T-Phenotype: Discovering Phenotypes of Predictive Temporal Patterns in Disease Progression

Yuchao Qin
University of Cambridge, UK

Mihaela van der Schaar
University of Cambridge, UK
Alan Turing Institute, UK

Changhee Lee
Chung-Ang University, South Korea

Abstract

Clustering time-series data in healthcare is crucial for clinical phenotyping to understand patients' disease progression patterns and to design treatment guidelines tailored to homogeneous patient subgroups. While rich temporal dynamics enable the discovery of potential clusters beyond static correlations, two major challenges remain outstanding: i) discovery of predictive patterns from many potential temporal correlations in the multi-variate time-series data and ii) association of individual temporal patterns to the target label distribution that best characterizes the underlying clinical progression. To address such challenges, we develop a novel temporal clustering method, *T-Phenotype*, to discover phenotypes of predictive temporal patterns from labeled time-series data. We introduce an efficient representation learning approach in frequency domain that can encode variable-length, irregularly-sampled time-series into a unified representation space, which is then applied to identify various temporal patterns that potentially contribute to the target label using a new notion of path-based similarity. Throughout the experiments on synthetic and real-world datasets, we show that T-Phenotype achieves the best phenotype discovery performance over all the evaluated baselines. We further demonstrate the utility of T-Phenotype by uncovering clinically meaningful patient subgroups characterized by unique temporal patterns.

1 INTRODUCTION

Discovering predictive patterns of disease progression has been a long pursuit in healthcare. Clinicians have considered specific clinical (disease) status and the associated patterns as a *phenotype* to uncover the heterogeneity of diseases and to design therapeutic guidelines tailored to homogeneous subgroups (Hripcsak and Albers, 2013; Richesson et al., 2016). While rule-based phenotypes identified by domain experts have been widely used (Denny et al., 2013; Richesson et al., 2016), designing and validating such rules require tremendous effort. Unfortunately, disease progression can manifest through a broad spectrum of clinical factors, collected as a sequence of measurements in electronic health records (EHRs), that may vary greatly across individual patients. This makes it even more daunting for domain experts to transform such raw and complex clinical observations into clinically relevant and interpretable patterns.

Temporal clustering has been recently used as a data-driven framework for phenotyping to partition patients with sequences of observations into homogeneous subgroups. To discover different temporal patterns, traditional notions of similarity focus on either adjusting similarity measures (Zhang et al., 2019; Baytas et al., 2017) or finding low-dimensional representations (Ho et al., 2014; Giannoula et al., 2018) for longitudinal observations. These approaches are purely unsupervised and discard valuable information about the disease status that is often available in the clinical data. More recently, predictive clustering methods (Lee and van der Schaar, 2020; Lee et al., 2020; 2022; Aguiar et al., 2022) have introduced a new notion of similarity such that each cluster shares similar disease status to provide a better prognostic value. Despite the effort to understand temporal dynamics in their mutual context, these clustering methods fail to capture the full picture of disease progression as reflected by covariate trajectories of prognostic characteristics, i.e., temporal patterns associated with specific disease status. Figure 1 illustrates a pictorial depiction of the notion of phenotypes behind different temporal clustering methods.

Contribution. In this paper, we propose a novel temporal clustering method to correctly uncover predictive temporal

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

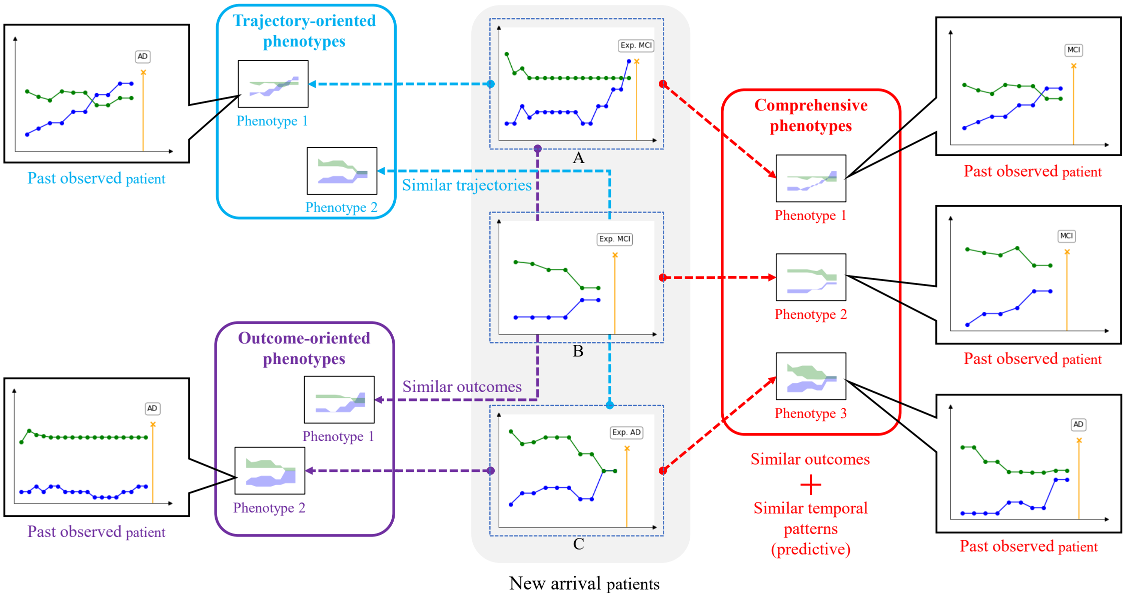


Figure 1: Different Notions of Temporal Phenotypes. Purely unsupervised clustering approaches focus on trajectory-oriented phenotypes (blue) and disregard the valuable information in patient outcomes. Predictive clustering methods aim at discovering outcome-oriented phenotypes (purple) which may not reflect the heterogeneity in patient trajectories despite the same diagnosis outcome. A desirable phenotyping method shall address both types of similarity and discover comprehensive phenotypes (red).

patterns descriptive of the underlying disease progression from the labeled time-series data. First, we formally define the notion of temporal phenotypes as predictive temporal patterns. Then, the association of individual temporal patterns with the target disease status is assessed by proposing a novel path-based similarity score. For effective evaluation of the path-based similarity, we introduce a representation learning approach based on the Laplace transform to convert variable-length, irregularly sampled time-series data into unified embeddings. Finally, based on the resulting path-based similarity graph, we formulate the task of temporal phenotyping as a temporal predictive clustering problem that can be efficiently solved by adopting the graph-constrained K-means clustering.

We validate our approach through experiments on synthetic and real-world time-series datasets. Our method discovers temporal phenotypes that provide superior prediction performance compared to state-of-the-art benchmarks, and we corroborate the interpretability of our discovered phenotypes with supporting medical and scientific literature.

2 TEMPORAL PHENOTYPING

Suppose disease progression manifests through a multivariate continuous-time trajectory $\boldsymbol{x}(t) \in \mathcal{X}$ defined on $t \in [0, 1]$, where $\mathcal{X}$ is the functional space of all possible patient trajectories.[1] Each trajectory consists of dim_x-dimensional time-varying covariates, i.e., $\boldsymbol{x}(t) = [x_1(t), \dots, x_{\text{dim}_x}(t)]^\top$, each of which can be described by a continuous-time function x_i in $L^2_{[0,1]}$ (i.e., L^2-space under the interval $[0, 1]$).[2] Thus, the considered trajectory space can be given as $\mathcal{X} = \bigotimes_{\text{dim}_x} L^2_{[0,1]}$. Each trajectory $\boldsymbol{x}$ is correlated with a target label vector $\boldsymbol{y} = [y_1, \dots, y_{\text{dim}_y}]^\top \in \mathcal{Y}$ that describes the clinical status of the underlying disease progression (e.g., clinical endpoints). Throughout the paper, we focus our description on the case where the outcome of interest $\boldsymbol{y}$ is categorical and represented by a one-hot vector, i.e., $\mathcal{Y} = \{0, 1\}^{\text{dim}_y}$.

Let $p(\boldsymbol{x}, \boldsymbol{y})$ be the joint distribution of the continuous-time trajectory and the label vector. To discover temporal patterns that are predictive of the clinical status of patients, we first define a vector-valued function $g(\boldsymbol{x}) = [p(y_1|\boldsymbol{x}), \dots, p(y_{\text{dim}_y}|\boldsymbol{x})]^\top$ which implies the categorical conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$. We assume the clinical status conditioned on a patient trajectory can be represented by one of the δ-*separable modes* in $g(\boldsymbol{x})$. These modes are δ-separable such that they can be separated based on a proper distance metric d_y with some threshold $\delta > 0$. Here, we choose the Jensen–Shannon (JS) divergence as our distance metric, i.e., $\text{d}_y(\boldsymbol{v}, \boldsymbol{u}) = \frac{1}{2}KL(g(\boldsymbol{v})||\boldsymbol{m}) + \frac{1}{2}KL(g(\boldsymbol{u})||\boldsymbol{m})$, where KL is the Kullback-Leibler divergence, $\boldsymbol{m} = \frac{g(\boldsymbol{v})+g(\boldsymbol{u})}{2}$.

[1] Trajectories defined within the interval $\mathbb{R}_+$ can be simply scaled to the unit interval $[0, 1]$.

[2] In many practical scenarios, the continuous-time functions for time-varying covariates are bounded and fall into the L^2-space which has a natural extension of Euclidean distance.

2.1 Phenotypes: Predictive Temporal Patterns

In this subsection, we introduce the formal definition of *phenotypes* as temporal patterns that are predictive of disease progression. To this goal, we start by describing how the temporal patterns in continuous-time trajectories can be discovered and how the specific disease progression can be associated with each individual pattern.

Temporal Patterns. A temporal pattern characterizes some temporal dynamics that are shared by a subset of trajectories in $\mathcal{X}$. Here, we introduce a novel definition to describe temporal patterns in the general form based on connectivity in trajectory space $\mathcal{X}$. Given two trajectories $\boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathcal{X}$, we define a translation from $\boldsymbol{x}^1$ to $\boldsymbol{x}^2$, denoted as $\Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2)$, as a continuous path Γ connecting the two trajectories in space $\mathcal{X}$. Typically, $\Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2)$ can continuously morph the shape of $\boldsymbol{x}^1$ into that of $\boldsymbol{x}^2$. Then, we formally define a *temporal pattern* as a connected set $\Phi \subset \mathcal{X}$ such that all the trajectories in Φ can be inter-connected by translations within Φ. That is, there exists a series of translations from any trajectory to any other trajectory in Φ.

Phenotypes. Considering multivariate continuous-time trajectories, a variety of temporal patterns may exist in $\mathcal{X}$ while only a few of them are relevant to the target label. In the meantime, the clinical status marked by the same target label may manifest in patient trajectories through different temporal characteristics. For instance, in lung transplant referral of cystic fibrosis patients, (i) low lung function score, (ii) rapid declining lung function score, and (iii) multiple exacerbations requiring intravenous antibiotics are identified as distinct predictive temporal patterns (Ramos et al., 2019) among various temporal dynamics.

To provide insights on disease progression, desirable phenotypes shall be defined based on distinct predictive temporal patterns. In line with such notion of phenotypes, we propose a new path-based similarity score that measures the variation of conditional label distribution (described by function $g(\boldsymbol{x})$) along a translation between two trajectories. Specifically, consider two continuous-time trajectories $\boldsymbol{x}^1, \boldsymbol{x}^2$ and a translation $\Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2)$, the score function evaluates the similarity between $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ via their impact on label $\boldsymbol{y}$ through path Γ as follows:

$$\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2) = \max_{\substack{\boldsymbol{x} \in \Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2) \\ i \in \{1,2\}}} \mathrm{d}_y(g(\boldsymbol{x}), g(\boldsymbol{x}^i)). \tag{1}$$

Small value of $\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2)$ indicates that trajectories $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ share similar clinical status $\boldsymbol{y}$ and contain similar temporal patterns that are predictive of their associated label.

Finally, we provide a formal definition of *phenotype* as a predictive temporal pattern associated with a distinct clinical status as follows:

Definition 1. (Phenotype) *Let $\boldsymbol{v}$ be the centroid of a δ-separable mode in $g(\boldsymbol{x})$. Then, there exists a unique phenotype, denoted as a tuple $(\boldsymbol{v}, \Phi)$ with Φ as a set of trajectories, that satisfies the following two properties:*

$$\textit{(Similar clinical status)} \qquad \max_{\boldsymbol{x} \in \Phi} \mathrm{d}_y(g(\boldsymbol{x}), \boldsymbol{v}) \leq \frac{\delta}{2},$$

$$\textit{(Similar predictive pattern)} \qquad \max_{\substack{\boldsymbol{x}^1, \boldsymbol{x}^2 \in \Phi \\ \Gamma \subseteq \Phi}} \mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2) \leq \delta,$$

and any trajectory $\boldsymbol{x} \in \mathcal{X} \setminus \Phi$ is either not connected to Φ or has a different mode.

Intuitively, the homogeneity of each phenotype $(\boldsymbol{v}, \Phi)$ guarantees that the continuous-time trajectories exhibiting a similar temporal pattern will lead to a similar clinical status, which in turn provides a prognostic value on the underlying disease progression.

2.2 Predictive Temporal Clustering

In practice, the continuous-time trajectories of a patient are systematically collected in EHRs as discrete observations with irregular intervals during his/her regular follow-ups or stay at hospital. Hence, we focus this subsection on formulating the task of discovering phenotypes given discrete observations of trajectories as a novel clustering problem.

Suppose we have a dataset $\mathcal{D} = \{(\boldsymbol{t}^i, \boldsymbol{X}^i, \boldsymbol{y}^i)\}_{i=1}^N$ comprising discrete observations on the underlying continuous-time trajectories and target labels. Here, we denote discrete observations as time-series $\boldsymbol{X} = [\boldsymbol{x}(t_1), \boldsymbol{x}(t_2), \ldots, \boldsymbol{x}(t_T)]$ which contains sequential observations of a trajectory $\boldsymbol{x}$ at observation time stamps $\boldsymbol{t} = [t_1, t_2, \ldots, t_T]^\top$ with $0 \leq t_1 \leq \ldots \leq t_T \leq 1$. The label vector $\boldsymbol{y} \in \mathcal{Y}$ describes the clinical status sampled from the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$. From this point forward, we will slightly abuse the notation and interchangeably write $\boldsymbol{X}$ to denote the discrete time-series and the associated time stamps.

Path-Based Connectivity. Note that the property of a phenotype in Definition 1 requires all trajectories in that phenotype share a similar predictive pattern. Consider two time-series $\boldsymbol{X}^1, \boldsymbol{X}^2$ with underlying continuous-time trajectories $\boldsymbol{x}^1, \boldsymbol{x}^2$ from the same phenotype $(\boldsymbol{v}, \Phi)$. There must exist a translation Γ from trajectory $\boldsymbol{x}^1$ to $\boldsymbol{x}^2$ such that the condition in $\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2) \leq \delta$ holds. Violating such a condition implies a significant difference between the two trajectories suggesting they are from different phenotypes. Therefore, we utilize the *path-based connectivity test*, i.e., $\exists\, \Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2),\ \mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2) \leq \delta$, to assesses the phenotype similarity between two given trajectories $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$. This enables discovery of predictive temporal patterns without access to the ground-truth phenotypes. Evaluation of the path-based connectivity on all possible pairs of time-series in dataset $\mathcal{D}$ generates a distance matrix $\boldsymbol{S}$. Element-wise comparison of $\boldsymbol{S}$ and threshold δ yields a similarity graph $\mathcal{G}_\delta$ with edges between similar samples. We will discuss how we can approximately achieve the path-based connectivity test based on the discrete observations in the next section.

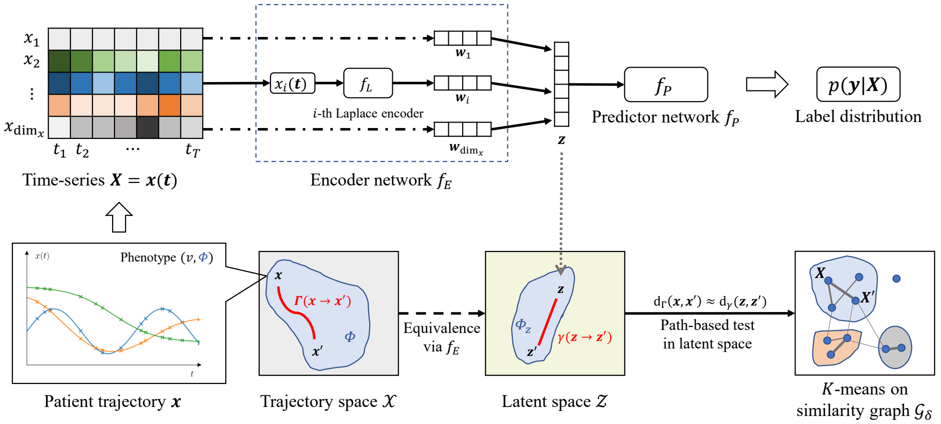

x_1
x_2
$\vdots$
$x_{\dim_x}$
t_1 t_2 $\cdots$ t_T
Time-series $\boldsymbol{X} = \boldsymbol{x}(\boldsymbol{t})$
$x_i(\boldsymbol{t})$
f_L
i-th Laplace encoder
$\boldsymbol{w}_1$
$\boldsymbol{w}_i$
$\boldsymbol{w}_{\dim_x}$
Encoder network f_E
$\boldsymbol{z}$
f_P
Predictor network f_P
$p(\boldsymbol{y}|\boldsymbol{X})$
Label distribution
$x(t)$
Phenotype (v, Φ)
t
Patient trajectory $\boldsymbol{x}$
$\boldsymbol{x}$
$\Gamma(\boldsymbol{x} \to \boldsymbol{x}')$
$\boldsymbol{x}'$
Φ
Trajectory space $\mathcal{X}$
Equivalence via f_E
$\boldsymbol{z}$
Φ_z
$\gamma(\boldsymbol{z} \to \boldsymbol{z}')$
$\boldsymbol{z}'$
Latent space $\mathcal{Z}$
$\mathrm{d}_\Gamma(\boldsymbol{x}, \boldsymbol{x}') \approx \mathrm{d}_\gamma(\boldsymbol{z}, \boldsymbol{z}')$
Path-based test in latent space
$\boldsymbol{X}$
$\boldsymbol{X}'$
K-means on similarity graph $\mathcal{G}_\delta$

Figure 2: Overview of T-Phenotype.

Temporal Phenotyping. To discover phenotypes from dataset $\mathcal{D}$, we assume that we have a proper approximator $f(\boldsymbol{X})$ of the conditional label distribution $g(\boldsymbol{x})$ from discrete observations in $\boldsymbol{X}$. Thus, similarity graph $\mathcal{G}_\delta$ can be constructed based on the path-based connectivity test with approximator $f(\boldsymbol{X})$. Now, we formulate the task of temporal phenotyping as a predictive clustering problem (Lee and van der Schaar, 2020) to group time-series into different clusters on top of $\mathcal{G}_\delta$. More specifically, the clusters (with distinct phenotypes) are discovered by solving the following constrained optimization problem:

$$\begin{aligned} \min_{\mathcal{C}} \quad & \sum_{C_k \in \mathcal{C}} \sum_{\boldsymbol{X} \in C_k} \mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}_k), \\ s.t. \quad & \forall \boldsymbol{X}^1, \boldsymbol{X}^2 \in C_k, \quad \boldsymbol{X}^1 \overset{\mathcal{G}_\delta}{\longleftrightarrow} \boldsymbol{X}^2, \end{aligned} \tag{2}$$

where $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ is a feasible set of $K \in \mathbb{N}$ clusters each of which has a centroid $\boldsymbol{v}_k$ as the average density $f(\boldsymbol{X})$, Since threshold δ is usually unknown in advance, we set its value according to $\delta = 2\max_{C_k \in \mathcal{C}, \boldsymbol{X} \in C_k} \mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}_k)$ for consistency with Definition 1. Here, $\boldsymbol{X}^1 \overset{\mathcal{G}_\delta}{\longleftrightarrow} \boldsymbol{X}^2$ implies that there exists a path over graph $\mathcal{G}_\delta$ such that $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ are interconnected. In (2), the objective function encourages the cluster centroids to be clearly distinguished in approximated label distribution $f(\boldsymbol{X})$ while the constraint on similarity graph $\mathcal{G}_\delta$ ensures that samples in the same cluster are of similar phenotypes. Each discovered cluster C_k represents a unique phenotype with centroid $\boldsymbol{v}_k$ describing the associated clinical status and allows us to explain the predictive temporal pattern in terms of the collection of time-series in C_k.

Unfortunately, the optimization problem in (2) is highly non-trivial due to the following two challenges: First, it requires to learn a proper approximation of the conditional label distribution from irregularly-sampled discrete time-series. Second, an efficient evaluation of the path-based connectivity test is required to construct similarity graph $\mathcal{G}_\delta$ given discrete time-series in $\mathcal{D}$.

3 METHOD: T-PHENOTYPE

In this section, we propose a novel temporal clustering framework, *T-Phenotype*, that effectively discovers phenotypes from discrete time-series data. To estimate the conditional label distribution from discrete time-series, we introduce two networks, an encoder and a predictor. The encoder, f_E, comprises $\dim_x$ feature-wise Laplace encoders, each of which transforms a single feature dimension of discrete time-series $\boldsymbol{X}$ into a fixed-length latent embedding. The predictor, f_P, takes embeddings from $\dim_x$ Laplace encoders as the input $\boldsymbol{z}$ in the latent space and estimates the conditional label distribution. The proposed Laplace encoders, f_L, allow us to establish (approximately) equivalence translation in the latent space and thereby to efficiently evaluate the path-based connectivity test between discrete time-series in dataset $\mathcal{D}$. Then, given an approximate similarity graph $\mathcal{G}_\delta$ constructed from the result of pair-wise connectivity test, we propose a graph-constrained K-means algorithm to discover distinct phenotypes. The overview of steps involved in T-Phenotype is illustrated in Figure 2.

3.1 Time-Series Embedding via Laplace Encoder

Now, we introduce a novel time-series encoder which encodes each dimension of a given discrete time-series into a unified parametric function in the frequency domain as an approximation of the Laplace transform.

Laplace Encoder. Let $x(\boldsymbol{t}) = [x(t_1), \ldots, x(t_T)]^\top \in \mathbb{R}^T$ be a time-series of discrete observations on a univariate trajectory $x(t)$ at time stamps $\boldsymbol{t} = [t_1, \ldots, t_T]^\top$ in the unit interval. The Laplace encoder (parameterized by θ_L), $f_L : \mathbb{R}^T \to \mathbb{C}^{n(d+1)}$, encodes discrete time-series $x(\boldsymbol{t})$ into a rational function on the complex plane with $n \in \mathbb{N}$ poles of maximum degree of $d \in \mathbb{N}$ as follows:

$$F_w(s) = \sum_{m=1}^{n} \sum_{l=1}^{d} \frac{c_{m,l}}{(s - p_m)^l}, \quad c_{m,l}, p_m \in \mathbb{C}. \tag{3}$$

Here, $\boldsymbol{w} \triangleq f_L(x(\boldsymbol{t})) = [p_1, \ldots, p_n, c_{1,1}, \ldots, c_{n,d}]^\top$ is the Laplace embedding comprising the poles and the corresponding coefficients. Note that the poles in (3) are distinct and are in a lexical order, i.e., $p_m \leq p_{m+1}$ for $m = 1, \ldots, n-1$ where $p_m \leq p_n$ if and only if $\mathrm{Re}(p_m) < \mathrm{Re}(p_n)$ or $\mathrm{Re}(p_m) = \mathrm{Re}(p_n) \wedge \mathrm{Im}(p_m) \leq \mathrm{Im}(p_n)$ holds. Then, the time-domain function can be efficiently reconstructed via the inverse Laplace transform:

$$\hat{x}(t) = \frac{1}{2\pi j} \lim_{T\to\infty} \int_{\sigma - jT}^{\sigma + jT} e^{st} F_w(s) \mathrm{d}s, \tag{4}$$

where $j^2 = -1$ and σ is some suitable complex number such that $\mathrm{Re}(\sigma) > \max_{p_m \in \boldsymbol{w}} \mathrm{Re}(p_m)$. With a sufficient number of poles, the Laplace embedding $\boldsymbol{w}$ becomes an equivalent description of the underlying trajectory $x(t)$. That is, the orthonormal basis $\{e^{2\pi jmt}, m \in \mathbb{Z}\}$ of $L^2_{[0,1]}$ is covered by the reconstruction $\hat{x}(t)$ when $n \to \infty$.

Given a dataset of N discrete univariate time-series, i.e., $\{x^i(\boldsymbol{t})\}_{i=1}^N$, we train the Laplace encoder utilizing the following loss function that consists of the time-series reconstruction error and the regularization term specifically designed to encourage unique Laplace embeddings:

$$\mathcal{L}_{\text{laplace}}(\theta_L) = \mathcal{L}_{\text{mse}}(\theta_L) + \alpha \mathcal{L}_{\text{unique}}(\theta_L) \tag{5}$$

where α is a balancing coefficient. The former term, i.e., $\mathcal{L}_{\text{mse}}(\theta_L) = \frac{1}{N}\sum_{i=1}^N \|x^i(\boldsymbol{t}) - \hat{x}^i(\boldsymbol{t})\|_2^2$, is the reconstruction error from our Laplace embeddings, and the latter term, i.e., $\mathcal{L}_{\text{unique}}(\theta_L) = \frac{1}{N(N-1)}\sum_{i\neq j} \ell_{\text{unique}}(\hat{x}^i(\boldsymbol{t}), \hat{x}^j(\boldsymbol{t}))$, encourages the uniqueness of the Laplace embedding. More specifically, ℓ_{unique} focuses on three aspects – (i) the obtained poles are distinct, (ii) the reconstructed trajectories are real-valued, and (iii) no two distinct Laplace embeddings generate the same trajectory. We further elaborate the uniqueness regularization in the Appendix.

From Trajectory Space to Latent Space. Utilizing $\dim_x$ feature-wise Laplace encoders as our encoder, f_E, any discrete observations of a continuous-time trajectory $\boldsymbol{x} \in \mathcal{X}$ can be transformed into a fixed-length embedding $\boldsymbol{z} \in \mathcal{Z}$ in the latent space as a composition of $\dim_x$ Laplace embeddings, i.e., $\boldsymbol{z} \triangleq [f_L(x_1(\boldsymbol{t})), \ldots, f_L(x_{\dim_x}(\boldsymbol{t}))]^\top$. The following proposition builds a strong connection between the trajectory space $\mathcal{X}$ and the latent space $\mathcal{Z}$:

Proposition 1. *Without loss of generality, consider univariate continuous-time trajectories $\hat{x}^1, \hat{x}^2 \in \mathcal{X}$ and their corresponding latent embeddings $\boldsymbol{z}^1, \boldsymbol{z}^2 \in \mathcal{Z}$, respectively. Then, the distance between two trajectories can be bounded by $\|\hat{x}^1 - \hat{x}^2\|^2_{L^2_{[0,1]}} \leq \psi \|\boldsymbol{z}^1 - \boldsymbol{z}^2\|_2^2$, where $\psi > 0$ is a constant and $\|x(t)\|^2_{L^2_{[0,1]}} = \int_0^1 x(t)\overline{x(t)}\mathrm{d}t$.*

The detailed proof can be found in the Appendix. Consider a subset of latent variables Φ_z and the corresponding trajectory set Φ of their time-domain representations. The upper bound in Proposition 1 implies that continuity of Φ_z in the latent space leads to the continuity of Φ in the trajectory space. This property allows efficient evaluation of the path-based connectivity test in the latent space as illustrated in the following subsection.

3.2 Efficient Evaluation of Path-based Similarity

Construction of similarity graph $\mathcal{G}_\delta$ involves iterative evaluation of the path-based similarity score d_Γ in (1) for all possible pairs of time-series samples in $\mathcal{D}$. This requires a substantial number of computations in both constructing translation Γ and calculating conditional $g(\boldsymbol{x})$ on all available continuous-time trajectories $\boldsymbol{x} \in \Gamma$. Instead, we efficiently approximate the similarity graph $\mathcal{G}_\delta$ via path-based connectivity test in the latent space and estimate the conditional $g(\boldsymbol{x})$ via neural networks.

Translation in Latent Space. Consider two trajectories $\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{x}}^2 \in \mathcal{X}$ with the corresponding latent embedding $\boldsymbol{z}^1, \boldsymbol{z}^2 \in \mathcal{Z}$. For any translation $\Gamma(\hat{\boldsymbol{x}}^1 \to \hat{\boldsymbol{x}}^2) \subseteq \mathcal{X}$ in trajectory space, we can always find a continuous path in the latent space, i.e., $\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2) \subseteq \mathcal{Z}$, such that the distance between its time-domain reconstruction and Γ is minimized. We consider γ to be an (approximately) equivalent translation of Γ.[3] This enables us to capitalize on the translation in the latent space without constructing intermediate trajectories along path Γ, which significantly reduces computations in obtaining the path-based similarity in (1).

Predictor. To estimate the function $g(\boldsymbol{x})$, we utilize the time-series encoder f_E, which consists of $\dim_x$ Laplace encoders, and a predictor f_P (an MLP parameterized by θ_P) to construct the approximator as $f(\boldsymbol{X}) \triangleq f_P \circ f_E(\boldsymbol{X}) \approx g(\boldsymbol{x})$ where $\boldsymbol{X}$ is the discrete observation of trajectory $\boldsymbol{x}$. The predictor f_P is trained based on the cross-entropy loss:

$$\mathcal{L}_{\text{predictor}}(\theta_P) = -\frac{1}{N}\sum_{i=1}^N \sum_{c=1}^{\dim_y} \boldsymbol{y}_c^i \log f_P(\boldsymbol{z}^i)_c, \tag{6}$$

where $\boldsymbol{z} = f_E(\boldsymbol{X})$ and subscript c indicates the c-th element in the output space. To maintain the property of the Laplace encoders, we only update the predictor via the signal from the label during training.

Consider a trajectory translation $\Gamma(\hat{\boldsymbol{x}}^1 \to \hat{\boldsymbol{x}}^2)$ and its equivalent translation $\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2)$ in latent space, the path-based similarity can be approximately calculated as

$$\mathrm{d}_\Gamma(\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{x}}^2) \approx \mathrm{d}_\gamma(\boldsymbol{z}^1, \boldsymbol{z}^2) = \max_{\boldsymbol{z}\in\gamma, i=1,2} \mathrm{d}_y(f_P(\boldsymbol{z}), f_P(\boldsymbol{z}^i)). \tag{7}$$

Hence, given two discrete time-series $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$, the path-based connectivity test can be efficiently performed along translation γ in the latent space without assessing the corresponding translation in the trajectory space $\mathcal{X}$.

[3] The equivalence is strict when all trajectories along translation Γ have rational Laplace transform as described in (3).

Approximate Similarity Graph. Consider a phenotype $(\boldsymbol{v}, \Phi)$ where centroid $\boldsymbol{v}$ represents a specific clinical status and Φ is the associated predictive temporal pattern. The encoder f_E is trained to map time-series $\boldsymbol{X}$ sampled from trajectories in Φ into a connected area Φ_z in latent space $\mathcal{Z}$ via Laplace encoders. Given time-series $\boldsymbol{X}$ that is observed from trajectory $\boldsymbol{x} \in \Phi$, Definition 1 implies that we have $\mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}) \leq \frac{\delta}{2}$ where $f(\boldsymbol{X}) = f_P(\boldsymbol{z})$ and $\boldsymbol{z} = f_E(\boldsymbol{X}) \in \Phi_z$. Hence, for two embeddings $\boldsymbol{z}^1, \boldsymbol{z}^2 \in \Phi_z$, there always exist a translation $\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2) \subseteq \Phi_z$ such that $\mathrm{d}_\gamma(\boldsymbol{z}^1, \boldsymbol{z}^2) \leq \delta$ due to the connectivity of Φ_z in the latent space. If two latent embeddings $\boldsymbol{z}^1, \boldsymbol{z}^2$ are located in the same convex subset of Φ_z, linear path $\bar{\gamma}(\boldsymbol{z}^1 \to \boldsymbol{z}^2) = \{\boldsymbol{z} | (1-a)\boldsymbol{z}^1 + a\boldsymbol{z}^2, a \in [0,1]\}$ suffices the connectivity test. When $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ are in different convex subsets, the connectivity of Φ_z guarantees that there exists a series of intermediate points $\boldsymbol{z}^{m_1}, \boldsymbol{z}^{m_2}, \ldots, \boldsymbol{z}^{m_l}$ such that composite path $\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2) = \bar{\gamma}(\boldsymbol{z}^1 \to \boldsymbol{z}^{m_1}) \cup \ldots \cup \bar{\gamma}(\boldsymbol{z}^{m_l} \to \boldsymbol{z}^2)$ is inside Φ_z and can be used for connectivity test. Therefore, in this work, we simplify the path-based connectivity test to the linear paths between latent variables as the similarity between two time-series can be inferred based on these linear paths. Overall, given two time-series $\boldsymbol{X}^i$ and $\boldsymbol{X}^j$, we calculate the approximate distance $\mathrm{d}_{\bar{\gamma}}(f_E(\boldsymbol{X}^i), f_E(\boldsymbol{X}^j))$ via discrete points along path $\bar{\gamma}$, which is stored in element S_{ij} of *path-based distance matrix* $\boldsymbol{S}$. The approximate similarity graph $\mathcal{G}_\delta$ is then constructed with edges between samples $\boldsymbol{X}^i$ and $\boldsymbol{X}^j$ if and only if $S_{i,j} \leq \delta$.

3.3 Predictive Clustering on Similarity Graph

Unfortunately, solving the clustering objective in (2) is a NP-hard combinatorial problem. Thus, we introduce a greedy approach to discover the temporal clusters from the path-based distance matrix $\boldsymbol{S}$ defined in the previous subsection.

The objective function in (2) has the following upper bound:

$$\begin{aligned} J &\triangleq \sum_{C_k \in \mathcal{C}} \sum_{\boldsymbol{X} \in C_k} \mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}_k), \\ &\leq \sum_{C_k \in \mathcal{C}} \frac{1}{|C_k|} \sum_{\boldsymbol{X}^i, \boldsymbol{X}^j \in C_k} \mathrm{d}_y(f(\boldsymbol{X}^i), f(\boldsymbol{X}^j)), \\ &\leq \sum_{C_k \in \mathcal{C}} \sum_{\boldsymbol{X}^i, \boldsymbol{X}^j \in C_k} \mathrm{d}_{\bar{\gamma}}(\boldsymbol{z}^i, \boldsymbol{z}^j), \\ &= \sum_{C_k \in \mathcal{C}} \sum_{\boldsymbol{X}^i, \boldsymbol{X}^j \in C_k} S_{ij} \triangleq \bar{J}(\boldsymbol{S}), \end{aligned} \tag{8}$$

where $\boldsymbol{z}^i = f_E(\boldsymbol{X}^i)$, latent translation $\bar{\gamma}$ is a linear path connecting two embeddings $\boldsymbol{z}^i$ and $\boldsymbol{z}^j$. The first inequality comes from the convexity of the JS divergence, and the second inequality establishes from equation (7) and the fact that $|C_k| \geq 1$. Local minimum of the upper bound $\bar{J}(\boldsymbol{S})$ can be achieved via a greedy K-partitioning algorithm based on pair-wise sample distances in matrix $\boldsymbol{S}$.

Utilizing the approximate solution in (8) as warm-start, we propose a graph-constrained K-means clustering approach to solve problem (2) via a greedy breadth-first search algorithm GK-means (details in Appendix). The overview of our predictive clustering method, T-Phenotype, is given in Algorithm 1. More details about the algorithm are provided in the Appendix.

Algorithm 1 T-Phenotype

```
Input: dataset $\mathcal{D}$, number of clusters $K$
Output: $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$
  calculate distance matrix $\boldsymbol{S}$ based on (7)
  $\mathcal{C} \leftarrow \arg\min_{\mathcal{C}} \bar{J}(\boldsymbol{S})$                    ▷ warm-start
  $\delta \leftarrow \log(2)$                    ▷ upper bound of $\mathrm{d}_{\mathrm{JS}}$
  while not converged do
      for $k = 1, 2, \ldots, K$ do
          update cluster seed $e_k$ via (9)
      end for
      $\delta' \leftarrow 2\max_{C_k \in \mathcal{C}, \boldsymbol{X} \in C_k} \mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}_k)$
      $\delta \leftarrow \min(\delta, \delta')$                    ▷ upper bound $J \leq N\delta$
      create similarity graph $\mathcal{G}_\delta$ from $S_{i,j} \leq \delta$
      $\mathcal{C} \leftarrow$ GK-means$(J | e_1, e_2, \ldots, e_K, \mathcal{G}_\delta)$
  end while
```

The cluster seeds in Algorithm 1 are used to perform greedy cluster expansion over similarity graph $\mathcal{G}_\delta$. For the k-th cluster, the cluster seed $e_k = (\boldsymbol{v}_k, \boldsymbol{X}^{(k)})$ can be given as

$$\boldsymbol{v}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{X} \in C_k} f(\boldsymbol{X}), \quad \boldsymbol{X}^{(k)} = \underset{\boldsymbol{X} \in C_k}{\arg\min}\, \mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}_k), \tag{9}$$

where $\boldsymbol{v}_k$ is the cluster centroid and $\boldsymbol{X}^{(k)}$ is the representative time-series in cluster C_k with closest conditional to that of the centroid.

4 RELATED WORK

Different strands of clustering methods have been increasingly investigated for knowledge discovery from time-series data with various similarity notions accustomed to specific application scenarios. One strand is unsupervised clustering methods that adopt the traditional notion of similarity into the time-series setting. To flexibly incorporate with variable-length irregularly-sampled time-series observations, the traditional methods applied K-means clustering by either finding fixed-length and low-dimensional representations using deep learning-based sequence-to-sequence model (Ma et al., 2019; Zhang et al., 2019) or on modifying the similarity measure such as dynamic time warping (DTW) (Giannoula et al., 2018) and the associated graph Laplacian (Lei et al., 2019; Hayashi et al., 2005). Alternatively, Bahadori et al. (2015) focused on sample affinities to conduct spectral clustering, and Chen et al. (2022) proposed a deep generative model whose parametric space is then used for clustering. Further, advanced hidden Markov models (Ceritli et al., 2022) and Gaussian processes (Schulam et al., 2015) have also been utilized together with hierarchical graph models in disease subtype discovery. In general, these methods are

Table 1: Comparison of Temporal Clustering Methods. The difference in the notion of phenotypes and similarity measure are highlighted together with two desiderata: (i) clusters are outcomes associated; and (ii) with interpretable insights on cluster assignment.

METHOD	PHENOTYPE	SIMILARITY MEASURE	(I)	(II)
Deep temporal K-means	Distance-based	Euclidean distance	✓	✗
Bahadori et al. (2015)	Affinity-based	Self-expression	✗	✗
Chen et al. (2022)	Pattern-oriented	Latent distance	✗	✓
Aguiar et al. (2022)	Attention&outcome-oriented	KL-divergence	✓	✓
Lee and van der Schaar (2020)	Outcome-oriented	KL-divergence	✓	✗
T-Phenotype (Ours)	Predictive pattern-oriented	Path-based connectivity	✓	✓

limited by some model specifications such as the linear subspace assumptions and graphical models for the underlying data generation process.

Clusters identified through these methods are purely unsupervised – they do not account for patients' clinical outcomes that are often available in EHRs – which may lead to heterogeneous outcomes even for patients in the same cluster. To overcome this issue, another strand of clustering methods combine predictions on the future outcomes with clustering. Lee and van der Schaar (2020) proposed an actor-critic approach to divide time-series of patient trajectories into subgroups based on their associated clinical status. The discovered patient subgroups allow clinicians to investigate the temporal patterns related to the transition of disease stages. Aguiar et al. (2022) extended it to capture phenotype-related feature contributions by employing an attention mechanism. Given predicted clusters, visualizing the associated attention map provides additional interpretability about the underlying disease progression.

Unfortunately, actionable information that can be inferred from the aforementioned temporal predictive clusters is still limited. These methods primarily focus on finding the discrete representations that can best describe the outcome labels rather without properly associating with temporal patterns that can be found among time-series samples. In this paper, we propose a novel temporal clustering method to correctly uncover predictive temporal patterns descriptive of the underlying disease progression from the labeled time-series data. Therefore, our method not only can provide clusters that have a prognostic value but also can offer interpretable information about the disease progression patterns.

5 EXPERIMENTS

In this section, we evaluate the clustering performance and the prognostic value of T-Phenotype with one synthetic dataset and two real-world datasets (detailed statistics are provided in the Appendix).

Synthetic Dataset. We construct a synthetic dataset of $N = 1200$ samples with ground truth cluster labels. Each sample comprises discrete observations of a 2-dimensional trajectory $\boldsymbol{x}(t)$ and the target binary outcome. We design the two elements $x_1(t)$ and $x_2(t)$ to model trend and periodicity of a trajectory, respectively: we set $x_1(t) = \iota \cdot \mathrm{sigmoid}(a \cdot (t - b - \varphi))$ with sign $\iota \in \{-1, 1\}$, $a = 10$, $b = 0.5$, and $\varphi \sim \exp(\frac{3}{10})$ and set $x_2(t) = \sin(c \cdot (t - \varphi))$ with $c \in \{4, 6, 8\}$ and φ identical to that of x_1. The trajectory $\boldsymbol{x} = [x_1, x_2]^\top$ is irregularly observed over 20 time stamps in $t \in [0, 2]$ with a white noise $\mathcal{N}(0, 0.1^2)$ for each variable. We set c as the ground truth phenotype label representing different periodicity and set the target outcome label y as $y = 0$ when $c = 6$ and $y = 1$ otherwise.

ADNI Dataset. The Alzheimer's Disease Neuroimaging Initiative[4] (ADNI) dataset includes records on the progression of Alzheimer's disease (AD) of $N = 1346$ patients with regular follow-ups every six months. Each patient is associated with various biomarkers, evaluation of MRI and PET images, and cognitive tests results. We set the target outcome at each time stamp as the three diagnostic groups – i.e., normal brain functioning (NL), mild cognitive impairment (MCI), and AD – which is used to indicate different stages of AD progression. We focus on three important temporal variables – i.e., the genetic biomarker of apolipoprotein (APOE) $\varepsilon 4$ gene, the hippocampus evaluation from MRI, and the cognitive test result of CDRSB – to predict the AD progression.

ICU Dataset. The PhysioNet ICU[5] (Goldberger et al., 2000) dataset contains temporal observations on 42 covariates of adult patients over the first 48 hours of ICU stay. We extract $N = 1554$ records of adult patients admitted to the medical or surgical ICU. Temporal covariates used in the experiments are age, gender, Glasgow Coma Scale (GCS), and partial pressure of arterial CO2 (PaCO2) with a time resolution of 1 hour, and we set patient mortality as the target binary outcome of interest.

Baselines. We compare the performance of T-Phenotype with the following benchmarks ranging from traditional method to recently developed deep learning-based methods, where each clustering method reflects a different notion of temporal phenotypes: 1) K-means with warping-based distance (KM-DTW); 2) deep temporal K-means with the encoder-predictor (E2P) structure introduced in (Lee and

[4] https://adni.loni.usc.edu

[5] https://physionet.org/content/challenge-2012/

van der Schaar, 2020), i.e., KM-E2P(z) and KM-E2P(y); 3) K-means on top of our proposed Laplace encoder (KM-$\mathcal{L}$); 4) sequence-to-sequence with K-means friendly representation space (SEQ2SEQ); and 5) the state-of-the-art temporal clustering approach AC-TPC (Lee and van der Schaar, 2020). Detailed description can be found in Appendix. In addition, we consider the ablation study of T-Phenotype with joint optimization for the Laplace encoders and predictor f_P and denote such model with T-Phenotype (J).

Throughout the experiments, time stamps of discrete time-series are scaled into $t \in [0, 1]$. For the synthetic and ADNI datasets, we use 64/16/20 train/validation/test splits in experiments. To get reliable clustering performance measurement on the ICU dataset, we use 48/12/40 train/validation/test splits for experiments. Hyperparameters of T-Phenotype and baselines are optimized through 3-fold cross-validation. For comparison of clustering performance, the number of clusters K for each dataset is shared by all methods. We select K as a hyperparameter of T-Phenotype, and the optimal cluster numbers are determined to be $K = 3$ (ground truth), $K = 4$ and $K = 3$ for the synthetic, ADNI and ICU dataset, respectively. Details can be found in the Appendix.

Table 2: Clustering Performance on the Synthetic Dataset.

METHOD	PURITY	RAND	NMI
KM-E2P(y)	0.663±0.019	0.477±0.033	0.569±0.045
KM-E2P(z)	0.677±0.029	0.418±0.024	0.485±0.047
KM-DTW	0.469±0.017	0.068±0.021	0.077±0.022
KM-$\mathcal{L}$	0.687±0.033	0.395±0.058	0.447±0.059
SEQ2SEQ	0.378±0.008	-0.003±0.003	0.005±0.003
AC-TPC	0.659±0.020	0.487±0.035	0.596±0.043
T-Phenotype (J)	0.655±0.021	0.440±0.051	0.543±0.064
T-Phenotype	**0.965±0.018**‡	**0.902±0.048**‡	**0.875±0.050**‡

Purity score, Rand index and normalized mutual information (NMI) are used to evaluate the clustering performance with ground truth phenotype labels. Best performance is highlighted in **bold**, and ‡ indicates p-value < 0.01.

Table 3: Benchmark Result on Two Real-world Datasets.

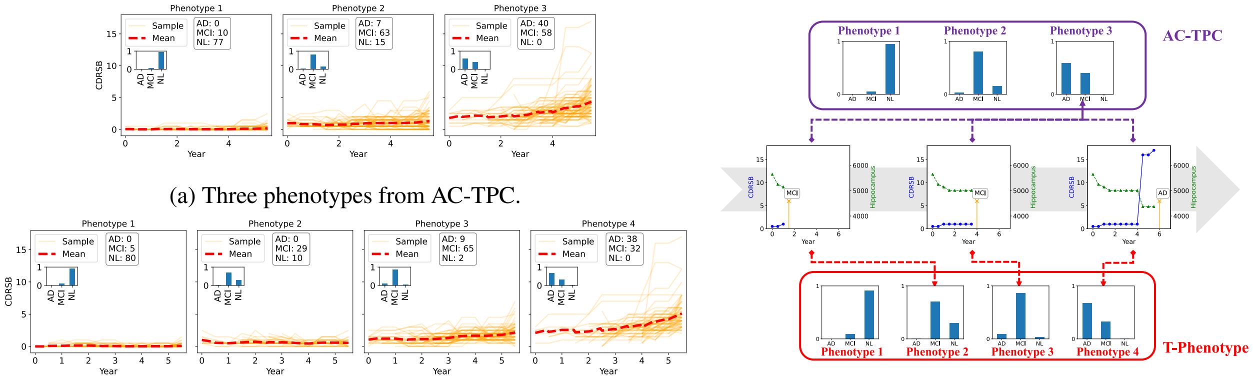
	METHOD	AUROC	AUPRC	H_{ROC}	H_{PRC}
ADNI	KM-E2P(y)	**0.893±0.005**	**0.728±0.017**	0.770±0.013	0.701±0.012
	KM-E2P(z)	0.884±0.012	0.711±0.020	0.763±0.018	0.690±0.013
	KM-DTW	0.743±0.013	0.522±0.020	0.752±0.027	0.618±0.021
	KM-$\mathcal{L}$	0.697±0.029	0.465±0.019	0.753±0.019	0.593±0.018
	SEQ2SEQ	0.775±0.023	0.550±0.030	0.773±0.012	0.642±0.022
	AC-TPC	0.861±0.012	0.665±0.020	0.788±0.014	0.694±0.013
	T-Phenotype (J)	0.867±0.020	0.679±0.040	0.768±0.011	0.684±0.021
	T-Phenotype	0.891±0.005	0.716±0.015	**0.791±0.013**	**0.713±0.009**‡
ICU	KM-E2P(y)	**0.697±0.014**	0.593±0.012	0.682±0.029	0.628±0.025
	KM-E2P(z)	0.677±0.030	0.579±0.018	0.686±0.031	0.633±0.024
	KM-DTW	0.539±0.030	0.515±0.011	0.636±0.023	0.621±0.021
	KM-$\mathcal{L}$	0.577±0.019	0.532±0.009	0.682±0.009	0.649±0.004
	SEQ2SEQ	0.592±0.024	0.539±0.012	0.690±0.011	**0.653±0.004**
	AC-TPC	0.660±0.008	0.573±0.003	0.695±0.014	0.644±0.011
	T-Phenotype (J)	**0.697±0.025**	**0.595±0.017**	0.691±0.056	0.636±0.048
	T-Phenotype	0.681±0.017	0.585±0.015	**0.703±0.007**	0.648±0.008

The area under the curve of receiving-operator characteristic (AUROC) and area under the curve of precision-recall (AUPRC) are used to assess the prognostic value of the discovered clusters on predicting target outcomes. Two composite metrics H_{ROC} and H_{PRC}, calculated as harmonic means between predictive accuracy (AUROC or AUPRC) and a cluster consistency metric AUSIL, are used to measure the phenotype discovery performance. Please refer to the Appendix for details. Best performance is highlighted in **bold**, and ‡ indicates p-value < 0.01.

Benchmark. The clustering performance of T-Phenotype is compared with six baselines, with all results reported using 5 random train/validation/test splits of the corresponding dataset. Benchmark results on synthetic dataset and two real-world datasets are provided in Table 2 and Table 3, respectively. Complete benchmark tables are available in the Appendix. On the synthetic dataset, T-Phenotype outperforms all baselines with significant gaps in considered clustering accuracy metrics. Similarly, T-Phenotype has the best (or very close to best) outcome prediction performance on both ADNI and ICU datasets and outperforms AC-TPC and most other baselines in phenotype discovery on the two datasets. The baseline of KM-E2P(y) directly discovers clusters over predicted outcome distributions and achieves the best prediction performance on the ADNI dataset, which is within expectation. However, its clustering performance, particularly H_{ROC}, is inferior to that of T-Phenotype due to the negligence of similarity in temporal patterns. On the ICU dataset, while T-Phenotype has close phenotype discovery performance H_{PRC} to baseline SEQ2SEQ, the clusters discovered by our method provide greater prognostic values as reflected in the outcome prediction accuracy.

Phenotypes of AD Progression. The CDRSB score measures the impairment on both cognitive abilities and brain function (Coley et al., 2011) and is widely used in AD progression assessment and staging (Kim et al., 2020; O'Bryant et al., 2008). The temporal patterns in CDRSB trajectory vary in different disease stages and show stable prognostic power on patient outcomes (Delor et al., 2013). On the ADNI dataset, four phenotypes are discovered by T-Phenotype. We examine these phenotypes by plotting the CDRSB scores of $N_{test} = 270$ test samples separately in corresponding clusters. As shown in Figure 3b, normal and high-risk patients with divergent cognitive test trajectories are correctly identified in phenotype 1 and 4 by T-Phenotype. In the meantime, for the predicted outcome of MCI, two subtypes of patients are clearly separated into two phenotypes (2 and 3) with different growth rates in CDRSB score. In comparison, AC-TPC fails to distinguish between these two subtypes as illustrated in Figure 3a, which impedes the prognostic value of clusters discovered by AC-TPC.

Prognostic Value of T-Phenotype. We further demonstrate the prognostic value of T-Phenotype with the temporal phenotyping results obtained on a typical patient from the ADNI dataset. The studied patient had a positive biomarker of APOE $\varepsilon 4$ gene which contributes to an increased risk of AD (Yamazaki et al., 2019). Consecutive observations of patient covariates at three time stamps are plotted in Figure 3c. Hippocampus volume (green triangle) and CDRSB score (blue dot) are displayed together with diagnosis obtained at the next follow-up (yellow bar). The temporal phenotype assignment via T-Phenotype is shown at the bottom. As a predictive factor of early-stage AD (Rao et al., 2022), fast decrease in hippocampus volume leads to the initial diag-



(a) Three phenotypes from AC-TPC.

(b) Four phenotypes from T-Phenotype.

(c) Prognostic values of T-Phenotype and AC-TPC.

Figure 3: Comparison of Phenotypes Discovered by T-Phenotype and AC-TPC on the ADNI Dataset.

nosis of phenotype 2 (MCI) in Figure 3b by T-Phenotype despite a low CDRSB score from cognitive test. Then, with a clear trend of increase appearing in CDRSB trajectory, the studied patient is classified into phenotypes ($2 \to 3 \to 4$) that reflect the growing risk in developing AD. In contrast, as shown on the top of Figure 3c, AC-TPC simply assigns the same phenotype to the patient throughout the considered time period and is unable to provide comparable insights on AD progression from the patient trajectory.

6 CONCLUSION

In this paper, we propose a novel phenotype discovery approach T-Phenotype to uncover predictive patterns from labeled time-series data. A representation learning method in frequency-domain is developed to efficiently embed the variable-length, irregularly sampled time-series into a unified latent space that provides insights on their temporal patterns. With our new notion of path-based phenotype similarity, a graph-constrained K-means approach is utilized to discover clusters representing distinct phenotypes. Throughout experiments on synthetic and real-world datasets, we show that T-Phenotype outperforms all baselines in phenotype discovery. The utility of T-Phenotype to discover clinically meaningful phenotypes is further demonstrated via comparison with the the state-of-the-art temporal phenotyping method AC-TPC on real-world healthcare datasets.

7 LIMITATIONS

Our proposed method, T-Phenotype, leverages Laplace encoders as a general approach to capture temporal patterns from time-series data as distinct Laplace embeddings. However, there may exist some complex temporal patterns, e.g., interactions between patient covariates at two specific time points, that cannot be encoded in this manner. To address this issue, additional representations (e.g., representation via attention mechanism) from the input time-series can be introduced to augment the Laplace embedding, which we leave as a future work. In the meantime, the phenotype discovery performance of T-Phenotype is highly dependent on the quality of predictor f_P. Unstable predictions from f_P will directly lead to inaccuracies in phenotype assignment. Thus, effective regularization of the predictor network would be another important future direction.

8 SOCIETAL IMPACT

Discovery of phenotypes from disease trajectories is a long pursuit in healthcare. In line with the target of precision medicine, the phenotype connects temporal patterns in patient trajectory and clinical outcomes is of great prognostic value since it allows clinicians to make more accurate diagnosis and issue the most appropriate treatment to their patients. By combining notions of similarity in both patient trajectories and clinical outcomes, our method, T-Phenotype, can effectively identify phenotypes of desired property. The discovered patient subgroups can be used to improve current clinical guidelines and help clinicians to better understand the disease progression of their patients. Nevertheless, the association between temporal patterns and clinical outcome in a phenotype cannot be interpreted as causal relationship without careful tests and examinations. Application of T-Phenotype without audits from human experts may lead to undesirable outcome of patients in certain edge cases.

Acknowledgements

Yuchao Qin was supported by the Cystic Fibrosis Trust. Changhee Lee was supported through the IITP grant funded by the Korea government (MSIT) (No. 2021-0-01341, AI Graduate School Program, CAU). We thank all reviewers for their time in helping us to evaluate our work. Their insightful comments are greatly appreciated.

References

H. Aguiar, M. Santos, P. Watkinson, and T. Zhu. Learning of cluster-based feature importance for electronic health record time-series. In *International Conference on Machine Learning*, pages 161–179. PMLR, 2022.

M. T. Bahadori, D. Kale, Y. Fan, and Y. Liu. Functional subspace clustering with application to time series. In *International Conference on Machine Learning*, pages 228–237. PMLR, 2015.

P. G. Bastos, X. Sun, D. P. Wagner, A. W. Wu, and W. A. Knaus. Glasgow coma scale score in the evaluation of outcome in the intensive care unit: findings from the acute physiology and chronic health evaluation iii study. *Critical Care Medicine*, 21(10):1459–1465, 1993.

I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, 2017.

S. Blot, M. Cankurtaran, M. Petrovic, D. Vandijck, C. Lizy, J. Decruyenaere, C. Danneels, K. Vandewoude, A. Piette, G. Vershraegen, et al. Epidemiology and outcome of nosocomial bloodstream infection in elderly critically ill patients: a comparison between middle-aged, old, and very old patients. *Critical Care Medicine*, 37(5):1634–1641, 2009.

T. Ceritli, A. P. Creagh, and D. A. Clifton. Mixture of input-output hidden markov models for heterogeneous disease progression modeling. In *Workshop on Healthcare AI and COVID-19*, pages 41–53. PMLR, 2022.

I. Y. Chen, R. G. Krishnan, and D. Sontag. Clustering interval-censored time-series for disease phenotyping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6211–6221, 2022.

K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

N. Coley, S. Andrieu, M. Jaros, M. Weiner, J. Cedarbaum, and B. Vellas. Suitability of the clinical dementia rating-sum of boxes as a single primary endpoint for alzheimer's disease trials. *Alzheimer's & Dementia*, 7(6):602–610, 2011.

I. Delor, J.-E. Charoin, R. Gieschke, S. Retout, P. Jacqmin, and A. D. N. Initiative. Modeling alzheimer's disease progression using disease onset time and disease trajectory concepts applied to cdr-sob scores from adni. *CPT: Pharmacometrics & Systems Pharmacology*, 2(10):1–10, 2013.

J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1111, 2013.

A. Giannoula, A. Gutierrez-Sacristían, A. Bravo, F. Sanz, and L. I. Furlong. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. *Scientific Reports*, 8(4216), 2018.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000.

L. E. Haas, L. Van Dillen, D. de Lange, D. Van Dijk, and M. Hamaker. Outcome of very old patients admitted to the icu for sepsis: a systematic review. *European Geriatric Medicine*, 8(5-6):446–453, 2017.

A. Hayashi, Y. Mizuhara, and N. Suematsu. Embedding time series data for classification. In *Machine Learning and Data Mining in Pattern Recognition: 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005. Proceedings 4*, pages 356–365. Springer, 2005.

J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.

G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

K. W. Kim, S. Y. Woo, S. Kim, H. Jang, Y. Kim, S. H. Cho, S. E. Kim, S. J. Kim, B.-S. Shin, H. J. Kim, et al. Disease progression modeling of alzheimer's disease according to education level. *Scientific Reports*, 10(1):1–9, 2020.

C. Lee and M. van der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In *International Conference on Machine Learning*, pages 5767–5777. PMLR, 2020.

C. Lee, J. Rashbass, and M. Van der Schaar. Outcome-oriented deep temporal phenotyping of disease progression. *IEEE Transactions on Biomedical Engineering*, 68 (8):2423–2434, 2020.

C. Lee, A. Light, E. S. Saveliev, M. van der Schaar, and V. J. Gnanapragasam. Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate cancer. *npj Digital Medicine*, 5(1):110, 2022.

Q. Lei, J. Yi, R. Vaculin, L. Wu, and I. S. Dhillon. Similarity preserving representation learning for time series clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019.

J. Leitgeb, W. Mauritz, A. Brazinova, M. Majdan, I. Janciak, I. Wilbacher, and M. Rusnak. Glasgow coma scale score at intensive care unit discharge predicts the 1-year outcome of patients with severe traumatic brain injury. *European Journal of Trauma and Emergency Surgery*, 39 (3):285–292, 2013.

Q. Ma, J. Zheng, S. Li, and G. W. Cottrell. Learning representations for time series clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J. S. Reisch, R. Doody, T. A. R. Consortium, et al. Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study. *Archives of Neurology*, 65(8):1091–1095, 2008.

K. J. Ramos, P. J. Smith, E. F. McKone, J. M. Pilewski, A. Lucy, S. E. Hempstead, E. Tallarico, A. Faro, D. B. Rosenbluth, A. L. Gray, et al. Lung transplant referral for individuals with cystic fibrosis: Cystic fibrosis foundation consensus guidelines. *Journal of Cystic Fibrosis*, 18(3): 321–333, 2019.

Y. L. Rao, B. Ganaraja, B. Murlimanju, T. Joy, A. Krishnamurthy, and A. Agrawal. Hippocampus and its involvement in alzheimer's disease: a review. *3 Biotech*, 12(2): 55, 2022.

R. L. Richesson, J. Sun, J. Pathak, A. N. Kho, and J. C. Denny. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial Intelligence in Medicine*, 71:57–61, 2016.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

P. Schulam, F. Wigley, and S. Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

D. Steinley. Properties of the hubert-arable adjusted rand index. *Psychological Methods*, 9(3):386, 2004.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080, 2009.

J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487. PMLR, 2016.

Y. Yamazaki, N. Zhao, T. R. Caulfield, C.-C. Liu, and G. Bu. Apolipoprotein e and alzheimer disease: pathobiology and targeting strategies. *Nature Reviews Neurology*, 15 (9):501–518, 2019.

B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *International Conference on Machine Learning*, pages 3861–3870. PMLR, 2017.

X. Zhang, J. Chou, J. Liang, C. Xiao, Y. Zhao, H. Sarv, C. Henchcliffe, and F. Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: A cohort study. *Scientific Reports*, 9(797), 2019.

Appendix

The appendix is organized in the following structure.

A Detailed discussion of the Laplace encoder.

B Proof of Proposition 1 and relevant discussions.

C The graph-constrained K-means algorithm

D Experiment setup.

E Hyperparameter Selection.

F Complete benchmark results.

G Additional analyses of results obtained on the two real-world datasets.

A summary of major notations used in this paper is provided below.

NOMENCLATURE

$\boldsymbol{x}$	Continuous-time disease trajectory of a patient
$\boldsymbol{y}$	Label vector indicating clinical status of a patient
$\boldsymbol{t}$	A vector of time stamps
$\boldsymbol{z}$	A vector of latent variables
$\boldsymbol{w}$	A vector of Laplace embedding
$\boldsymbol{X}$	Discrete-time observation to disease trajectory $\boldsymbol{x}(t)$
Φ	A connected set of patient trajectories which represents a temporal pattern
$g(\boldsymbol{x})$	Vector-valued function that describes the conditional distribution $p(\boldsymbol{y}\|\boldsymbol{x})$
$\mathrm{d}_y(\cdot,\cdot)$	Distance metric of two label distributions
$\Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2)$	A translation from trajectory $\boldsymbol{x}^1$ to $\boldsymbol{x}^2$
$\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2)$	A translation from latent representation $\boldsymbol{z}^1$ to $\boldsymbol{z}^2$
$\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2)$	Path-based similarity score between trajectories $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$
$\mathrm{d}_\gamma(\boldsymbol{z}^1, \boldsymbol{z}^2)$	Proxy of path-based similarity score $\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2)$ in latent space
$\boldsymbol{S}$	A distance matrix of path-based similarity score between samples in a dataset
$\mathcal{G}_\delta$	A graph generated from matrix $\boldsymbol{S}$ with threshold δ
K	Number of clusters
$\mathcal{C}$	A set of K clusters
f_L	A Laplace encoder
f_E	A composite encoder with feature-wise Laplace encoders
f_P	A predictor for label distribution

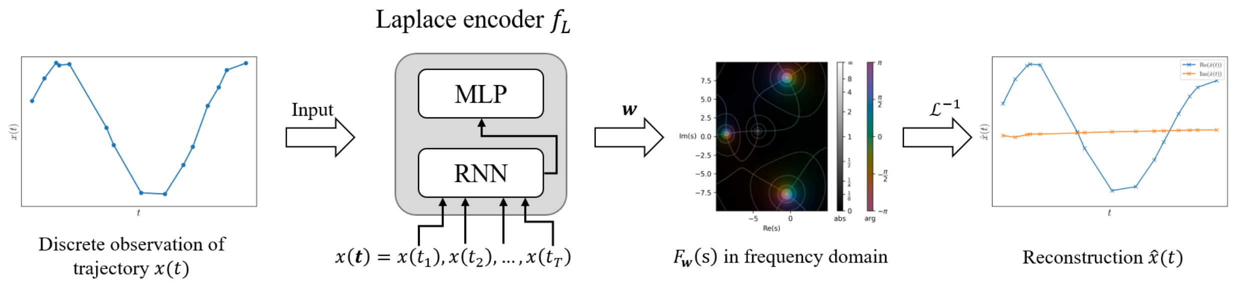
Code Availability. The source code of T-Phenotype can be found in the two GitHub repositories listed below:

- The van der Schaar lab repo: `https://github.com/vanderschaarlab/tphenotype`
- The author's personal repo: `https://github.com/yvchao/tphenotype`

A ANALYSIS OF THE LAPLACE ENCODER

A.1 Implementation Details

The proposed Laplace encoder is implemented with a RNN-based neural network f_L parameterized by θ_L. As shown in Figure A.1, given discrete time-series of a one-dimension trajectory $x(t)$, the Laplace encoder first generates a summary of time-series $x(\boldsymbol{t})$ via the RNN. With the summary as input, the MLP outputs a representation $\boldsymbol{w} \in \mathbb{C}^{n(d+1)}$. Elements in $\boldsymbol{w}$ can be divided into two groups: poles and coefficients, which are further used to construct a function in the frequency domain, $F_{\boldsymbol{w}}(s)$, as defined in (3). Changing the order of poles (and associated coefficients) in $\boldsymbol{w}$ has no effect on $F_{\boldsymbol{w}}(s)$ since it is permutation-invariant to the poles in $\boldsymbol{w}$. As discussed in the main manuscript and in the next paragraph, we impose a lexical order on poles in the MLP output $\boldsymbol{w}$ to make it a unique representation of $F_{\boldsymbol{w}}(s)$. The trajectory $x(t)$ can be reconstructed as $\hat{x}(t)$ through the inverse Laplace transform (4) on $F_{\boldsymbol{w}}(s)$. Here, the reconstruction $\hat{x}(t)$ is a function and its value can be evaluated everywhere in $t \in [0, 1]$. This allows us to compare input time-series with variable-length and irregularly-sampled observations in a unified latent space. For the sake of convenience, we denote with $\mathcal{L}^{-1}(\boldsymbol{w}) = \hat{x}(t) = \mathcal{L}^{-1}[F_{\boldsymbol{w}}(s)](t)$ the transform that maps embedding $\boldsymbol{w}$ to its time-domain reconstruction $\hat{x}(t)$.



Discrete observation of trajectory $x(t)$

$x(\boldsymbol{t}) = x(t_1), x(t_2), \ldots, x(t_T)$

$F_{\boldsymbol{w}}(s)$ in frequency domain

Reconstruction $\hat{x}(t)$

Figure A.1: Laplace Encoder.

Robust Lexical Order of Poles. Due to the summation in (3), $F_{\boldsymbol{w}}(s)$ is permutation-equivariant with respect to the poles in $\boldsymbol{w}$. Thus, we impose a lexical order ($p_m \leq p_{m+1}$ for $m = 1, \ldots, n-1$) on the poles to obtain a unique Laplace embedding $\boldsymbol{w}$ as discussed in the manuscript. To guarantee this property, we transform the unordered representation (output of the MLP in Laplace encoder) into the final unique Laplace embedding $\boldsymbol{w}$ by sorting the poles (together with their associated coefficients) in a lexical order. To achieve a stable ordering that is robust to inevitable noise in $\boldsymbol{w}$, we encourage any pair of two poles to be sufficiently different to avoid abrupt changes in their order. Hence, given two poles p_m, p_l, we say $p_m \leq p_l$ if and only if $(\mathrm{Re}(p_m) < \mathrm{Re}(p_l)) \wedge (|\mathrm{Re}(p_m) - \mathrm{Re}(p_l)| > \delta_{pole})$ or $(|\mathrm{Re}(p_m) - \mathrm{Re}(p_l)| \leq \delta_{pole}) \wedge (\mathrm{Im}(p_m) \leq \mathrm{Im}(p_l))$, and $p_m > p_l$ otherwise, where $\delta_{pole} \geq 0$ is a threshold that controls the robustness of the lexical order. The best threshold δ_{pole} is search as a hyperparameter in our experiment.

Ranges of Poles and Coefficients. Each pole p_m in embedding $\boldsymbol{w}$ is located on the complex plane $\mathbb{C}$. The real part $\mathrm{Re}(p_m)$ indicates the increase or decay speed of the corresponding component ($e^{\mathrm{Re}(p_m)t}$) in the time-domain reconstruction $\hat{x}(t) = \mathcal{L}^{-1}(\boldsymbol{w})$. Too large or small value of $\mathrm{Re}(p_m)$ leads to unrealistic signals. In the meantime, The imaginary part $\mathrm{Im}(p_m)$ represents the frequency of oscillations in the related component $(\cos(\mathrm{Im}(p_m)t) + j\sin(\mathrm{Im}(p_m)t), j^2 = -1)$ in reconstruction $\hat{x}(t)$. Very high-frequency oscillation in the input time-series $x(\boldsymbol{t})$ are usually caused by random noise and should be discarded in reconstruction $\hat{x}(t)$. In our experiment, we limit the range of poles to the area of $\{p \mid |\mathrm{Re}(p)| \leq r_{max}, |\mathrm{Im}(p)| \leq freq_{max}\}$, where r_{max} limits that maximum increase or decrease speed of signals in reconstruction $\hat{x}$, $freq_{max}$ is the maximum allowed frequency such that high-frequency signals above $freq_{max}$ are considered as a noise component in time-series $x(\boldsymbol{t})$ and, thus, discarded when constructing $F_{\boldsymbol{w}}(s)$. In our experiments, we set $r_{max} = 10$ and $freq_{max} = 20Hz$. Similarly, the coefficient $c_{m,l}$ in embedding $\boldsymbol{w}$ is limited to a square area of $\{c \mid |\mathrm{Re}(c)| \leq c_{max}, |\mathrm{Im}(c)| \leq c_{max}\}$. We set $c_{max} = 5$ which is sufficient for normalized time-series (via min-max or normal scaling). The range of poles and coefficients in $\boldsymbol{w}$ can be adjusted accordingly based on needs in practical application scenarios. When fed into the predictor network f_P, the poles and coefficients in embedding $\boldsymbol{w}$ are normalized by the corresponding maximum allowed values to facilitate the learning process.

Embedding of Static Features. In order to improve computation efficiency, when the d-th feature dimension x_d of trajectory $\boldsymbol{x}$ is known to be constant over time, i.e., $x_d(t) \equiv x_d(0)$, instead of training a Laplace encoder, the static value $x_d(0)$ is directly used to represent $x_d(t)$, and the d-th component $\boldsymbol{w}_d$ in latent variable $\boldsymbol{z}$ is replaced by $x_d(0)$.

Regularization Terms. Apart from the lexical order imposed on the embedding $\boldsymbol{w}$, we further introduce three regularization terms that encourage the Laplace encoder to provide a unique and consistent Laplace representation given an input time-series. These regularization terms are combined into the second term of $\mathcal{L}_{\text{unique}}$ in (5); we will describe each in turn.

The first regularizer, l_{sep}, penalizes the case where two poles in embedding $\boldsymbol{w}$ are nearly identical – that is, p_m and p_l are considered as an identical pole when $|p_m - p_l| \leq \delta_{pole}$ – based on the following hinge loss:

$$l_{\text{sep}}(\hat{x}(\boldsymbol{t})) = \sum_{m \neq l} \max(0, \delta_{pole} - |p_m - p_l|). \tag{10}$$

Here, p_m and p_l are two poles in the associated embedding $\boldsymbol{w}$ given the input time-series $x(\boldsymbol{t})$, i.e., $\boldsymbol{w} = f_L(x(\boldsymbol{t}))$, and the threshold $\delta_{pole} > 0$ for robust pole sorting is reused here as a pole separation threshold.

The second regularizer, l_{real}, ensures that the reconstructed trajectory $\hat{x}(t)$ is real-valued on $[0, 1]$ by suppressing the imaginary part of the reconstructed trajectory $\hat{x}(t)$ via the following loss:

$$l_{\text{real}}(\hat{x}(\boldsymbol{t})) = \frac{1}{T}\|\text{Im}(\hat{x}(\boldsymbol{t}))\|_2^2, \tag{11}$$

where $\boldsymbol{t} = [t_1, \ldots, t_T]^\top$ includes time stamps randomly sampled over $t_j \in [0, 1]$ for $j = 1, \ldots, T$. Specifically, $t_j = \text{clamp}(\frac{j}{T} + \frac{1}{2T}\varepsilon, \min = 0, \max = 1), \varepsilon \sim \text{Normal}(0, 1)$.

The last regularizer, l_{distinct}, encourages that no two distinct Laplace embeddings generate the same trajectory based on the following loss:

$$l_{\text{distinct}}(\hat{x}^i(\boldsymbol{t}), \hat{x}^j(\boldsymbol{t})) = \|\boldsymbol{w}^i - \boldsymbol{w}^j\|_2^2 e^{-\|\hat{x}^i(\boldsymbol{t}) - \hat{x}^j(\boldsymbol{t})\|_2^2}, \tag{12}$$

where the radial basis similarity function $e^{-\|\hat{x}^i(\boldsymbol{t}) - \hat{x}^j(\boldsymbol{t})\|_2^2}$ is used to discover similar trajectories, $\boldsymbol{w}^i$ and $\boldsymbol{w}^j$ are embeddings of input time-series while $\hat{x}^i(\boldsymbol{t})$ and $\hat{x}^j(\boldsymbol{t})$ are their time-domain reconstructions. Since the input time-series may be of different lengths and sampling intervals, we use the reconstructed trajectories for pair-wise comparison between time-series here.

Overall, we construct $\mathcal{L}_{\text{unique}}(\theta_L)$ as a combination of the three regularization terms introduced above:

$$\mathcal{L}_{\text{unique}}(\theta_L) = \sum_{d=1}^{\dim_x} \left(\frac{1}{N}\sum_i l_{\text{sep}}(\hat{x}_d^i(\boldsymbol{t})) + \frac{\alpha_1}{\alpha} l_{\text{real}}(\hat{x}_d^i(\boldsymbol{t})) + \frac{\alpha_2}{\alpha}\frac{1}{N(N-1)}\sum_{i \neq j} l_{\text{distinct}}(\hat{x}_d^i(\boldsymbol{t}), \hat{x}_d^j(\boldsymbol{t})) \right), \tag{13}$$

where α is the coefficient for $\mathcal{L}_{\text{unique}}(\theta_L)$ in (5), α_1 and α_2 are balancing coefficients that trade-off different uniqueness properties in the Laplace encoder. In the experiment, due to the high computational complexity, the last term l_{distinct} is only evaluated on a subset of 10 randomly selected time-series in each training batch. In addition, since l_{distinct} relies on the reconstructed time-series $\hat{x}(t)$ which may be inaccurate in the beginning of training, we fix α_2 to 0.01 such that it majorly takes effect after the reconstruction error is small enough.

A.2 Quantitative Analysis

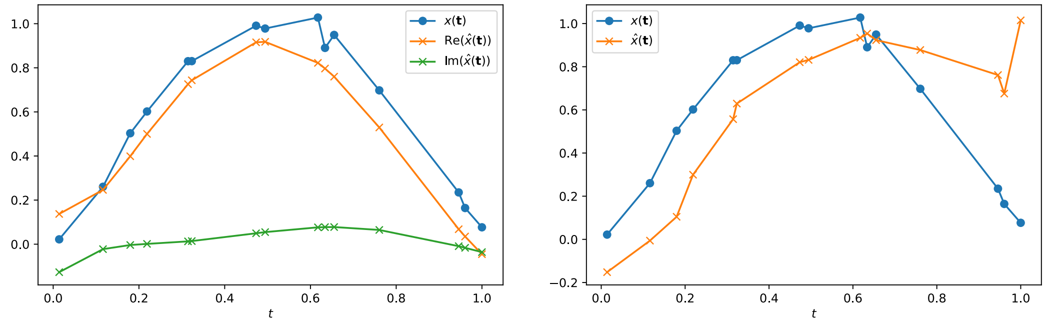
Comparison with Regular Auto-encoder. We provide a toy example to demonstrate the advantage of our proposed Laplace encoder over regular auto-encoders in time-series reconstruction. A Laplace encoder composed of a 1-layer GRU (Cho et al., 2014) and a 1-layer MLP with 10 hidden units in each layer is considered in the following discussion. Other parameters of the Laplace encoder are set as $n = 4, d = 1, \alpha = 1.0, \alpha_1 = 0.1, \alpha_2 = 0.01, \delta_{pole} = 1.0$. A regular time-series auto-encoder is used for comparison. The auto-encoder has a 1-layer GRU network as the encoder. The decoder contains a 1-layer MLP on top of another 1-layer GRU. Each layer in the auto-encoder includes 10 hidden units. The auto-encoder maps the input time-series to a latent variable. Then, the latent variable is provided to the decoder network for reconstruction of the entire time-series.

Consider a toy dataset with $N = 1000$ irregularly sampled time-series in $t \in [0, 1]$. Each sample contains $T = 15$ observations from one of the following four types of trajectories:

- Type 1: $x(t) = \cos(2\pi(t - \phi))$.
- Type 2: $x(t) = \cos(\pi(t - \phi))$.

- Type 3: $x(t) = \sin(\pi(t - \phi))$.
- Type 4: $x(t) = \sin(2\pi(t - \phi))$.

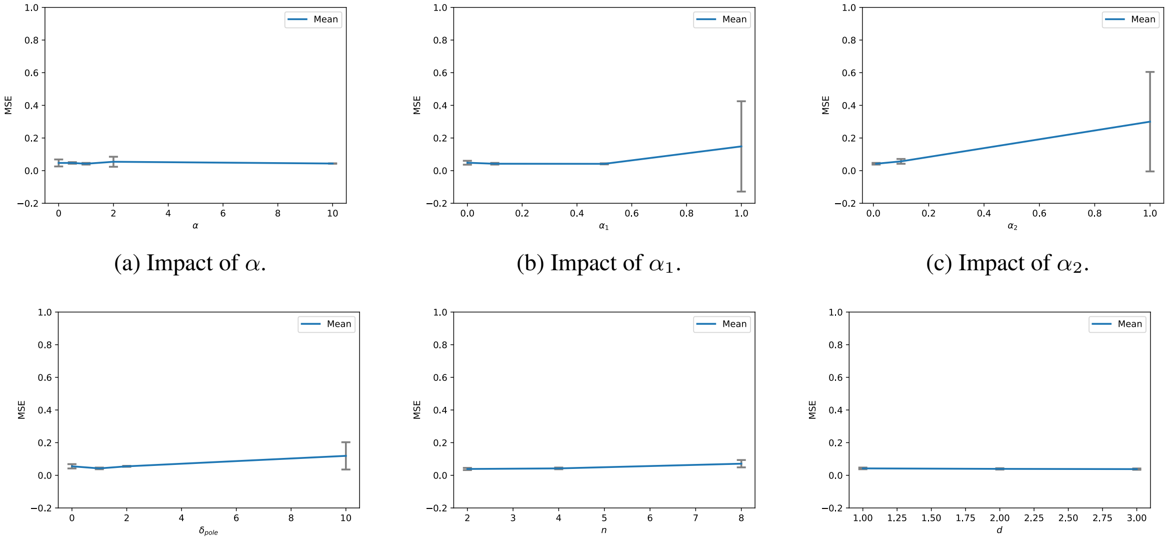
Delay term $\phi \sim \mathrm{Exp}(\frac{1}{2})$. Gaussian noise sampled from $\mathrm{Normal}(0, 0.03^2)$ is independently introduced to the observations at different time points. The mean squared error (MSE) in time-series reconstruction of the considered Laplace encoder and auto-encoder network is evaluated over 5 random splits of the toy dataset with the train/validation/test ratio of 64/16/20. Our proposed Laplace encoder achieves the best performance of $\mathrm{MSE} = 0.039 \pm 0.008$. The auto-encoder has a much higher reconstruction error of $\mathrm{MSE} = 0.108 \pm 0.019$. Comparison of typical reconstruction outcomes of the Laplace encoder and the auto-encoder is illustrated in Figure A.2.



(a) Trajectory reconstruction via Laplace encoder.

(b) Trajectory reconstruction via Auto-encoder.

Figure A.2: Comparison of Time-series Reconstruction Outcomes of Laplace Encoder and Auto-encoder.



(a) Impact of α.

(b) Impact of α_1.

(c) Impact of α_2.

(d) Impact of δ_{pole}.

(e) Impact of pole number n.

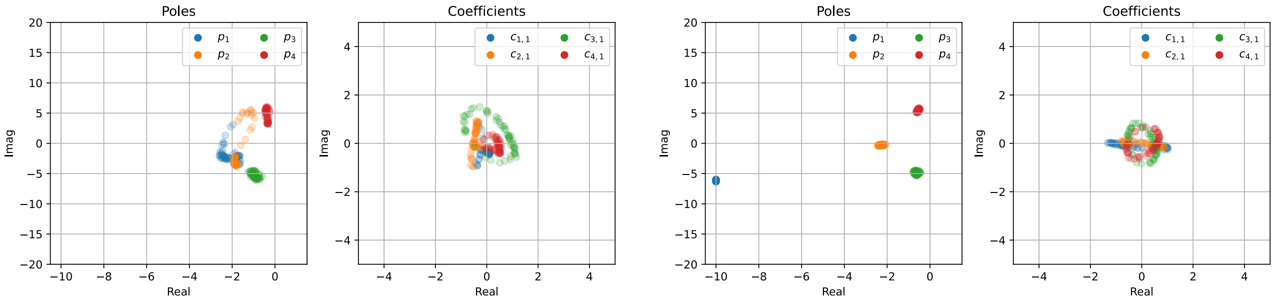
(f) Impact of maximum pole degree d.

Figure A.3: Sensitivity of Laplace Encoder with Respect to Different Hyperparameters. Error bars are calculated via evaluation on 3 random splits of the toy dataset.

Sensitivity Analysis. We further conduct a sensitivity analysis of the Laplace encoder f_L under different hyperparameters on the toy dataset. The default hyperparameters are set as $n = 4, d = 1, \alpha = 1.0, \alpha_1 = 0.1, \alpha_2 = 0.01, \delta_{pole} = 1.0$. To evaluate the impact of individual hyperparameter on the Laplace encoder, in each test, we only alter the value of one hyperparameter and keep other hyperparameters the same as default setting. The parameter sensitivity is measured via the reconstruction error (MSE), and the sensitivity test result is given in Figure A.3.

It can be found that our proposed Laplace encoder f_L has relatively stable time-series reconstruction performance under different hyperparameters. As mentioned earlier, the regularizer l_{ditinct} may generate wrong gradients in the beginning of training due to the large reconstruction error. The increased MSE for larger α_2 in Figure A.3c is within expectation, and we choose to set α_2 to 0.01 such that it only takes effect when the reconstruction error is small enough.

In addition, the effect of pole separation threshold δ_{pole} on the Laplace embedding is illustrated in Figure A.4. When $\delta_{pole} = 0.0$, the order of poles in Laplace embedding $\boldsymbol{w}$ can easily be affected by random noise in input time-series, which makes it difficult to ensure the uniqueness of $\boldsymbol{w}$. In contrast, setting $\delta_{pole} = 1.0$ effectively improves the representations learned by the Laplace encoder, and different components in the Laplace transform $F_{\boldsymbol{w}}(s)$ are clearly represented by distinct poles (marked with different colors).



(a) $\delta_{pole} = 0.0$.

(b) $\delta_{pole} = 1.0$.

Figure A.4: Distribution of Laplace Embedding $\boldsymbol{w}$ under Different Thresholds of δ_{pole}. The Laplace embeddings of trajectory $x(t) = \cos(2\pi(t - \phi)), \phi \sim \text{Exp}(\frac{1}{2})$ are plotted as poles and coefficients on the complex plane with different values of δ_{pole}.

Impact of Sampling Rate in Input Data. The Nyquist Sampling Theorem states that a band-limited signal (maximum frequency of B) can be perfectly reconstructed from sequential observations with (average) sampling rate above $2B$. It provides a lower bound on the number of time-series observations required for our proposed Laplace encoder to work. Thus, we assume that the sampling rate in real-world datasets is sufficiently large so that important temporal patterns can be correctly identified. To validate the above statement, we conduct a synthetic experiment on time-series data generated by $x(t) = \sin(2\pi t + \varphi)$ where $\varphi \sim \text{Exp}(\frac{1}{2})$ with different sampling rates. Figure A.5 demonstrates that the reconstruction error of the Laplace encoder converges to zero when the sampling rate is sufficiently large.

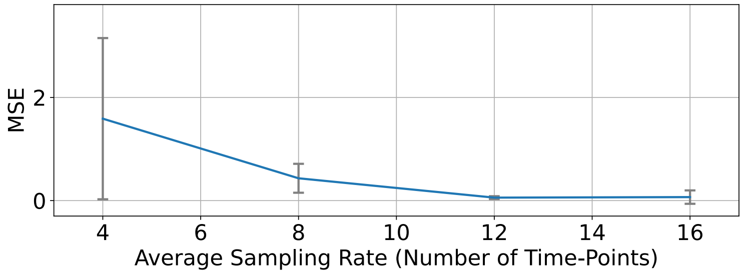


Figure A.5: Impact of Sampling Rate on Time-series Reconstruction via Laplace Encoder.

B PROOF OF PROPOSITION 1

Proposition 1 states that, given two Laplace embeddings $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ in latent space $\mathcal{Z}$, the distance between their corresponding time-domain trajectories $\hat{\boldsymbol{x}}^1$ and $\hat{\boldsymbol{x}}^2$ is upper-bounded by $\psi\|\boldsymbol{z}^1 - \boldsymbol{z}^2\|_2^2$ with some scalar $\psi > 0$. The proof of Proposition 1 can be derived as the following:

Proof. Let us first consider the uni-variate case. Given two arbitrary Laplace embeddings $\boldsymbol{w}^1, \boldsymbol{w}^2 \in \mathbb{C}^{n(d+1)}$, their time-domain reconstructions can be obtained via inverse Laplace transform, i.e., $\hat{x}^i(t) = \mathcal{L}^{-1}(\boldsymbol{w}^i) \triangleq \mathcal{L}^{-1}[F_{\boldsymbol{w}^i}(s)](t), i = 1, 2.$

According to (3) and (4), we have

$$\hat{x}^i(t) = \sum_{m=1}^{n} \sum_{l=1}^{d} \frac{c_{m,l}^i t^{l-1}}{\Gamma(l)} e^{p_m^i t}, t \geq 0, \tag{14}$$

where $\Gamma(l) = (l-1)!$ is the Gamma function, $\boldsymbol{w}^i = [p_1^i, p_2^i, \ldots, c_{1,1}^i, \ldots, c_{n,d}^i]^\top, i = 1, 2$.

Difference in One Coefficient. Suppose $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$ only differ at one coefficient $c_{m,l}$, which leads to the result $\|\boldsymbol{w}^1 - \boldsymbol{w}^2\|_2^2 = |c_{m,l}^1 - c_{m,l}^2|^2$. Then,

$$\begin{aligned} \|\hat{x}^1 - \hat{x}^2\|_{L^2_{[0,1]}}^2 &= \int_0^1 |\hat{x}^1(t) - \hat{x}^2(t)|^2 \mathrm{d}t, \\ &= \int_0^1 |c_{m,l}^1 - c_{m,l}^2|^2 \left| \frac{t^{l-1}}{\Gamma(l)} e^{p_m t} \right|^2 \mathrm{d}t, \\ &\leq |c_{m,l}^1 - c_{m,l}^2|^2 \psi_{m,l}^c = \psi_{m,l}^c \|\boldsymbol{w}^1 - \boldsymbol{w}^2\|_2^2, \end{aligned} \tag{15}$$

where $\psi_{m,l}^c$ is some suitable constant.

Difference in One Pole. Now, let us consider the case where $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$ only differ at one pole p_m which gives $\|\boldsymbol{w}^1 - \boldsymbol{w}^2\|_2^2 = |p_m^1 - p_m^2|^2$. Without loss of generality, we assume $p_m^2 - p_m^1 = r + j\theta$, where $r \leq 0, j^2 = -1$. The following inequality can be established when $t \in [0, 1]$:

$$\begin{aligned} |1 - e^{(p_m^2 - p_m^1)t}|^2 &= |1 - e^{rt}(\cos(\theta t) - j\sin(\theta t))|^2, \\ &= (1 - e^{rt})^2 + 2e^{rt}(1 - \cos(\theta t)), \\ &\leq (1 - e^{rt})^2 + e^{rt}\theta^2 t^2, && \text{(via } e^{rt} > 0 \text{ and } 1 - \cos(x) \leq \frac{x^2}{2}) \\ &\leq r^2 t^2 + e^{rt}\theta^2 t^2, && \text{(via } r \leq 0 \text{ and } 0 \leq 1 - e^{rt} \leq (-r)t) \\ &\leq (r^2 + \theta^2)t^2, \\ &= |p_m^1 - p_m^2|^2 t^2. \end{aligned} \tag{16}$$

Hence, we have

$$\begin{aligned} \|\hat{x}^1 - \hat{x}^2\|_{L^2_{[0,1]}}^2 &= \int_0^1 |\hat{x}^1(t) - \hat{x}^2(t)|^2 \mathrm{d}t, \\ &= \int_0^1 \left| \frac{c_{i,l} t^{l-1}}{\Gamma(l)} \right|^2 |e^{p_m^1 t}|^2 |1 - e^{(p_i^2 - p_i^1)t}|^2 \mathrm{d}t, \\ &\leq \int_0^1 \left| \frac{c_{i,l} t^{l-1}}{\Gamma(l)} \right|^2 |e^{p_m^1 t}|^2 |p_m^1 - p_m^2|^2 t^2 \mathrm{d}t, \\ &\leq |p_m^1 - p_m^2|^2 \psi_m^p = \psi_m^p \|\boldsymbol{w}^1 - \boldsymbol{w}^2\|_2^2, \end{aligned} \tag{17}$$

where ψ_m^p is some suitable constant.

General Cases. Now, we define an operator $S_i(\boldsymbol{w}^1, \boldsymbol{w}^2)$ that generates a new composite vector $\bar{\boldsymbol{w}}_i$ from $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$. The first i elements of the composite vector $\bar{\boldsymbol{w}}_i$ are taken from $\boldsymbol{w}^2$ while the latter $n(d+1) - i$ elements of $\bar{\boldsymbol{w}}_i$ are obtained from $\boldsymbol{w}^1$. For instance, we have $S_0(\boldsymbol{w}^1, \boldsymbol{w}^2) = \boldsymbol{w}^1$, $S_1(\boldsymbol{w}^1, \boldsymbol{w}^2) = [p_1^2, p_2^1, p_3^1, \ldots, c_{1,1}^1, \ldots, c_{n,d}^1]^\top$, $S_2(\boldsymbol{w}^1, \boldsymbol{w}^2) = [p_1^2, p_2^2, p_3^1, p_4^1 \ldots, c_{1,1}^1, \ldots, c_{n,d}^1]^\top$, …, and $S_{n(d+1)}(\boldsymbol{w}^1, \boldsymbol{w}^2) = \boldsymbol{w}^2$. It is easy to see that $S_i(\boldsymbol{w}^1, \boldsymbol{w}^2)$ and $S_{i+1}(\boldsymbol{w}^1, \boldsymbol{w}^2)$ only differ at one pole or one coefficient, and $\|S_i(\boldsymbol{w}^1, \boldsymbol{w}^2) - S_{i+1}(\boldsymbol{w}^1, \boldsymbol{w}^2)\|_2^2 = |p_{i+1}^1 - p_{i+1}^2|^2$ when $0 \leq i \leq n-1$ and $\|S_i(\boldsymbol{w}^1, \boldsymbol{w}^2) - S_{i+1}(\boldsymbol{w}^1, \boldsymbol{w}^2)\|_2^2 = |c_{m,l}^1 - c_{m,l}^2|^2$ otherwise, where $m = \lfloor \frac{i-n}{d} + 1 \rfloor, l = i - n - (m-1)d + 1$. Each composite vector $\bar{\boldsymbol{w}}_i = S_i(\boldsymbol{w}^1, \boldsymbol{w}^2)$ yields a time-domain trajectory $\mathcal{L}^{-1}(\bar{\boldsymbol{w}}_i)$ via inverse Laplace transform of $F_{\bar{\boldsymbol{w}}_i}(s)$.

Note that $\mathcal{L}^{-1}(\bar{\boldsymbol{w}}_0) = \mathcal{L}^{-1}(\boldsymbol{w}^1) = \hat{x}^1$ and $\mathcal{L}^{-1}(\bar{\boldsymbol{w}}_{n(d+1)}) = \mathcal{L}^{-1}(\boldsymbol{w}^2) = \hat{x}^2$. Based on the triangular inequality,

$$\begin{aligned}
\|\hat{x}^1 - \hat{x}^2\|^2_{L^2_{[0,1]}} &= \| \sum_{i=0}^{n(d+1)-1} \mathcal{L}^{-1}(S_i(\boldsymbol{w}^1, \boldsymbol{w}^2)) - \mathcal{L}^{-1}(S_{i+1}(\boldsymbol{w}^1, \boldsymbol{w}^2)) \|^2_{L^2_{[0,1]}} \\
&\leq \sum_{i=0}^{n(d+1)-1} \|\mathcal{L}^{-1}(S_i(\boldsymbol{w}^1, \boldsymbol{w}^2)) - \mathcal{L}^{-1}(S_{i+1}(\boldsymbol{w}^1, \boldsymbol{w}^2))\|^2_{L^2_{[0,1]}}, \\
&\leq \sum_{i=0}^{n(d+1)-1} \psi \|S_i(\boldsymbol{w}^1, \boldsymbol{w}^2) - S_{i+1}(\boldsymbol{w}^1, \boldsymbol{w}^2)\|_2^2, \\
&= \sum_{m=1}^{n} \psi |p^1_m - p^2_m|^2 + \sum_{m=1}^{n} \sum_{l=1}^{d} \psi |c^1_{m,l} - c^2_{m,l}|^2, \\
&= \psi \|\boldsymbol{w}^1 - \boldsymbol{w}^2\|_2^2,
\end{aligned} \tag{18}$$

where we take $\psi = \max_{m,l}(\psi^p_m, \psi^c_{m,l})$.

Finally, for the multivariate case, let us consider two latent embeddings $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ as well as their associated time-domain reconstructions $\hat{\boldsymbol{x}}^1$ and $\hat{\boldsymbol{x}}^2$. We define the distance between trajectories $\hat{\boldsymbol{x}}^1$ and $\hat{\boldsymbol{x}}^2$ as

$$\|\hat{\boldsymbol{x}}^1 - \hat{\boldsymbol{x}}^2\|^2_{L^2_{[0,1]}} \triangleq \sum_{d=1}^{\dim_x} \int_0^1 |\hat{x}^1_d(t) - \hat{x}^2_d(t)|^2 \mathrm{d}t, \tag{19}$$

where $\hat{x}^m_d$ is the d-th dimension of trajectory, $\hat{\boldsymbol{x}}^m = \mathcal{L}^{-1}(\boldsymbol{w}^m_d) = \mathcal{L}^{-1}[F_{\boldsymbol{w}^m_d}(s)]$ for $m = 1, 2$, $\boldsymbol{w}^m_d$ is the d-th component of $\boldsymbol{z}^m$. According to (18), we have the following bound for each dimension d.

$$\int_0^1 |\hat{x}^1_d(t) - \hat{x}^2_d(t)|^2 \mathrm{d}t = \|\hat{\boldsymbol{x}}^1_d - \hat{\boldsymbol{x}}^2_d\|^2_{L^2_{[0,1]}} \leq \psi_d \|\boldsymbol{w}^1_d - \boldsymbol{w}^2_d\|_2^2, \tag{20}$$

where $\psi_d > 0$ is some suitable scalar. Since $\|\boldsymbol{z}^1 - \boldsymbol{z}^2_2\|_2^2 = \sum_{d=1}^{\dim_x} \|\boldsymbol{w}^1_d - \boldsymbol{w}^2_d\|_2^2$, the distance between the two reconstructed trajectories $\hat{\boldsymbol{x}}^1$ and $\hat{\boldsymbol{x}}^2$ can be upper-bounded as follows with some suitable $\psi > 0$.

$$\|\hat{\boldsymbol{x}}^1 - \hat{\boldsymbol{x}}^2\|^2_{L^2_{[0,1]}} \leq \sum_{d=1}^{\dim_x} \psi_d \|\boldsymbol{w}^1_d - \boldsymbol{w}^2_d\|_2^2 \leq \psi \|\boldsymbol{z}^1 - \boldsymbol{z}^2\|_2^2. \tag{21}$$

□

Corollary 1. *Given a continuous set Φ_z in latent space, the set Φ, which consists of reconstructed trajectories of $\boldsymbol{z} \in \Phi_z$, is also a continuous set in trajectory space $\mathcal{X}$.*

Proof. Consider a trajectory $\hat{\boldsymbol{x}} \in \Phi$ and its corresponding latent embedding $\boldsymbol{z} \in \Phi_z$. For any $\varepsilon > 0$, due to the continuity of Φ_z, there must exist another embedding $\boldsymbol{z}' \in \Phi_z$ such that $\|\boldsymbol{z} - \boldsymbol{z}'\|_2^2 < \delta\varepsilon$, where $\delta > 0$ is a scalar. Let us denote the time-domain reconstruction of $\boldsymbol{z}'$ as $\hat{\boldsymbol{x}}' \in \Phi$. According to Proposition 1, $\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}'\|^2_{L^2_{[0,1]}} \leq \psi \|\boldsymbol{z} - \boldsymbol{z}'\|_2^2$ holds for some $\psi > 0$. Setting $\delta = \frac{1}{\psi}$ leads to the inequality $\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}'\|^2_{L^2_{[0,1]}} \leq \varepsilon$ which indicates the continuity of set Φ. □

Equivalent Translation in the Latent Space. Consider two trajectories $\boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathcal{X}$ with the corresponding latent embeddings $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ in the latent space. We construct a set $P_z = \{\tilde{\gamma}(\boldsymbol{z}^1 \to \boldsymbol{z}^2)\}$ of all possible continuous path $\tilde{\gamma}$ in the latent space that connects $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$. Let $g_E : \mathcal{Z} \to \mathcal{X}$ be a function that maps latent embedding $\boldsymbol{z}$ back to its time-domain reconstruction $\hat{\boldsymbol{x}}$ in the trajectory space. Then, given a translation $\Gamma(\boldsymbol{x}^1 \to \boldsymbol{x}^2)$ in the trajectory space, we can define the (approximately) equivalent translation in the latent space as

$$\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2) \triangleq \mathop{\arg\min}_{\tilde{\gamma} \in P_z} \min_{\boldsymbol{z} \in \tilde{\gamma}} \max_{\boldsymbol{x} \in \Gamma} \|\boldsymbol{x} - g_E(\boldsymbol{z})\|^2_{L^2_{[0,1]}}, \tag{22}$$

where $\min_{\boldsymbol{z} \in \tilde{\gamma}} \max_{\boldsymbol{x} \in \Gamma} \|\boldsymbol{x} - g_E(\boldsymbol{z})\|^2_{L^2_{[0,1]}}$ measures the minimum distance between translation, i.e., Γ, and the time-domain reconstruction of latent path $\tilde{\gamma}$, i.e., $\tilde{\Gamma} = \{g_E(\boldsymbol{z}) \mid \boldsymbol{z} \in \tilde{\gamma}\}$. In general, γ is the closet projection of Γ within the latent space

$\mathcal{Z}$, and the equivalence of trajectory translation is approximate. If every trajectory $\boldsymbol{x} \in \Gamma$ has a rational Laplace transform with no more than n poles and maximum degree of d as described in (3), the equivalence becomes strict. Without loss of generality, let us consider the uni-variate case. Given a translation Γ, we assume each $x \in \Gamma$ can be exactly described by the Laplace transform $F_{\boldsymbol{w}}(s)$ in (3), where $\boldsymbol{w} = f_L(x(\boldsymbol{t}))$, $\boldsymbol{t}$ is a vector of some suitable sampling time stamps. For any two trajectories $x, x' \in \Gamma$ that satisfy $|x(t) - x'(t)| \leq \delta$ almost everywhere in $t \in [0, 1]$, we have

$$\begin{aligned} |F_{\boldsymbol{w}}(s) - F_{\boldsymbol{w}'}(s)|^2 &= \left| \int_0^\infty (x(t) - x'(t)) e^{-st} \mathrm{d}t \right|^2, \\ &\leq \int_0^\infty |x(t) - x'(t)|^2 |e^{-st}|^2 \mathrm{d}t, \\ &\leq \delta^2 \int_0^\infty |e^{-st}|^2 \mathrm{d}t, \\ &= \frac{\delta^2}{2\mathrm{Re}(s)}, \end{aligned} \tag{23}$$

holds for $\mathrm{Re}(s) > 0$. When $\delta \to 0$, we have $x' \to x$ and $F_{\boldsymbol{w}'} \to F_{\boldsymbol{w}}$. Note that $F_{\boldsymbol{w}} - F_{\boldsymbol{w}'}$ is rational and can be determined with a sufficient number of observations in its region of convergence, e.g., $\mathrm{Re}(s) > 0$. The equivalence in Laplace transform, i.e., $|F_{\boldsymbol{w}}(s) - F_{\boldsymbol{w}}(s)|^2 \equiv 0$, implies that $\boldsymbol{w}' = \boldsymbol{w}$.[6] Thus, $x' \to x$ also leads to $\boldsymbol{w}' \to \boldsymbol{w}$, which means that the collection of Laplace embeddings $\{\boldsymbol{w} | \boldsymbol{w} = f_L(x(\boldsymbol{t})), x \in \Gamma\}$ is in fact a continuous path γ in the latent space. Thereby, path γ is a latent translation that exactly yields the trajectory translation Γ. Similar results can be easily extended to the multi-variate trajectory setting.

Justification for Latent Path-based Test. The path-based connectivity test $\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2)$ is defined based on the oracle model $g(\boldsymbol{x})$ of conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$. In our proposed method T-Phenotype, a predictor is built upon the Laplace embedding, i.e., $f(\boldsymbol{X}) = f_P \circ f_E(\boldsymbol{X})$, to approximate the oracle conditional distribution such that $f(\boldsymbol{X}) \approx g(\boldsymbol{x})$ given time-series $\boldsymbol{X}$ sampled from $\boldsymbol{x}$. Thus, we have $\mathrm{d}_\Gamma(\boldsymbol{x}^1, \boldsymbol{x}^2) \approx \max_{\boldsymbol{x} \in \Gamma, i=1,2} \mathrm{d}_y(f(\boldsymbol{x}(\boldsymbol{t})), f(\boldsymbol{X}^i))$, where $\boldsymbol{t}$ is a vector of some suitable observation time stamps. Further, note that translation Γ in trajectory space can be approximated by $\hat{\Gamma}$ as time-domain reconstruction of latent translation $\gamma(\boldsymbol{z}^1 \to \boldsymbol{z}^2)$ in $\mathcal{Z}$, where $\boldsymbol{z}^i = f_E(\boldsymbol{X}^i)$ for $i = 1, 2$. Then, we have

$$\max_{\boldsymbol{x} \in \Gamma, i=1,2} \mathrm{d}_y(f(\boldsymbol{x}(\boldsymbol{t})), f(\boldsymbol{X}^i)) \approx \max_{\hat{\boldsymbol{x}} \in \hat{\Gamma}, i=1,2} \mathrm{d}_y(f(\hat{\boldsymbol{x}}(\boldsymbol{t})), f(\boldsymbol{X}^i)) \approx \max_{\boldsymbol{z} \in \gamma, i=1,2} \mathrm{d}_y(f_P(\boldsymbol{z}), f_P(\boldsymbol{z}^i)), \tag{24}$$

which leads to the latent path-based test in (7).

C GRAPH-CONSTRAINED K-MEANS ALGORITHM IN T-PHENOTYPE

The graph-constrained K-means iteration in Algorithm 1 is provided in Algorithm C.1. After each run via GK-means, the objective function J in (2) is re-evaluated. The main algorithm of T-Phenotype stops after 5 iterations with no improvement in objective J under maximum of 1,000 iterations. Alternatively, T-Phenotype stops when the improvement is below certain tolerance $\mathrm{tol} = 10^{-7}$, i.e., $|\Delta J| \leq \mathrm{tol}$.

D EXPERIMENT SETUP

D.1 Datasets and Statistics

For the two real-world medical datasets, we want to capture recent temporal patterns and associated target outcomes. Thus, we utilize a sliding window of size 6 years and 24 hours to extract sub-sequences containing temporal predictive patterns among most recent observations for ADNI and ICU datasets, respectively. Statistics of major feature variables in the ADNI dataset and ICU dataset can be found in Table D.1 and Table D.2, respectively.

D.2 Baselines

We compare the performance of T-Phenotype with the following five benchmarks ranging from traditional method to state-of-the-art deep learning-based methods, where each clustering method reflects a different notion of temporal phenotypes:

[6] When $F_{\boldsymbol{w}}(s)$ and $F_{\boldsymbol{w}'}(s)$ have less than n poles, $\boldsymbol{w}$ and $\boldsymbol{w}'$ may take value from multiple alternative embeddings. However, we can always select the combination such that $\boldsymbol{w}' = \boldsymbol{w}$.

Algorithm C.1 GK-means (Single K-means iteration over similarity graph $\mathcal{G}_\delta$)

```
Input: $J, e_1, e_2, \ldots, e_K, \mathcal{G}_\delta$    ▷ $J$ objective, $e_k$ cluster seed, $\mathcal{G}_\delta$: similarity graph
Output: $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$
for $k = 1, 2, \ldots, K$ do
    $\boldsymbol{v}_k, \boldsymbol{X}^{(k)} \leftarrow e_k$
    $C_k \leftarrow \{\boldsymbol{X}^{(k)}\}$    ▷ Initialize cluster $C_k$ with seed $e_k$
end for
$D_{\text{free}} \leftarrow \{\boldsymbol{X} | \boldsymbol{X} \notin C_k, \forall C_k \in \mathcal{C}\}$    ▷ Get the set of unclustered samples
while $|D_{\text{free}}| > 0$ do
    for $\boldsymbol{X} \in D_{\text{free}}$ do
        $C^* \leftarrow \arg\min_{C_k \in \mathcal{C}, \boldsymbol{X} \overset{\mathcal{G}_\delta}{\longleftrightarrow} C_k} \mathrm{d}_y(f(\boldsymbol{X}), \boldsymbol{v}_k)$    ▷ Find the best cluster assignment
        $C^* \leftarrow C^* \cup \{\boldsymbol{X}\}$
        $D_{\text{free}} \leftarrow D_{\text{free}} \setminus \{\boldsymbol{X}\}$
    end for
    for $k = 1, 2, \ldots, K$ do
        $\boldsymbol{v}_k \leftarrow \frac{1}{|C_k|} \sum_{\boldsymbol{X} \in C_k} f(\boldsymbol{X})$    ▷ Update cluster centroid
    end for
end while
```

Table D.1: Statistics of ADNI Dataset.

STATIC COVARIATES		**TYPE**	**MEAN**	**MIN/MAX (MODE)**		**TYPE**	**MEAN**	**MIN/MAX (MODE)**
Demographic	Race	Cat.	0.93	White	Ethnicity	Cat.	0.97	Not Hisp/Latino
	Education	Cat.	16.13	16	Marital Status	Cat.	0.75	Married
Genetic	APOE $\varepsilon 4$	Cat.	0.44	0				
TIME-VARYING COVARIATES		**TYPE**	**MEAN**	**MIN/MAX (MODE)**		**TYPE**	**MEAN**	**MIN/MAX (MODE)**
Demographic	Age	Cont.	73.62	55/91.4				
Biomarker	Entorhinal	Cont.	3.6E+3	1.0E+3/6.7E+3	Mid Temp	Cont.	2.0E+4	8.9E+3/3.2E+4
	Fusiform	Cont.	1.7E+4	9.0E+3/2.9E+4	Ventricles	Cont.	4.1E+4	5.7E+3/1.6E+5
	Hippocampus	Cont.	6.9E+4	2.8E+3/1.1E+4	Whole Brain	1.0E+6	6.5E+5/1.5E+6	
	Intracranial	Cont.	1.5E+6	2.9E+2/2.1E+6				
Cognitive	CDRSB	Cont.	1.21	0.0/17.0	Mini Mental State	Cont.	27.84	2.0/30.0
	ADAS-11	Cont.	8.58	0.0/70.0	ADAS-13	Cont.	13.60	0.0/85.0
	RAVLT Immediate	Cont.	38.26	0.0/75.0	RAVLT Learning	Cont.	4.65	-5.0/14.0
	RAVLT Forgetting	Cont.	4.19	-12.0/15.0	RAVLT Percent	Cont.	51.68	-500.0/100.0

K-means with Warping-based Distance. The technique of dynamic time warping (DTW) provides one way to measure time-series similarity regardless of the observation interval. Time-series with similar temporal patterns usually leads to smaller DTW distances. We apply conventional K-means with the DTW-based similarity measure to discover clusters representing different temporal patterns. We denote this approach as KM-DTW.

Deep Temporal K-means. Embedding (i.e., hidden representations) from RNNs can provide meaningful information to measure the similarity between time-series. With the encoder-predictor (E2P) structure introduced in (Lee and van der Schaar, 2020), we include the baseline of KM-E2P that performs clustering in a representation space via K-means. We denote the baseline as KM-E2P(z) when The representation space is formed by the latent embeddings from an encoder network. The discovered cluster will capture both similarities in input time-series and the output label prediction due to the E2P structure. When the representation space is selected to be the output (label prediction) of the predictor network, we refer to the method as KM-E2P(y). In this case, the discovered clusters are aligned to major modes in the label distribution and are not necessarily associated with certain temporal patterns in trajectory space.

K-means with Laplace Encoder. Similar to the baseline of KM-DTW, the time-series embedding from Laplace encoder provides a unified representation of (potentially) irregularly sampled time-series. The Euclidean distance between Laplace embeddings can thus be used as a similarity measure for different patient trajectories. In practice, the longitudinal observations of patients are first converted to a latent space via the Laplace encoder. Then, K-means algorithm is performed over the latent representations to identify patient subgroups based on their similarity in temporal patterns.

Toward K-means Friendly Spaces using Sequence-to-sequence. Sequence-to-sequence (SEQ2SEQ) learning paradigm

Table D.2: Statistics of ICU Dataset.

Static Covariates		**TYPE**	**MEAN**	**MIN/MAX (MODE)**		**TYPE**	**MEAN**	**MIN/MAX (MODE)**
Demographic	Age	Cont.	67.25	15.0/90.0	Gender	Cat.	0.56	Male
Admission	ICU Type	Cat.	2.76	Medical ICU				
TIME-VARYING COVARIATES		**TYPE**	**MEAN**	**MIN/MAX (MODE)**		**TYPE**	**MEAN**	**MIN/MAX (MODE)**
Blood Test	Albumin	Cont.	2.92	1.0/5.3	ALP	Cont.	1.2E+2	1.2E+1/2.2E+3
	ALT	Cont.	3.9E+3	1.0/1.2E+4	AST	Cont.	5.1E+2	4.0/1.8E+4
	Bilirubin	Cont.	2.91	0.1/47.7	BUN	Cont.	27.41	0.0/197.0
	Cholesterol	Cont.	156.52	28.0/330.0	Creatinine	Cont.	1.50	0.1/22.1
	Glucose	Cont.	1.4E+3	1.0E+1/1.1E+3	Lactate	Cont.	2.88	0.3/29.3
	HCO3	Cont.	23.12	5.0/50.0	pH	Cont.	7.49	1.0/735.0
	K	Cont.	4.14	1.8/22.9	Mg	Cont.	2.03	0.6/9.9
	Na	Cont.	139.07	98.0/177.0	HCT	Cont.	30.69	9.0/61.8
	TroponinI	Cont.	7.15	0.3/49.2	TroponinT	Cont.	1.20	0.01/24.91
	Platelets	Cont.	1.9E+2	6.0/1.0E+3	White Blood Cell	Cont.	12.67	0.1/187.5
Monitoring	Heart Rate	Cont.	86.80	0.0/199.5	Respiratory Rate	Cont.	19.64	0.0/98.0
	SysABP	Cont.	119.57	0.0/273.0	NISysABP	Cont.	119.20	0.0/247.0
	DiasABP	Cont.	59.54	0.0/268.0	NIDiasABO	Cont.	58.18	0.0/180
	MAP	Cont.	80.23	0.0/295.0	NIMAP	Cont.	77.13	0.0/194.0
	GCS	Cont.	11.41	3.0/15.0	Temperature	Cont.	37.07	-17.8/42.1
	Urine	Cont.	12E+2	0.0/1.1E+5				
Oxygen	FiO2	Cont.	0.54	0.21/1.0	PaCO2	Cont.	40.41	11.0/100.0
	PaO2	Cont.	147.82	0.0/500.0	SaO2	Cont.	96.65	26.0/100.0

allows the learning of a representation space that is easier to perform clustering compared to the original time-series data. Such baseline reflects the recent trend of combining conventional clustering methods, e.g., K-means, with dimension reduction using deep learning technique (Xie et al., 2016; Baytas et al., 2017). With different temporal patterns encoded in a low-dimension representation space, K-means clustering is applied to discover clusters that represent various temporal feature interactions in input time-series data. In the experiment, we use a modified version of DCN (Yang et al., 2017) as the SEQ2SEQ baseline.

AC-TPC. AC-TPC (Lee and van der Schaar, 2020) is one of the state-of-the-art temporal clustering approach that discovers outcome-oriented clusters. AC-TPC learns a cluster assignment policy in the latent space based on an encoder network. The cluster assignment policy is trained with the actor-critic loss from reinforcement learning to find the optimal clusters that represent typical label distributions learned by a predictor network. Similar to KM-E2P(y), there is no guarantee on the association between temporal patterns and clusters discovered by AC-TPC.

D.3 Training Procedure of T-Phenotype

To fit the model of T-Phenotype on a dataset, the Laplace encoder for each trajectory dimension is firstly pre-trained based on (5) calculated at each time step. Then, we fit the predictor f_L with observed patient outcomes y. Finally, the temporal clusters are discovered via graph-constrained K-means algorithm C.1 based on the output from the predictor. Latent embeddings from the Laplace encoder has a clear mathematical meaning. Thus, we freeze the pre-trained Laplace encoder to be isolated from gradients due to outcome predictions. Joint optimization of the encoder and predictor may lead to slower convergence and lower performance as shown in Table 2 and Table 3 with "T-Phenotype (J)" as the ablation study.

D.4 Performance Metrics

Prediction Performance. Area under the curve of receiving-operator characteristic (AUROC) and area under the curve of precision-recall (AUPRC) are used to assess the prognostic value of the discovered clusters on predicting the target label $\boldsymbol{y}$. For non-binary (category larger than 2) labels, these scores are calculated individually for each category and averaged over the entire categories.

Clustering Performance. For synthetic data, we evaluate the clustering performance in terms of the purity score (Lee and van der Schaar, 2020), adjusted Rand index (RAND) (Steinley, 2004), and normalized mutual information (NMI) (Vinh et al., 2009) as the ground-truth cluster label is available. For the real-world dataset, there is no ground-truth of cluster label. In such a case, the Silhouette coefficient (Rousseeuw, 1987) is commonly used as a measure of cluster consistency by assessing the homogeneity within each cluster and heterogeneity across different clusters. More specifically, the traditional Silhouette index assumes convex clusters and uses the average intra-cluster distance (a) and inter-cluster distance (b) to evaluate the consistency between cluster assignment and pattern distribution as $s = \frac{|b-a|}{\max(a,b)}$. Averaging s over all samples gives the Silhouette index S.

In this paper, the clusters are identified via predictive temporal patterns and are not necessarily in convex shapes. To better reflect our new notion of clusters, we instead use an m-nearest neighbor version of Silhouette index, i.e., S^m. Specifically, suppose there are K clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$. Given a time-series $\boldsymbol{X}$ in cluster C_k, we only consider its m nearest samples in the corresponding cluster when calculating intra- and inter-cluster distances a^m and b^m as given below:

$$a^m = \frac{1}{|N_m(\boldsymbol{X}, C_k)|} \sum_{\boldsymbol{X}' \in N_m(\boldsymbol{X}, C_k)} \|\boldsymbol{X} - \boldsymbol{X}'\|_2^2, \quad b^m = \min_{i \neq k} \frac{1}{|N_m(\boldsymbol{X}, C_i)|} \sum_{\boldsymbol{X}' \in N_m(\boldsymbol{X}, C_i)} \|\boldsymbol{X} - \boldsymbol{X}'\|_2^2, \tag{25}$$

where $N_m(\boldsymbol{X}, C_k)$ indicates the set of m nearest neighbors of $\boldsymbol{X}$ in cluster C_k. Then, the clustering consistency in our variant Silhouette index is calculated as $s^m = \frac{|b^m - a^m|}{\max(a^m, b^m)}$. The average score S^m of all samples is used to measure the overall clustering consistency. Note that when $m \geq \max_{C_k \in \mathcal{C}} |C_k|$, the variant S^m is identical to the original Silhouette index, i.e., $S^m = S$.

Focusing on m closest samples allows us to effectively evaluate pattern consistency in non-convex and irregularly shaped clusters. Nonetheless, when multiple temporal patterns are put into the same cluster, S^m may still generate a high score due to the focus on local similarity. To address this issue, we use another connectivity-based metric P^m to evaluate the purity of a cluster in terms of temporal patterns. Consider a cluster C_k, a connectivity graph over time-series in C_k can be derived via m-nearest neighbor discovery. We use the count p_k of connected subgraphs to estimate the number of temporal pattern included in cluster C_k and calculate the temporal pattern purity via $P^m = \frac{1}{K} \sum_{C_k \in \mathcal{C}} \frac{1}{p_k}$. It is clear that $P^m = 1$ when m is sufficiently large and each cluster only contains a single temporal pattern, and $P^m = \frac{1}{K} \sum_{C_k \in \mathcal{C}} \frac{1}{|C_k|}$ when $m = 0$.

To get an overall assessment of cluster consistency, we normalize S^m into $[0, 1]$ and calculate the summary metric *AUSIL* as the area under the curve of S^m verses P^m for $m = 1, 2, \ldots, M, M \in \mathbb{N}$. For the evaluation of phenotype discovery, we combine the prediction accuracy (AUROC and AUPRC) and cluster consistency (AUSIL) into two composite metrics H_{ROC} and H_{PRC}. Similar to the F1-score in classification, these composite metrics are defined respectively as

$$H_{\text{ROC}} \triangleq 2\frac{\text{AUROC} \cdot \text{AUSIL}}{\text{AUROC} + \text{AUSIL}}, \quad H_{\text{PRC}} \triangleq 2\frac{\text{AUPRC} \cdot \text{AUSIL}}{\text{AUPRC} + \text{AUSIL}}. \tag{26}$$

E HYPERPARAMETER SELECTION

In the experiment, T-Phenotype, KM-E2P(y), KM-E2P(z) are implemented with PyTorch and are trained with learning rate of 0.1 in 50 epochs. AdamW optimizer is used to tune the network parameters. The K-means clustering in KM-E2P(y), KM-E2P(z) and KM-DTW is performed with K-means++ initialization based on implementation in PyClustering.[7] The baselines of AC-TPC and SEQ2SEQ are implemented in TensorFlow. They are trained with Adam optimizer with training epochs set to 200 due to different learning rates in their implementation.

We perform hyperparameter selection on each dataset via 3-fold cross-validation. For T-Phenotype, the best hyperparameters of the Laplace encoders are searched to minimize the average reconstruction error over all temporal dimensions. For each real-world dataset, the best number of clusters K is searched via maximizing the composite metric H_{PRC} of T-Phenotype. The selected best cluster number K is used for all baselines on the same dataset. For baselines of KM-E2P(y) and KM-E2P(z), the hyperparameters for each dataset are search to maximize H_{PRC} (or purity score on the synthetic dataset) given the selected cluster number K. The hyperparameters of AC-TPC and SEQ2SEQ are set to be the same with the original implementation in (Lee and van der Schaar, 2020) (dropout layers are disabled to ensure reproducibility). The hyperparameter space considered in our experiment is discussed as follows.

E.1 Hyperparameter Selection of T-Phenotype

Laplace Encoder. In the experiment, each Laplace encoder f_L in T-Phenotype contains a 1-layer GRU and a 1-layer MLP with 10 hidden units in each layer. Given a time-series input, each Laplace encoder generates an embedding with $n = 4$ poles and maximum degree of $d = 1$. As mentioned earlier, coefficient α_2 for regularization term l_{distinct} is set to 0.01 throughout the experiment. The rest hyperparameters are searched in the parameter space as follows.

- Coefficient for pole separation loss l_{sep}: $\alpha \in \{1.0, 10.0\}$.

[7] https://pyclustering.github.io/

- Coefficient loss l_{real}: $\alpha_1 \in \{0.1, 1.0\}$.
- Threshold for pole sorting and the separation loss: $\delta_{pole} \in \{1.0, 2.0\}$.

To address the complex temporal patterns in the ICU dataset, the maximum degree of poles d is also added to the search space, and the range of $d \in \{1, 2\}$ is considered. The best hyperparameter for Laplace encoder on the three datasets are given as follows.

- Synthetic dataset: $\alpha = 1.0, \alpha_1 = 0.1, \delta_{pole} = 1.0$.
- ADNI dataset: $\alpha = 1.0, \alpha_1 = 0.1, \delta_{pole} = 2.0$.
- ICU dataset: $\alpha = 1.0, \alpha_1 = 0.1, \delta_{pole} = 2.0, d = 2$.

Predictor. The predictor f_P is composed of a 3-layer MLP with 10 hidden units in each layer.

Cluster Number K. The best number of K for each dataset is selected based on the optimal Laplace encoder and predictor structures selected above. We use the ground truth cluster number $K = 3$ for the synthetic dataset. For the two real-world datasets, the cluster number is searched among $K \in \{2, 3, 4, 5\}$ to maximize the composite clustering performance H_{PRC}. The optimal cluster number selection result is given below.

- Synthetic dataset: $K = 3$ (we directly use the ground truth).
- ADNI dataset: $K = 4$.
- ICU dataset: $K = 3$.

E.2 Hyperparameter Selection of Baselines

KM-E2P(y). The KM-E2P(y) model includes a 1-layer GRU network to extract temporal features from input time-series. A 2-layer MLP is stacked on top of the GRU network to form an encoder. Given the encoder output, another 2-layer MLP is used to predict the categorical label $\boldsymbol{y}$. All layers in the GRU and MLP share the same number h of hidden units. Hyperparameters of $h \in \times\{5, 10, 20\}$ is searched in each dataset basedd on the corresponding K determined above. By maximizing the composite metric H_{PRC} or purity score, the hyperparameter selection result is obtained as follows.

- Synthetic dataset: $h = 20$.
- ADNI dataset: $h = 20$.
- ICU dataset: $h = 20$.

KM-E2P(z). Similar to KM-E2P(y), the KM-E2P(z) model is composed of a encoder with 2-layer MLP on top of a 1-layer GRU network to extract temporal features from input time-series. The encoder outputs a r-dimension latent vector, which is then used by a 2-layer MLP-based predictor for label prediction. All layers in the GRU and MLP share the same number h of hidden units. Given the best cluster numbers of K found by T-Phenotype, on each dataset, the optimal combination of h and r are search in the space of $(h, r) \in \{10, 20\} \times \{5, 10, 20\}$ to maximize the composite metric H_{PRC} or purity score when ground truth cluster label is available. The hyperparameter selection result is given as follows.

- Synthetic dataset: $h = 10, r = 10$.
- ADNI dataset: $h = 10, r = 20$.
- ICU dataset: $h = 20, r = 10$.

KM-$\mathcal{L}$. The baseline KM-$\mathcal{L}$ simply shares hyperparameters with T-Phenotype for its Laplace encoders on each dataset.

F COMPLETE BENCHMARK RESULT

The complete benchmark result on synthetic dataset is shown in Table F.1. T-Phenotype has significantly better clustering performance (purity score, adjusted Rand index, normalized mutual information) over all baselines on the synthetic data. In the meantime, the advantage of T-Phenotype over other baselines (except for AC-TPC) is clearly demonstrated via the proposed phenotype discovery performance metrics of H_{ROC} and H_{PRC}. An extra baseline of KM-Laplacian (K-means on graph Laplacian calculated via dynamic time warping) is included in Table F.1 for reference. We note that this method has two major drawbacks: 1) there is not a stable and consistent representation space for cluster assignment for new samples; and 2) the distance matrix computation complexity in dynamic time warping could be extremely high, which makes this baseline infeasible for the two real-world datasets.

Table F.1: Complete Benchmark Result on the Synthetic Dataset.

METHOD	AUROC	AUPRC	PURITY	RAND	NMI	H_{ROC}	H_{PRC}
KM-E2P(y)	0.973±0.014	**0.962±0.019**	0.663±0.019	0.477±0.033	0.569±0.045	0.846±0.012	0.842±0.010
KM-E2P(z)	0.963±0.012	0.948±0.011	0.677±0.029	0.418±0.024	0.485±0.047	0.879±0.011	0.873±0.009
KM-DTW	0.722±0.033	0.649±0.028	0.469±0.017	0.068±0.021	0.077±0.022	0.787±0.020	0.742±0.019
KM-Laplacian	0.736±0.024	0.663±0.017	0.490±0.021	0.086±0.011	0.094±0.010	0.797±0.016	0.752±0.013
KM-$\mathcal{L}$	0.646±0.030	0.593±0.027	0.687±0.033	0.395±0.058	0.447±0.059	0.735±0.020	0.700±0.017
SEQ2SEQ	0.507±0.028	0.505±0.014	0.378±0.008	-0.003±0.003	0.005±0.003	0.630±0.022	0.628±0.011
AC-TPC	0.966±0.012	0.952±0.017	0.659±0.020	0.487±0.035	0.596±0.043	**0.931±0.011**	**0.925±0.014**
T-Phenotype (J)	0.967±0.020	0.954±0.025	0.655±0.021	0.440±0.051	0.543±0.064	0.845±0.064	0.840±0.064
T-Phenotype	**0.975±0.013**	0.960±0.024	**0.965±0.018**‡	**0.902±0.048** ‡	**0.875±0.050**‡	0.927±0.010	0.920±0.014

Best performance is highlighted in **bold**. Symbol ‡ indicates p-value < 0.01

The complete benchmark result on two real-world datasets is provided in Table F.2. T-Phenotype in general has the best (or second best) phenotype discovery performance (H_{ROC} and H_{PRC}) while achieving high accuracy in outcome prediction (AUROC and AUPRC), which demonstrates the prognostic value of the phenotypes discovered by T-Phenotype.

Table F.2: Complete Benchmark Result on Two Real-world Datasets.

	METHOD	AUROC	AUPRC	AUSIL	H_{ROC}	H_{PRC}
ADNI	KM-E2P(y)	**0.893±0.005**	**0.728±0.017**	0.677±0.019	0.770±0.013	0.701±0.012
	KM-E2P(z)	0.884±0.012	0.711±0.020	0.672±0.028	0.763±0.018	0.690±0.013
	KM-DTW	0.743±0.013	0.522±0.020	0.762±0.049	0.752±0.027	0.618±0.021
	KM-$\mathcal{L}$	0.697±0.029	0.465±0.019	**0.820±0.022**‡	0.753±0.019	0.593±0.018
	SEQ2SEQ	0.775±0.023	0.550±0.030	0.772±0.014	0.773±0.012	0.642±0.022
	AC-TPC	0.861±0.012	0.665±0.020	0.726±0.020	0.788±0.014	0.694±0.013
	T-Phenotype (J)	0.867±0.020	0.679±0.040	0.690±0.007	0.768±0.011	0.684±0.021
	T-Phenotype	0.891±0.005	0.716±0.015	0.711±0.023	**0.791±0.013**	**0.713±0.009**‡
ICU	KM-E2P(y)	**0.697±0.014**	0.593±0.012	0.668±0.046	0.682±0.029	0.628±0.025
	KM-E2P(z)	0.677±0.030	0.579±0.018	0.698±0.042	0.686±0.031	0.633±0.024
	KM-DTW	0.539±0.030	0.515±0.011	0.786±0.072	0.636±0.023	0.621±0.021
	KM-$\mathcal{L}$	0.577±0.019	0.532±0.009	**0.834±0.024**	0.682±0.009	0.649±0.004
	SEQ2SEQ	0.592±0.024	0.539±0.012	0.830±0.016	0.690±0.011	**0.653±0.004**
	AC-TPC	0.660±0.008	0.573±0.003	0.735±0.024	0.695±0.014	0.644±0.011
	T-Phenotype (J)	0.697±0.025	**0.595±0.017**	0.691±0.091	0.691±0.056	0.636±0.048
	T-Phenotype	0.681±0.017	0.585±0.015	0.726±0.015	**0.703±0.007**	0.648±0.008

Best performance is highlighted in **bold**. Symbol ‡ indicates p-value < 0.01

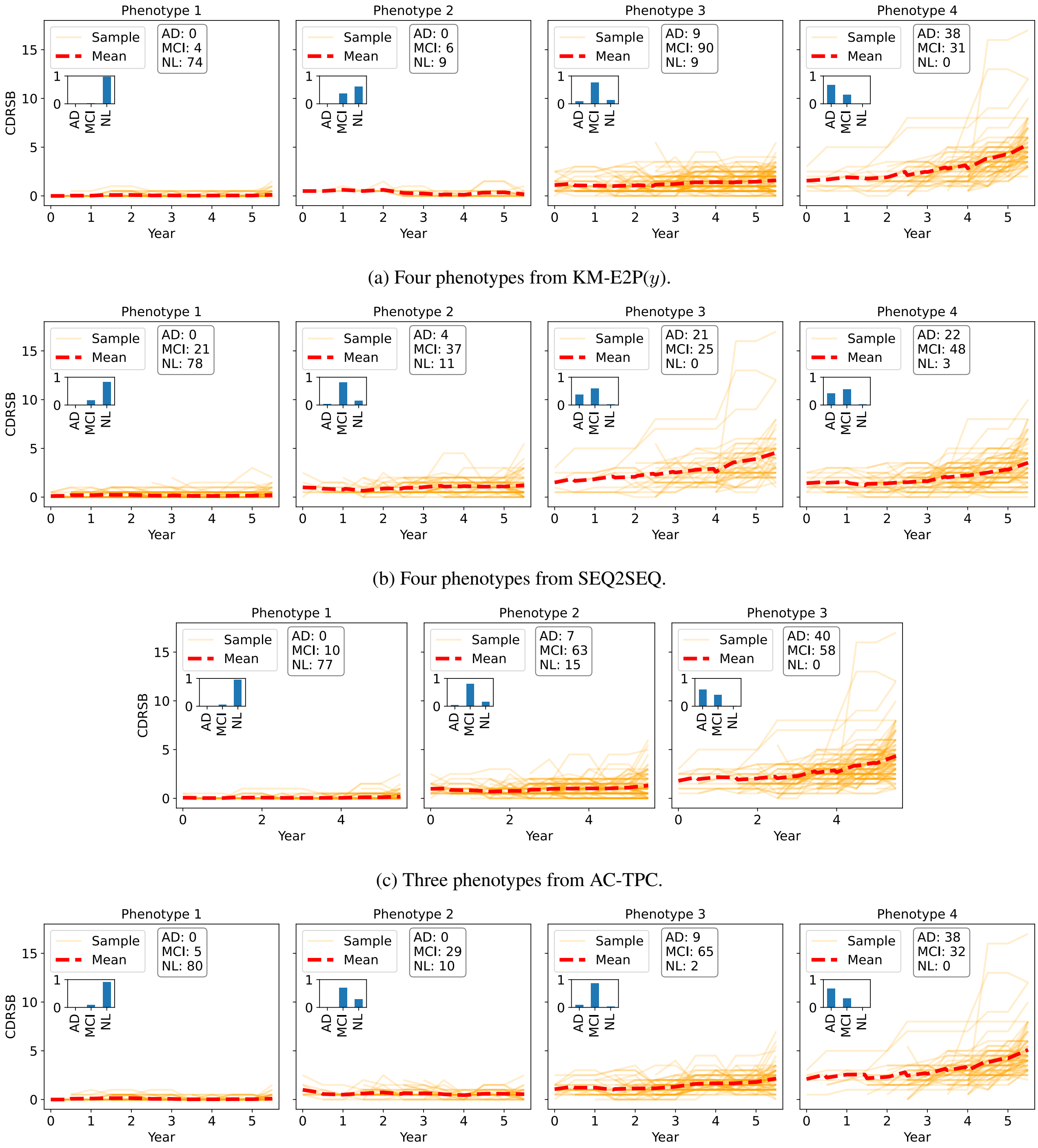
G FURTHER ANALYSIS ON PHENOTYPE DISCOVERY

Comparison of Cluster Assignments on ADNI Dataset. On the ADNI dataset, typical phenotypes from KM-E2P(y), SEQ2SEQ, AC-TPC and T-Phenotype are compared in Figure G.1. Due to the model design, KM-E2P(y) only focuses on the predicted outcome distribution when discovering phenotypes (as shown in Figure G.1a). Compared to T-Phenotype, KM-E2P(y) wrongly splits normal patients with the same temporal pattern (stable CDRSB trajectory) into two clusters under $K = 4$. Additionally, KM-E2P(y) fails to discover the two subtypes of patients with high-risk of MCI as illustrated in phenotype 2 and 3 in Figure G.1d. While the SEQ2SEQ method is able to capture temporal patterns exhibit in patient trajectories, it is incapable to properly associate these temporal patterns with patient outcomes. For instance, SEQ2SEQ wrongly splits high-risk patients with increasing CDRSB scores over time into two different subgroups with similar outcome distributions.

As discussed in the main manuscript, AC-TPC aims at discovering the minimum number of clusters that can sufficiently represent the outcome distribution. Thus, it only identifies three phenotypes under $K = 4$ and combines the two subtypes (Phenotype 2 and 3 in Figure G.1d) of MCI patients into the same cluster. In comparison, T-Phenotype discovers phenotypes based on both predicted outcome and the associated predictive temporal patterns. The two subgroups of patients with expected diagnosis of MCI are correctly identified by T-Phenotype, which demonstrates the prognostic value of our method over the considered baselines.

Phenotypes on ICU Mortality. On the ICU dataset, T-Phenotype is applied to identify phenotypes based on the patient's age, gender, GCS score and the fraction of PaCO2. Three major phenotypes are discovered by T-Phenotype, and the GCS trajectories of test samples in each subgroup are illustrated in Figure G.2. Based on the stability of their GCS trajectory, patients in each phenotype are plotted separately in two subfigures. The GCS score is predictive of patient mortality after ICU discharge (Leitgeb et al., 2013) and shows good discrimination accuracy on high- and low-risk patients admitted to ICU (Bastos et al., 1993). The predicted mortality rates in phenotypes 1, 2 and 3 are 15.3%, 3.2% and 32.4%, respectively. The GCS levels of patients in the three subgroups manifest a clear association to their corresponding mortality risks. For instance, many patients of phenotype 3 had lower GCS score (below 10) than the two other subgroups. In contrast, while having higher GCS levels, many patients in Phenotype 1 and 2 had an increase pattern (as shown in Figure G.2b) in their recent GCS measurements, which potentially contributes to their decreased risks of death. In the meantime, age is reported to be another risk factor for ICU mortality (Blot et al., 2009; Haas et al., 2017). With the average patient age of 63.0 (IQR: 53.0 – 76.0), 43.0 (IQR: 29.8 – 55.3) and 70.6 (IQR: 62.0 – 82.0) in the three identified subgroups, phenotype 1 and 2 are clearly separated.[8]

[8] Interquartile range (IQR) is the range defined by 25% and 75% quantiles of a variable.



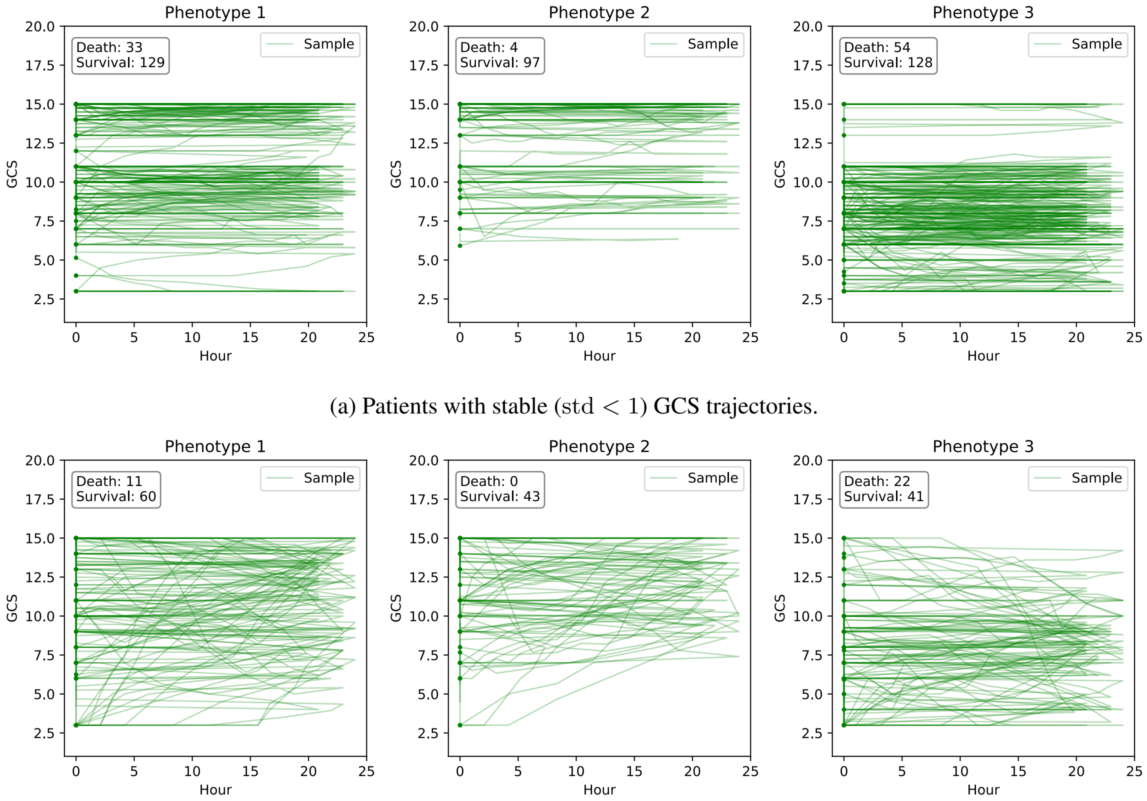
(a) Four phenotypes from KM-E2P(y).

(b) Four phenotypes from SEQ2SEQ.

(c) Three phenotypes from AC-TPC.

(d) Four phenotypes from T-Phenotype.

Figure G.1: Comparison of Phenotypes Discovered on the ADNI Dataset.



(a) Patients with stable ($\mathrm{std} < 1$) GCS trajectories.

(b) Patients with less stable ($\mathrm{std} \geq 1$) GCS trajectories.

Figure G.2: Three Phenotypes Discovered by T-Phenotype on ICU Dataset. The GCS trajectory of patients with different phenotypes are illustrated in the considered time period. All trajectories start at $t = 0$ and are smoothed with a rolling window of size 5.