/59b1deff341edb0b76ace57820cef237-Paper.pdf.

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

G. Kerrigan, G. Migliorini, and P. Smyth. Functional flow matching. In S. Dasgupta, S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3934–3942. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/kerrigan24a.html.

P. Kidger. On neural differential equations, 2022. URL https://arxiv.org/abs/2202.02435.

P. Kidger, J. Foster, X. Li, H. Oberhauser, and T. Lyons. Neural sdes as infinite-dimensional gans. In *International conference on machine learning*. PMLR, 2021a.

P. Kidger, J. Foster, X. C. Li, and T. Lyons. Efficient and accurate gradients for neural sdes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18747–18761. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9ba196c7a6e89eafd0954de80fc1b224-Paper.pdf.

K. Kuvshinova, O. Tsymboi, A. Kostromina, D. Simakov, and E. Kovtun. Towards foundation time series model: To synthesize or not to synthesize? *arXiv preprint arXiv:2403.02534*, 2024.

S. Lahlou, M. Jain, H. Nekoei, V. I. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=eGLdVRvvfQ. Expert Certification.

M. Lechner and R. Hasani. Learning long-term dependencies in irregularly-sampled time series. *arXiv preprint arXiv:2006.04418*, 2020.

Y. Lee, E. Jun, J. Choi, and H.-I. Suk. Multi-view integrative attention-based deep representation learning for irregular clinical time-series data. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4270–4280, 2022. doi: 10.1109/JBHI.2022.3172549.

X. Li, T.-K. L. Wong, R. T. Q. Chen, and D. K. Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In C. Zhang, F. Ruiz, T. Bui, A. B. Dieng, and D. Liang, editors, *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–28. PMLR, 08 Dec 2020. URL https://proceedings.mlr.press/v118/li20a.html.

Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Z. Li, S. Li, and X. Yan. Time series as images: Vision transformer for irregularly sampled time series. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 49187–49204. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a17c1eb808cf012065e9db47b7ca80d-Paper-Conference.pdf.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Z. C. Lipton, D. Kale, and R. Wetzel. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 253–270, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR. URL https://proceedings.mlr.press/v56/Lipton16.html.

G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar. I2sb: image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *The Eleventh International Conference on Learning Representations (ICLR)*, 2023b. URL https://par.nsf.gov/biblio/10445517.

X. Liu, L. Wu, M. Ye, and qiang liu. Learning diffusion bridges on constrained domains. In *The Eleventh International Conference on Learning Representations*, 2023c. URL https://openreview.net/forum?id=WH1yCa0TbB.

Y. Liu, H. Zhang, C. Li, X. Huang, J. Wang, and M. Long. Timer: Transformers for Time Series Analysis at Scale. *arXiv e-prints*, art. arXiv:2402.02368, Feb. 2024. doi: 10.48550/arXiv.2402.02368.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Y. Oh, D. Lim, and S. Kim. Stable neural stochastic differential equations in analyzing irregular time series data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4VIgNuQ1pY.

T. Pollard, A. Johnson, J. Raffa, L. A. Celi, O. Badawi, and R. Mark. eicu collaborative research database (version 2.0), 2019. URL https://doi.org/10.13026/C2WM1R.

A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Q. Chen. Multisample flow matching: straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1.

Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/42a6845a557bef704ad8ac9cb4461d43-Paper.pdf.

N. Seedat, F. Imrie, A. Bellot, Z. Qian, and M. van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19497–19521. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/seedat22b.html.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion schrödinger bridge matching. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 62183–62223. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c428adf74782c2092d254329b6b02482-Paper-Conference.pdf.

S. N. Shukla and B. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

V. R. Somnath, M. Pariset, Y.-P. Hsieh, M. R. Martinez, A. Krause, and C. Bunne. Aligned diffusion Schrödinger bridges. In R. J. Evans and I. Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1985–1995. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/somnath23a.html.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

A. Tong, K. FATRAS, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

C. Vorbach, R. Hasani, A. Amini, M. Lechner, and D. Rus. Causal navigation by continuous-time neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12425–12440. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/67ba02d73c54f0b83c05507b7fb7267f-Paper.pdf.

G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Yd8eHMY1wz.

H. Xue and F. D. Salim. PromptCast: A New Prompt-Based Learning Paradigm for Time Series Forecasting . *IEEE Transactions on Knowledge & Data Engineering*, 36(11):6851–6864, Nov. 2024. ISSN 1558-2191. doi: 10.1109/TKDE.2023.3342137. URL https://doi.ieeecomputersociety.org/10.1109/TKDE.2023.3342137.

Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

T. Zhou, P. Niu, x. wang, L. Sun, and R. Jin. One fits all: Power general time series analysis by pretrained lm. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43322–43355. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/86c17de05579cde52025f9984e6e2ebb-Paper-Conference.pdf.

J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.

A Proof of theorems

We first prove a Lemma which shows TFM learns valid flows between distributions with the target prediction reparameterization trick.

Lemma A.1. *If $p_t(x) > 0$, δ_{data} is Lipschitz continuous for all $x \in \mathbb{R}^d$ and $t \in [0, 1]$, $\mathcal{L}_{FM}$ and $\mathcal{L}_{TFM}$ are equal,*

$$\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{TFM}(\theta)$$

Proof. This proof is a simple extension of Lipman et al. [2023], Tong et al. [2024] which proved $\mathcal{L}_{CFM}$ and $\mathcal{L}_{FM}$ are equal under similar constraint.

Given $\delta_{data} = t_1 - t_0$, we have $u_t(x) = \frac{x_1 - x_0}{\delta_{data}}$ where t_0 is the previous time in the time series, and t_1 is the current time for inference. For the time series data, we are assuming it to be Lipschitz continuous there exist $L \geq 0$ such that for all $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| \leq L\|x - y\|$.

$$\nabla_\theta \mathbb{E}_{p_t(x)} \|v_\theta(t,x) - u_t(x)\|^2 = \mathbb{E}_{t,q(z),p_t(x|z)} \frac{1}{(1-t)^2} \|\hat{x}_\theta^1(t,x) - x_1\|^2 \tag{18}$$

$$= \nabla_\theta \mathbb{E}_{t,q(z),p_t(x|z)} \frac{1}{(1-t)^2} \left(\|\hat{x}_\theta^1(t,x)\|^2 - 2\left\langle \hat{x}_\theta^1(t,x), x_1 \right\rangle + x_1^2 \right) \tag{19}$$

$$= \nabla_\theta \mathbb{E}_{t,q(z),p_t(x|z)} \left(\frac{1}{(1-t)^2} \|\hat{x}_\theta^1(t,x)\|^2 - 2\left\langle \hat{x}_\theta^1(t,x), x_1 \right\rangle \right) \tag{20}$$

By bilinearity of the 2-norm and since x_1 is independent of θ. Next,

$$\begin{aligned}
\mathbb{E}_{p_t(x)} \frac{1}{(1-t)^2} \|\hat{x}_\theta^1(t,x)\|^2 &= \int \|\hat{x}_\theta^1(t,x)\|^2 p_t(x) dx \\
&= \iint \|\hat{x}_\theta^1(t,x)\|^2 p_t(x|z) q(z) dz dx \\
&= \mathbb{E}_{q(z),p_t(x|z)} \|\hat{x}_\theta^1(t,x)\|^2
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}_{p_t(x)} \left\langle \hat{x}_\theta^1(t,x), x_1 \right\rangle &= \int \left\langle \hat{x}_\theta^1(t,x), \frac{\int x_1 p_t(x|z) q(z) dz}{p_t(x)} \right\rangle p_t(x) dx \\
&= \int \left\langle \hat{x}_\theta^1(t,x), \int x_1 p_t(x|z) q(z) dz \right\rangle dx \\
&= \iint \left\langle \hat{x}_\theta^1(t,x), x_1 \right\rangle p_t(x|z) q(z) dz dx \\
&= \mathbb{E}_{q(z),p_t(x|z)} \left\langle \hat{x}_\theta^1(t,x), x_1 \right\rangle
\end{aligned}$$

Where we first substitute then change the order of integration for the final equality. Since at all times t the gradients of $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{TFM}$ are equal, $\nabla_\theta \mathcal{L}_{\text{FM}}(\theta) = \nabla_\theta \mathcal{L}_{TFM}$

by substitution.

$$\mathcal{L}\mathbb{E}_{t,q(z),p_t(x|z)} \|v_\theta(t,x) - u_t(x|z)\|^2 = \mathbb{E}_{t,q(z),p_t(x|z)} \frac{1}{(\lceil t \rceil - t)^2} \|\hat{x}_\theta^{\lceil t \rceil}(t,x) - x^{\lceil t \rceil}\|^2 \tag{21}$$

$$\mathbb{E}_{t,q(z),p_t(x|z)} \|v_\theta(t,x) - u_t(x|z)\|^2 = \mathbb{E}_{t,q(z),p_t(x|z)} \left\| \frac{\hat{x}_\theta^{\lceil t \rceil}(t,x) - x}{\lceil t \rceil - t} - \frac{x^{\lceil t \rceil} - x}{\lceil t \rceil - t} \right\|^2 \tag{22}$$

$$= \mathbb{E}_{t,q(z),p_t(x|z)} \frac{1}{(\lceil t \rceil - t)^2} \|\hat{x}_\theta^{\lceil t \rceil}(t,x) - x^{\lceil t \rceil}\|^2 \tag{23}$$

□

Lemma 3.1 *The SDE $dx_t = u_t(x|z)dt + \sigma^2 dW_t$ where u_t is defined in eq. 9 generates $p_t(x|z)$ in eq. 8 with initial condition $p_0 := \delta_{x_1}$ where δ is the Dirac delta function.*

Proof. For simplicity of notation we first show the case where $\lceil t \rceil = 1$.

$$dx_t = u_t(x|z)dt + \sigma^2 dW_t = \frac{1 - x_t}{1 - t}dt + \sigma^2 dW_t \tag{24}$$

which is equivalent to the d dimensional Brownian bridge which has marginal

$$\mathcal{N}((1-t)x_0 + tx_1, \sigma^2 t(1-t)) \tag{25}$$

which completes the proof for $\lceil t \rceil = 1$. □

Proposition 3.2 (Coupling Preservation) *Under mild regulatory criteria on $u_t(\cdot|z)$, p_t, and q, if*

$$\mathbb{E}_{t\sim\mathcal{U}(0,T),z\sim q(z),c\sim q(c|z),x_t\sim p_t(x_t|z)} \|u_t(x_t|z,c) - u_t(x_t|c)\|_2^2 = 0$$

and $z, q(z), p_t(x|z)$, and $u_t(x|z)$ are as defined in eqs. 6-9 then $\Pi(u)^\star = \Pi^\star(x_{1:T})$.

Proof. We prove the deterministic case with $T = 1$. The extensions to stochastic and $T > 1$ are evident. The couplings are equal if the marginal vector field $u_t(x_t|c) = u_t(x_t|z,c)$ everywhere as the coupling is governed by the push forward flows $\phi(x_0, c) = \int_0^1 u_t(x_t|c)dt$, and $\phi(x_0, c, z) = \int_0^1 (u_t(x_t|z,c)$. If

$$\mathbb{E}_{t\sim\mathcal{U}(0,T),z\sim q(z),c\sim q(c|z),x_t\sim p_t(x_t|z)} \|u_t(x_t|z,c) - u_t(x_t|c)\|_2^2 = 0$$

then $\phi(x_0, c, z) = \phi(x_0, c)$ for all x_0 and therefore the couplings of the optimal map are equivalent. We note that this requires exchange of integrals under the same conditions as **Lemma A.1**. □

Next we show how **(A1)-(A3)** satisfy Prop. 3.2.

- **(A1)** When $c = x_0$ and there exists $T : \mathcal{X} \to \mathcal{X}$ such that $T(x_0) = x_1$ if and only if $\Pi^\star(x_0, x_1)$. We note that this is equivalent to asserting the existence of a Monge map $T^\star$ for the coupling $\Pi^\star$.
  In the two timepoint case, $c = x_0$ is sufficient as long as there aren't two trajectories that have the same x_0 but different x_1s. Conditioning on this way ensures the conditions of of Prop. 3.2 as the uniqueness property ensures the uniqueness of $u_t(x_t|c)$.
- **(A2)** There exist no two trajectories x^i, x^j such that $x_t^i = x_t^j$ for $h + 1$ consecutive observations.
  In this case notice that this is simply a multi-timepoint extension of **A1** to $c = x_{t-h-1:t-1}$, i.e. conditioned on a history of length h. If this is the case then the same reasoning as **A1** applies.
- **(A3)** Trajectories are associated with unique conditional vectors c independent of t.
  This satisfies Prop 3.2 by definition.

Proposition 3.3 There exists a scaling function $c(t) : \mathbb{R}_+ \to \mathbb{R}$ such that $\mathcal{L}_{\text{target}}(\theta) = c(t)\mathcal{L}_{\text{match}}(\theta)$.

Proof. We start with the matching loss.

$$\mathbb{E}_{t,q(z),p_t(x|z)} \|v_\theta(t,x) - u_t(x|z)\|^2 = \mathbb{E}_{t,q(z),p_t(x|z)} \frac{1}{(\lceil t \rceil - t)^2} \|\hat{x}_\theta^{\lceil t \rceil}(t,x) - x^{\lceil t \rceil}\|^2 \tag{26}$$

by substitution,

$$\mathbb{E}_{t,q(z),p_t(x|z)} \|v_\theta(t,x) - u_t(x|z)\|^2 = \mathbb{E}_{t,q(z),p_t(x|z)} \left\| \frac{\hat{x}_\theta^{\lceil t \rceil}(t,x) - x}{\lceil t \rceil - t} - \frac{x^{\lceil t \rceil} - x}{\lceil t \rceil - t} \right\|^2 \tag{27}$$

$$= \mathbb{E}_{t,q(z),p_t(x|z)} \frac{1}{(\lceil t \rceil - t)^2} \|\hat{x}_\theta^{\lceil t \rceil}(t,x) - x^{\lceil t \rceil}\|^2 \tag{28}$$

completing the proof. □

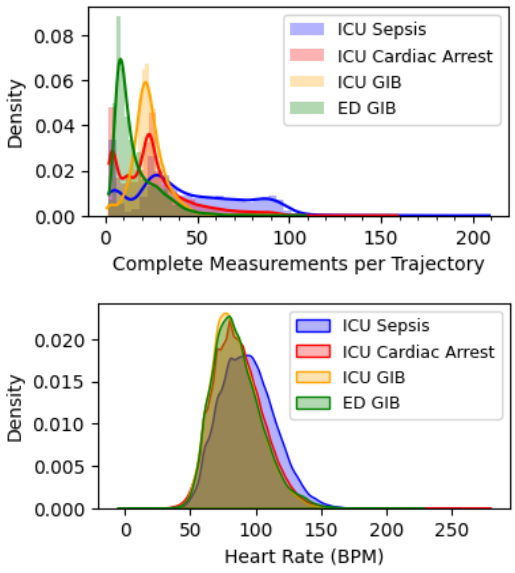

Figure 4: **Left**: Distribution of number of complete vital measurements per patient trajectory within the first 24 hours of admission in each clinical dataset. **Right**: Distribution of raw heart rate values in each clinical dataset.

B Experimental Details

B.1 1D Oscillators

The three oscillation trajectories correspond to $c = 0.25$ (the red trajectory in Figure 2), $c = 2$ (blue), and $c = 3.75$ (green). Before used as an input, t was scaled to between 0 and 1 by dividing by 10.

B.2 Clinical Datasets

B.2.1 Clinical Data Characteristics

In order to accurately model the perturbations in the physiologic signals (mean arterial pressure and heart rate) of the underlying patient states, we need to learn beyond the general trend of the data.

While the physiologic measurements themselves reflect patient status and drive clinical decision making, the degree of variation holds information that goes beyond the snapshot at a single time point. Our approach models the data distribution and stochasticity rather than just fitting the average trajectory. Other time-varying such as treatment conditions and non-time-varying covariates such as underlying disease states may also hold information that may impact the underlying state generating the physiologic signals. Our approach also incorporates this information to inform the trajectory modeling.

The data distribution in the ICU datasets reflect its status as the most resource-intensive clinical setting with increased measurement frequency and data distribution shift towards more abnormal physiologic values (Figure 4). The ED dataset reflects its status as the clinical setting focused on triaging patients, with sparser and physiologic measurements that fall in the normal range.

B.2.2 Clinical Data Preprocessing

For each clinical dataset, we modeled patient trajectories formed with heart rate and blood pressure measurements during the first 24 hours following admission. The timeline for each trajectory, originally in minutes, was scaled to a range between 0 and 1 by dividing by 1440. Additionally, heart rate and blood pressure values were z-score normalized to standardize the data.

Intensive Care Unit Sepsis (ICU Sepsis) Dataset The eICU Collaborative Research Database v2.0 [Pollard et al., 2019] is a database including deidentified information collected from over 200,000 patients in multiple intensive care units (ICUs) in the United States from 2014 to 2015. The ICU Sepsis Dataset was created by subsetting the eICU Database for 3362 patients with sepsis as the

primary admission diagnosis (2689 patients in training set, 336 in validation set, and 337 in test set). The following data fields were extracted: patient sex, age, heart rate, mean arterial pressure, norepinephrine dose and infusion rate, and a validated ICU score (APACHE-IV). Each patient's complete pair measurements of heart rate and mean arterial pressure over time form one trajectory to be modeled.

Norepinephrine infusion rates were calculated by converting drug doses or infusion rates to μg/kg/min, and where drug doses were not explicitly available, the dose was inferred from the free text given in the drug name. Start and end times for norepinephrine infusion were calculated by dividing the dose by the infusion rate. Where there appeared to be multiple infusions at the same time, the maximum infusion rate was taken as the infusion rate. As a conditional input to the models, the norepinephrine infusion doses are then scaled to between 0 and 1 by dividing by the maximum norepinephrine value in the dataset.

The APACHE-IV score, a validated critical care risk score, predicts individual patient mortality risk [Zimmerman et al., 2006]. In data preprocessing, we uses logistic regression of the score against binary hospital mortality data to generate a probability for each patient, serving as an additional input condition for models.

Intensive Care Unit Cardiac Arrest (ICU Cardiac Arrest) Dataset This dataset was extracted from the to eICU Collaborative Research Database v2.0 [Pollard et al., 2019] described above to reflect ICU patients at risk for cardiac arrest. This dataset excludes patients who presented with myocardial infarction (MI) and includes variables used in the Cardiac Arrest Risk Triage (CART) score [Churpek et al., 2012]: respiratory rate, heart rate, diastolic blood pressure, and age at the time of ICU admission. As an input to the model, the age was z-score normalized. 51671 patients were included in the training set, with 6459 patients each in the validation and test sets.

Intensive Care Unit Acute Gastrointestinal Bleeding (ICU GIB) Dataset The Medical Information Mart for Intensive Care III (MIMIC-III) critical care database contains data for over 40,000 patients in the Beth Israel Deaconess Medical Center from 2001 to 2012 requiring an ICU stay [Johnson et al., 2016]. We selected a cohort of 2602 ICU patients with the primary diagnosis of gastrointestinal bleeding to form the ICU GIB dataset, split into a training set of 2082 patients, and a validation set and a test set of 260 patients each. We extracted the following variables: age, sex, heart rate, systolic blood pressure, diastolic blood pressure, usage of vasopressor, usage of blood product, usage of packed red blood cells, and liver disease. Since the vasopressor and blood product usage are encoded as a binary value and may not represent actual infusion amount that are most likely decaying, we experimented with adding a Gaussian decay to them to use as conditional inputs. Likewise, trajectories to model consist of complete pairs of heart rate and mean arterial pressure (calculated from systolic blood pressure and diastolic blood pressure) measurements.

Emergency Department Acute Gastrointestinal Bleeding (ED GIB) Dataset This dataset reflects 3348 patients presenting with signs and symptoms of acute gastrointestinal bleeding to two hospital campuses in Yale New Haven Hospital between 2014 and 2018. The patients were split into a training set, a validation set, and a test set of 2636, 352, and 360 patients. Variables extracted include patient sex, age, heart rate, mean arterial pressure, initial measurements of 24 lab tests, and 17 pre-existing medical conditions as determined by ICD-10 codes. Like ICU Sepsis data, the trajectoires consist of complete pairs of heart rate and mean arterial pressure measurements.

Age, initial lab test measurements (three labs omitted due to missing data), and pre-existing medical conditions were used to train an XGBoost model [Chen and Guestrin, 2016] to predict the binary outcome variable indicating the need for hospital-based care. The resulting probabilities of requiring hospital-based care (outcome of 1) for each patient were then calculated using the trained model and used as conditional input to conditional models in experiments on this dataset.

Of note, the outcome variable was defined as 1 if a patient (1) requires red blood cell transfusion, (2) requires urgent intervention (endoscopic, interventional radiologic, or surgical) to stop bleeding or (3) all-cause 30-day mortality. Labs and medical conditions included in this dataset are listed below. Labs in bold were excluded from the XGBoost risk score calculation due to missing data.

- **Labs:** Sodium, Potassium, Chloride, Carbon Dioxide, Blood Urea Nitrogen, Creatinine, International Normalized Ratio, **Partial Thromboplastin Time**, White Blood Cell Count, Hemoglobin, Platelet Count, Hematocrit, Mean Corpuscular Volume, Mean Corpuscular Hemoglobin, Mean Corpuscular Hemoglobin Concentration, Red Cell Distribution Width,

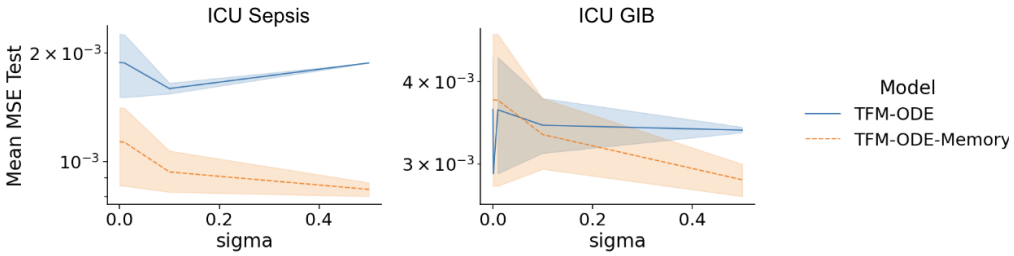

Figure 5: Sigma mean MSE comparison

Red Blood Cell Count, Aspartate Aminotransferase, Alanine Aminotransferase, Alkaline Phosphatase, Total Bilirubin, **Direct Bilirubin**, Albumin, **Lactate**.

- **Previous Medical Histories:** Charlson Comorbidity Index, Cerebrovascular Accident, Deep Vein Thrombosis, Pulmonary Embolism, Atrial Fibrillation, Upper Gastrointestinal Bleeding, Lower Gastrointestinal Bleeding, Unspecified Gastrointestinal Bleeding, Peptic Ulcer Disease, Helicobacter Pylori Infection, Coronary Artery Disease, Heart Failure, Hypertension, Type 2 Diabetes Mellitus, Chronic Kidney Disease, Alcohol Use Disorder, Cirrhosis.

B.3 Training

B.3.1 1D Oscillators

Since the trajectories in this dataset are deterministic and regularly sampled, we deployed only TFM-ODE and applied solely the $\mathcal{L}_{\text{match}}$ loss (i.e, no uncertainty or time predictive loss), as these methods sufficiently address the structured nature of the data to generate proof-of-concept results. The three models presented in Figure 2 all have hidden size of 256, σ of 0.1, trained under seed=0 with Adam optimizer with learning rate 1×10^{-3} for a maximum of 1000 epochs with early stopping (patience=3) monitoring validation loss.

B.3.2 Clinical Data

All the models for clinical data experiments are trained with Adam optimizer. A maximum training time and epochs are set to 48 hours and 300, with early stopping (patience=3) monitoring validation loss. All metrics reported were ran with 5 seeds (0,1,2,3,4) to ensure it is reproducible.

TFM, TFM-ODE, and ablations The TFM models were trained with learning rate 1×10^{-6} and had σ of 0.1. The complete models have hidden size of 256 and memory of 3, while ablation study with a hidden size of 64 and/or no memory was performed (Table 4). The noise parameter for the SDE implementation was set to 0.1 for ablations without $\mathcal{L}_{\text{uncertainty}}$. The hyperparameters $\sigma = 0.1$ and memory=3 for full models were selected through experiments with different values of σ and memory (Figure 5 and 6).

FM The FM baseline models were trained with learning rate 1×10^{-6}. All models had a hidden size of 64 with σ of 0.1.

Latent Neural ODE The latent Neural ODE models were trained with a learning rate of 1×10^{-3}. 100 GRU units were used for the encoder model and the number of latent dimensions was 2.

Baseline Neural SDE and Neural ODE Both baseline models were trained with learning rate 1×10^{-5} and had a hidden size of 64.

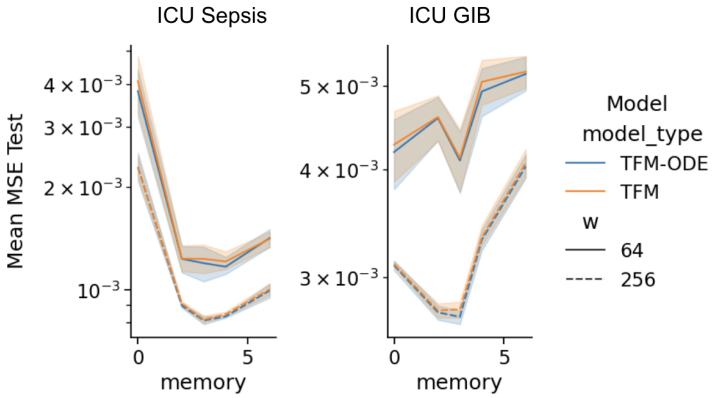

Figure 6: Memory Mean MSE comparison

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claim. The theoretical portions of the claim are supported in sections 2 and 3 of the main text and in appendix A; the claims on its performance are supported by section 4 of the main text. We discuss the limitations of our models context of the ablation study and in the related works section. See the next section for more details on the specifics on the discussion of limitations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limits of our current model depending on characteristics of the dataset. In the ablation study, we discuss the variable effect memory had across datasets of different data distributions and measurement frequency as a limitation to the current framework. We also provide context in the related work section that describes the limitation of our current framework in not being able to provide causal explanations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors

should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proof of theorems are included in section 3 and appendix A.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducibility is accomplished by providing a full description of the model in the main text in addition to providing publically available code at `https://github.com/nZhangx/TrajectoryFlowMatching`. The experimental data are available publicly online (ICU Sepsis, ICU GIB) or upon reasonable request (ED GIB) subject to institutional approval and HIPAA regulations. Data preprocessing steps are extensively outlined in Appendix B.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Anonymised code is available as supplementary material. The ICU Sepsis dataset is available publicly online via the eICU database and the ICU GIB dataset is available via the MIMIC-III database, whereas the ED GIB datasets are datasets which include identifiable information. A deidentified dataset may be made available upon reasonable request subject to guidelines set by the institutional review board and in accordance with HIPAA policy.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide an overview with the significance of each clinical dataset in the body of the paper. In the Appendix we have written detailed descriptions of each dataset, including the methodology for identifying the cohorts of patients, input variables extracted, number of patient encounters, summary statistics, as well as frequency distributions. We also detail the process of training and testing for all methods presented in the paper, from baseline to our novel method. We have also provided anonymized link to all code used for the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper includes a standard error of the mean for all the results reported in a numerical format. There are no figures with error bars for which the nature of error bar calculation would be relevant. There are no statistical tests that were perform which would make discussion of the validity of statistical tests relevant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the description of the compute resources utilized with available computing clusters detailed in the Experimental Results section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have all reviewed the NeurIPS Code of Ethics and striven to maintain and preserve anonymity.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In the conclusion section we discuss how improved modeling of Emergency Department and ICU physiology can lead to the positive societal impact of improving disease risk prediction which can enable medical teams to make better clinical decisions, provide patients and family members with increased information on the likely course of the illness, and improve resource allocation by identifying which patients may require costly or rare resources such a blood transfusions.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The current study proposes a predictive model for time-varying datasets and as such does not pose risks that are significantly higher in comparison to conventional models that fit data trends. There is no generative component in our model, and little risk that the results could be misused in a way that misleads or proves otherwise detrimental to the broader public. We therefore believe this item does not apply for this submission.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The publicly available datasets from MIMIC-III and eICU databases are properly credited, respected, mentioned and used under the PhysioNet Credentialed Health Data License Version 1.5.0. The ED GIB EHR data is owned by the authors.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: We will not be releasing new datasets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: While the data involves patient data, it does not include any experiments performed on humans but instead based on the use of retrospectively collected healthcare records generated as part of routine clinical care. Publicly available datasets were used in accordance in the regulations set out by the hosting institution, whereas collection of data from patients in non-publicly available datasets were performed in accordance with institutional review board guidance and HIPAA data protection regulations. As the data collection does not involve experiments on human subjects or crowdsourcing experiments, this item does not apply to our current study.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: We have obtained institutional IRB consent for the use of the ED GIB EHR data under an approved protocol that we are happy to provide upon request. MIMIC-III and eICU databases are pre-approved, de-identified, publicly available clinical data sources.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.