PCPs: Patient Cardiac Prototypes Enable Interpretable Medical Diagnosis, Dataset Distillation, and Patient Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Clinical deep learning systems often generate medical diagnoses which are population-based and difficult to interpret. This is in contrast to how primary care physicians make decisions for the unique patient under consideration. Inspired by the workflow of such physicians, we develop a framework for learning embeddings, referred to as patient cardiac prototypes (PCPs), capable of capturing information that is unique to an individual patient's electrocardiogram (ECG) data. Through rigorous evaluation on three publicly-available ECG datasets, we show that PCPs allow deep learning systems to generate more interpretable diagnoses. We also demonstrate that PCPs can be used in lieu of the full dataset, which is orders of magnitude larger, to achieve comparable diagnostic performance. We find that PCPs can also be exploited to retrieve similar and dissimilar patient data across clinical databases. Our framework contributes to the development of transparent and patient-specific clinical deep learning systems.

1 Introduction

Primary care physicians often make decisions for the unique patient under their care (Hamburg & Collins, 2010). In doing so, they attempt to account for the complex mosaic of a patient's demographics, physiological state, and treatment trajectory. Such a patient-specific approach is in contrast to, for example, randomized control trials, long considered the gold standard of evidence in medical research (Cartwright, 2007), where findings are population-based and may overlook patient-specific nuances (Akobeng, 2005). These characteristics are all too common in deep learning systems which, despite their success in automating the diagnosis of medical conditions such as cardiovascular diseases (Galloway et al., 2019; Attia et al., 2019a;b; Ko et al., 2020), continue to generate population-based predictions that are considered opaque and uninterpretable by physicians.

Inspired by the workflow of primary care physicians, we believe that a framework which incorporates the information of an individual patient's data into a clinical deep learning system can confer several benefits. First, such a framework can allow for more interpretable medical diagnoses. For example, a particular misdiagnosis can now be traced back to the specific patient information involved in the diagnosis. One such diagnosis is that of identifying abnormalities in the functioning of the heart, also known as cardiac arrhythmia diagnosis, based on cardiac time-series data. Second, a framework which captures information unique to an individual patient's data can allow for patient data retrieval, whereby a clinical database is searched in order to retrieve data that are most similar to (or dissimilar from) the data of an existing patient. Such retrieval capabilities can, for example, allow physicians to compare the diagnoses, prognoses, and treatment outcomes of patients, complementing their decision-making. They also enable medical educators to retrieve similar and dissimilar patients from a database, leveraging up-to-date real-world evidence as a means of teaching the next generation of medical students.

Previous studies have attempted to develop frameworks that capture information that is unique to an individual patient's data, primarily through the concept of contrastive learning (Cheng et al., 2020). Although

previous work has focused on learning patient-specific representations of cardiac signals (Kiyasseh et al., 2021b), it was limited exclusively to pre-training a network and did not allow for interpretable diagnoses nor patient retrieval. While researchers have learned representations of cardiac signals that are specific to a set of patient attributes such as sex, age, and disease class (Kiyasseh et al., 2021a), their work remains population-based and also lacks an element of interpretability. Others have attempted to quantify the similarity of patients (Sharafoddini et al., 2017) via patient similarity networks (PSNs) (Pai & Bader, 2018; Pai et al., 2019). That approach, however, does not allow for dataset distillation nor patient retrieval, as is done by our framework. The learning of prototypes has been introduced (Snell et al., 2017; Sung et al., 2018), although primarily in the context of few-shot learning, where deep learning systems are encouraged to learn a task quickly and with few data points.

The main contributions of this study are as follows.

- **Supervised contrastive learning framework** - we develop a framework that learns embeddings, which we refer to as patient cardiac prototypes (PCPs), that capture information unique to an individual patient's electrocardiogram (ECG) data. We incorporate these PCPs into a deep learning system which identifies cardiac arrhythmias (abnormalities in the functioning of the heart) based on single-lead or 12-lead ECG signals (Strouse et al., 1939; Carter, 1950).
- **Interpretable diagnoses** - through rigorous evaluation on three publicly-available ECG datasets, we demonstrate that PCPs can be exploited to better understand *why* a particular diagnosis is made by a deep learning system for an individual patient. This can contribute to the development of *transparent* clinical deep learning systems and instill trust in clinical stakeholders.
- **Dataset distillation** - we show that PCPs can be used in lieu of the full dataset, which is orders of magnitude larger, to achieve comparable diagnostic performance. This has the potential to reduce the amount of resources required to train deep learning systems.
- **Patient retrieval** - we demonstrate find that PCPs can reliably retrieve similar (and dissimilar) patient data across distinct clinical databases. This can facilitate the provision of education by medical educators to the next generation of medical students.

2 Methods

2.1 Motivation behind patient cardiac prototypes

Patient cardiac prototypes are embeddings that summarize an individual patient's ECG data. We first provide the motivation behind the design of such prototypes before outlining in detail how we set out to learn them (next section).

2.1.1 Leveraging spatial and temporal invariances

First, recent research has demonstrated the benefit of learning representations of cardiac signals that are invariant to spatial and temporal sources of variability while *pre-training* on large-scale cardiac datasets (Kiyasseh et al., 2021b). An invariance is an aspect of the input data which, when changed, does *not* alter the information exhibited by that data. Specifically, we showed that attracting representations of different ECG leads to one another or attracting representations of temporally-adjacent ECG signals to one another (e.g., on the order of seconds) can allow deep learning systems to perform cardiac arrhythmia classification with fewer training datapoints than would otherwise have required. Inspired by those findings, we decided, in this study, to leverage the same invariances to learn representations (PCPs) that are invariant to changes in an individual patient's ECG data over space and time. Precisely, we assume that data that belong to the *same* patient reflect the *same* type of information, otherwise known as intra-patient invariance.

We acknowledge that, in some cases, intra-patient invariance *can* be undesirable. An example is when an individual patient's cardiac signals are collected over a large time-span (e.g., on the order of years), during which their physiological state might have changed, and are thus likely to reflect distinct information. By collapsing this distinct information into a single embedding, as is done with PCPs, we would lose diagnostically

relevant insight. From a practical perspective, and in our experiments, we avoid this problematic scenario by exclusively considering data points that are on the order of seconds (e.g., 10 seconds) and which are collected during the same hospital visit (see Fig. 1 left). To achieve this, we used the patient ID meta information, and, where available, the date of the ECG recording. An intra-patient invariance is therefore likely to hold in light of the (a) short time-span over which the data recording took place and (b) low likelihood of an electrocardiogram signal exhibiting major morphological changes during this time.

2.1.2 Aiming for interpretable diagnoses

One of our goals was to design an interpretable cardiac arrhythmia classification system. Combining this goal with our inspiration from how primary care physicians tend to the individual patient at hand, we set out to learn embeddings that capture information unique to a patient's ECG data and which are explicitly relied on by the deep learning system for diagnosis (see algorithm development for this dependence).

2.2 Learning patient cardiac prototypes

2.2.1 Notation

Let us assume we have a dataset, $\mathcal{D} = \{\boldsymbol{x_i}, y_i\}_{i=1}^{N}$, comprising N instances, $\boldsymbol{x}$, and cardiac arrhythmia labels, y, corresponding to a total of Ω_{train} patients in the training set. Each patient is associated with $N/\Omega_{\text{train}} > 1$ instances. This could be due to the provision of multiple medical tests during the same hospital visit or several visits. We also define a learner, $f_\theta : \boldsymbol{x} \in \mathbb{R}^D \rightarrow \boldsymbol{h} \in \mathbb{R}^E$, parameterized by $\boldsymbol{\theta}$, that maps a D-dimensional instance, $\boldsymbol{x}$, to an E-dimensional representation, $\boldsymbol{h}$. We aim to learn a set of embeddings, each of which efficiently summarizes the cardiac state of a patient. To that end, we associate each patient in the training set with a unique, learnable embedding, $\boldsymbol{p} \in \mathbb{R}^E$, to form the set of embeddings, $P = \{\boldsymbol{p}_j\}_{j=1}^{\Omega_{train}}$. Hereafter, we refer to such embeddings as patient cardiac prototypes (PCPs). We next explain how to learn PCPs.

2.2.2 Contrastive learning

We exploit the contrastive learning framework which, in short, consists of a sequence of attractions and repulsions between representations of instances. The idea is to attract representations of instances of the same patient to the single PCP of that same patient, and to repel them from the PCPs of the remaining patients. Formally, we encourage the representation, $\boldsymbol{h}_i = f_\theta(\boldsymbol{x}_i)$, of an instance, $\boldsymbol{x}_i$, associated with the k-th patient to be similar to the k-th PCP, $\boldsymbol{p}_k$, and dissimilar from the remaining PCPs, $\boldsymbol{p}_j$, $j \neq k$ (see Fig. 1). To achieve this, we optimize the InfoNCE loss (equation 1). Intuitively, it penalizes the learner for placing less probability mass on the similarity, $s(\boldsymbol{h_i}, \boldsymbol{p}_k)$, of the representation and prototype pair that should be most similar (based on patient ID) than on the similarity of other pairs, of which there are $\Omega_{train} - 1$. We quantify the cosine similarity between such pairs with a temperature parameter, τ.

$$\mathcal{L}_{NCE} = -\sum_{i=1}^{B} \log \left[\frac{e^{s(\boldsymbol{h}_i, \boldsymbol{p}_k)}}{\sum_{j}^{\Omega_{train}} e^{s(\boldsymbol{h}_i, \boldsymbol{p}_j)}} \right] \tag{1}$$

$$s(\boldsymbol{h}_i, \boldsymbol{p}_j) = \frac{\boldsymbol{h}_i \cdot \boldsymbol{p}_j}{\|\boldsymbol{h}_i\| \|\boldsymbol{p}_j\|} \cdot \frac{1}{\tau}$$

As a result of this many-to-one mapping from representations to PCP, the latter will become invariant to *intra-patient* differences present in the data. For context, it is these prototypes, P, which are presented in Fig. 2. This outcome is desirable only if we assume that such intra-patient differences point to the same underlying physiological state of the patient. One way to satisfy this assumption is by, for example, exclusively considering patient data that spans a short time-frame (e.g., seconds). Please refer to the previous section for a discussion on this assumption.

2.3 Incorporating patient cardiac prototypes into diagnosis pipeline

Equipped with PCPs, we direct our attention to the question of *how do we exploit PCPs to generate diagnoses?* Before addressing this question, we note that neural network parameters are often deterministic; they are held constant during inference. Therefore, when making a prediction (e.g., medical diagnosis), a network depends almost exclusively on the instance. Although an instance is likely to reflect patient information, a network does *not* explicitly exploit such information. In light of this, we design a framework in which a subset of network parameters are explicitly conditioned on patient information. To achieve this, we exploited hypernetworks (Ha et al., 2016) alongside our PCPs, as outlined next.

2.3.1 Making a diagnosis with hypernetworks

A hypernetwork, a neural network in and of itself, generates parameters for another neural network. Formally, a hypernetwork is a function, $g_\phi : \boldsymbol{h} \in \mathbb{R}^E \rightarrow \boldsymbol{\omega} \in \mathbb{R}^{E \times C}$, parameterized by $\boldsymbol{\phi}$, that maps an E-dimensional representation, $\boldsymbol{h}$, to a matrix of parameters, $\boldsymbol{\omega}$, where C is the number of class labels (i.e., cardiac arrhythmia categories). The output parameters, $\boldsymbol{\omega}$, can now be used to parameterize a linear classification head, $p_\omega : \boldsymbol{h} \in \mathbb{R}^E \rightarrow \boldsymbol{y} \in \mathbb{R}^C$, which maps an E-dimensional representation, $\boldsymbol{h}$, to an output probability distribution, $\boldsymbol{y}$. To explicitly condition the parameters, $\boldsymbol{\omega}$, on patient information as we had initially desired, we exploit PCPs differently during the training and inference stages of the framework, as outlined next.

2.3.2 Retrieving PCPs during training

During training, each representation, $\boldsymbol{h_i} = f_\theta(\boldsymbol{x_i})$, of an instance, $\boldsymbol{x_i}$, serves multiple purposes (see Fig. 1 left). First, it is attracted to its corresponding PCP, $\boldsymbol{p_k}$, as outlined earlier. To do so, we optimize the InfoNCE loss. Second, it is used as an input to the hypernetwork to generate *instance-specific* parameters, $\boldsymbol{\omega_i} = g_\phi(\boldsymbol{h}_i)$. Third, the representation is input into the classification head, p_ω, as is usual with neural networks. Given the ground-truth disease class, c, of each instance in a mini-batch of size B, we can optimize the categorical cross-entropy loss ($\mathcal{L}_{CE}$). In summary, during the training stage, we learn the parameters of the feature extractor ($\boldsymbol{\theta}$), the hypernetwork ($\boldsymbol{\phi}$), and the PCPs ($\{\boldsymbol{p}_j\}_{j=1}^{\Omega_{train}}$), in an end-to-end manner by optimizing the combined loss ($\mathcal{L}_{combined}$).

$$\mathcal{L}_{CE} = -\sum_{i=1}^{B} \log p_{\omega_i}(\boldsymbol{y_i} = c|\boldsymbol{h_i}) \tag{2}$$

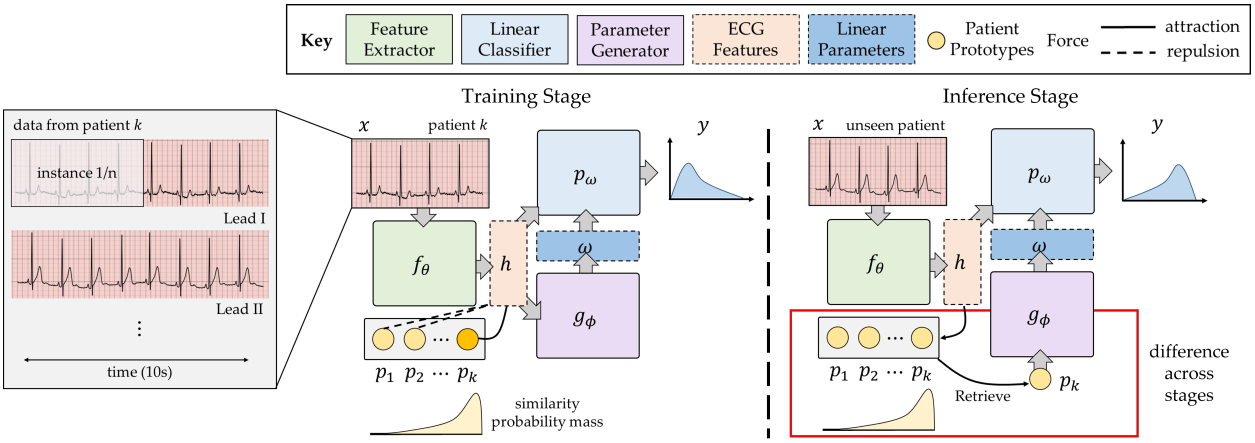

Figure 1: **Training and inference stages of cardiac arrhythmia diagnosis with patient cardiac prototypes. (Training Stage)** the representation, $\boldsymbol{h}$, of an instance, x, belonging to data from patient k is (a) fed into a hypernetwork, g_ϕ, to generate parameters, ω, for a linear classification layer, p_ω, (b) input directly into p_ω to output a class probability distribution, and (c) encouraged to be similar to the corresponding patient cardiac prototype (PCP), $\boldsymbol{p}_k$. **(Inference stage)** the representation of an instance associated with an unseen patient retrieves the nearest PCP, $\boldsymbol{p}_k$, which is then input into the hypernetwork. This generates linear parameters for classification.

$$\mathcal{L}_{combined} = \mathcal{L}_{CE} + \mathcal{L}_{NCE}$$

2.3.3 Retrieving PCPs during inference

During the inference stage, we propose several modifications to the pipeline (see Fig. 1 right). First, each representation, $\boldsymbol{h_i}$, is no longer attracted to or repelled from PCPs. This is primarily because the patients in the inference stage do *not* overlap with those in the training stage; they are mutually exclusive by design. Instead, each representation, $\boldsymbol{h_i}$, searches through the set of PCPs, P, and retrieves the single PCP to which it is closest, $\boldsymbol{p}_k = \text{argmax}_{\boldsymbol{p}_j} \; s(\boldsymbol{h_i}, \boldsymbol{p}_j)$, where s is some similarity metric such as cosine similarity. Second, the retrieved PCP, $\boldsymbol{p}_k$, (instead of $\boldsymbol{h_i}$, as was done during training) is now used as an input to the hypernetwork, g_ϕ. As a result, we generate parameters, $\boldsymbol{\omega_i} = g_\phi(\boldsymbol{p}_k)$, for each unseen instance in the held-out set of data. More precisely, we generate parameters conditioned on the single PCP associated with the patient in the training set that is deemed most similar to the patient observed during inference. In summary, the final diagnosis is underpinned by this retrieval and parameter-generation process.

2.4 Motivation behind retrieving PCPs during inference

We designed a framework in which, during inference, the representation of an unseen cardiac signal *retrieves* the PCP to which it is closest. We now outline the motivation behind this retrieval mechanism. Once a PCP is retrieved, it is fed into a hypernetwork to generate parameters (for a linear classification head) that directly influence the final diagnosis. Therefore, each diagnosis can now be traced back to the single (or handful of) PCP that was retrieved. This PCP (and its associated patient data) can be further inspected to help researchers better understand the diagnosis made by the network. For example, in the event of a correct diagnosis, inspection of the patient data associated with the retrieved PCP can act as a sanity check that the network is depending on clinically-relevant patients when making a diagnosis. On the other hand, and in the event of an incorrect diagnosis, inspection of the patient data associated with the retrieved PCP can lend insight into both *why* and *how* a network failed for this particular instance. We provide quantitative evidence in support of these claims in the Results section.

3 Experimental design

3.1 Electrocardiogram datasets

We conducted experiments on three datasets comprising electrocardiogram signals and cardiac arrhythmia labels. **CPSC** (Alday et al., 2020) consists of 12-lead ECG recordings from 6,877 patients alongside nine cardiac arrhythmia labels: AFIB, I-AVB, LBBB, Normal, PAC, PVC, RBBB, STD, and STE. **Chapman** (Zheng et al., 2020) consists of 12-lead ECG recordings from 10,646 patients alongside four high-level cardiac arrhythmia labels: AFIB, GSVT, Sinus Bradycardia, and Sinus Rhythm. **PTB-XL** (Wagner et al., 2020) consists of 12-lead ECG recordings from 18,885 patients alongside 71 different types of annotations provided by two cardiologists. We followed a previously-established training and evaluation protocol(Strodthoff et al., 2020) where we leveraged the 5 diagnostic class labels: Conduction Disturbance (CD), Hypertrophy (HYP), Myocardial Infarction (MI), Normal (NORM), and Ischemic ST-T Changes (STTC). We altered the original setup to only consider ECG segments with one label assigned to them and converted the task into a binary classification problem (NORM vs. Rest). Across all datasets, we split patients randomly into training, validation, and test sets, ensuring that there was no patient overlap between sets (see Appendix A).

3.2 Evaluation metrics

We evaluated our framework differently depending on the experimental scenario in which it was deployed. When evaluating the reliability of the diagnoses and the dataset distillation capabilities of our framework, we calculated the area under the receiver operating characteristic curve (AUC) of its predictions on instances in a held-out set of data. When evaluating the quality of the patient data retrieval system, we calculated the precision and negative predictive value (NPV) of the retrieved data points (those identified as relevant).

Precisely, we quantified whether a pair of patients identified as relevant by our framework were actually relevant, defined according to a match in their cardiac arrhythmia disease class. We qualitatively evaluated the relevance of these retrieved data points by comparing their corresponding patient information to one another.

3.3 Description of the ablation studies

To gain a better understanding of the marginal benefit of each of our framework's components on its overall performance, we conducted an extensive set of ablation studies, as outlined next.

3.3.1 Experimenting with variants of the retrieval mechanism

During inference on an unseen data point, our framework retrieves patient cardiac prototypes learned on exclusively on the *training* dataset (see Retrieving PCP during inference). We experimented with variants of this retrieval mechanism as part of an ablation study, which we outline next.

Retrieval variant 1 (mean) The first variant, which we refer to as *mean*, involved taking the average of all PCPs, $\bar{\boldsymbol{p}} = \Omega_{train}^{-1} \cdot \sum_{j=1}^{\Omega_{train}} \boldsymbol{p}_j$, regardless of the instance in the held-out set. Since PCPs and the hypernetwork are deterministic during the inference stage, this approach implies that the generated linear parameters are effectively reduced to a constant.

Retrieval variant 2 (similarity-weighted mean) The second variant, which we refer to as *similarity-weighted mean*, involved calculating a linear combination of the PCPs, weighted according to their similarity to the representation, $\boldsymbol{h}_i$, of an instance. Formally, $\bar{\boldsymbol{p}} = \sum_{j=1}^{\Omega_{train}} s(\boldsymbol{p}_j, \boldsymbol{h}_i) \cdot \boldsymbol{p}_j$. In effect, this approach exploited information unique to a patient's ECG data to down-weight, or up-weight, the contribution of each PCP.

Retrieval variant 3 (nearest) The third variant, which we refer to as *nearest*, involved the vanilla approach of retrieving the *single* PCP closest to the representation, $\boldsymbol{h}_i$. In effect, this approach retrieved the PCP of the patient in the training set that was deemed most similar to the representation of an instance associated with a different patient in the held-out set.

Retrieval variant 4 (nearest 10) We hypothesized that by restricting the framework to only select a single PCP, we were preventing the representation from exploiting potentially useful information contained in additional PCPs. For example, these additional PCPs could reflect patients with attributes (e.g., sex, age, and treatment outcome) that are shared with, and thus potentially useful for, the patient in the held-out set for whom the prediction is being made. Therefore, our fourth variant of the retrieval mechanism, which we refer to as *nearest 10*, involved taking the average of the ten PCPs closest to the representation, $\boldsymbol{h}_i$.

3.3.2 Experimenting with variants of the patient cardiac prototypes

To gain some better intuition as to whether PCPs were indeed capturing information unique to patient ECG data, we learned embeddings that differed slightly from our patient cardiac prototypes.

Optimizing the supervised loss alone In this ablation study, we trained our deep learning framework *without* a contrastive loss term (equation 1) and, therefore, did not learn the corresponding PCPs. In this scenario, we treated the average representation from each patient (instead of the PCP) as the descriptor of an individual patient's ECG data. In the Results, we refer to these embeddings as Mean Representations.

Optimizing the contrastive loss alone In this ablation study, we exclusively optimized the InfoNCE loss. Here, supervision manifested solely in the form of patient IDs. Importantly, cardiac arrhythmia disease labels are *not* seen by the network during training. After learning prototypes in this setting, we exploited them to perform cardiac arrhythmia classification. In contrast to previous experimental settings, this setting does not include a linear classification head that would have output a probability distribution over the possible

classes. Therefore, during inference, we decided to implement the k-nearest neighbours (knn) algorithm as a way to mimic the previously-introduced retrieval mechanism. Specifically, for each representation in the held-out set, we searched for its K nearest prototypes (using cosine similarity) and considered the most frequent cardiac arrhythmia ground-truth label of these prototypes as the diagnosis. To clarify, $K = 1$ and $K = 10$ can be thought of as being analogous to the nearest and nearest 10 retrieval mechanisms introduced earlier.

3.4 Description of the dataset distillation experiments

To conduct the dataset distillation experiments, we first trained our deep learning framework, as per usual, to learn the patient cardiac prototypes. We then fit a machine learning model (e.g., support vector machine, random forest) to these prototypes and evaluated its performance using representations of instances in the held-out set of data. We opted for these two models due to their simplicity in construction and implementation. The motivation was that if PCPs could perform well with such relatively simple models, then they are also likely to perform well with more complex non-linear neural networks.

3.5 Description of the patient data retrieval process

Patient cardiac prototypes have the potential to retrieve similar patient data within clinical databases. To substantiate this claim, we followed these steps.

Step 1 - calculate distance between patients When we searched for cardiac signals *within a dataset*, we first calculated the Euclidean distance, $d(\boldsymbol{p}_j, \boldsymbol{h}_i)$ between the j-th PCP, $\boldsymbol{p}_j$, and the representation, $\boldsymbol{h}_i$, of the i-th instance unseen during training (e.g., those in the validation set). Our motivation for choosing the Euclidean distance, as opposed to some other similarity metric, such as cosine similarity, will be discussed later. At this point, we had access to distances between a particular patient (in the form of a PCP) and representations. However, to obtain distances between patients and other patients, we averaged these distance values across representations, $\boldsymbol{h}_i$, of instances associated with the same patient. This process was then repeated for all PCPs. In contrast, when we searched for cardiac signals *across distinct datasets*, we simply calculated the Euclidean distance between the PCPs of these respective datasets. This immediately provided us with patient-patient distance values.

Step 2 - define relevance of retrieved patients Evaluating the relevance of the retrieved patient data is non-trivial. This is because the similarity (and dissimilarity) of patient data from a clinical perspective is nebulous. For example, patient data can be deemed similar based on attributes such as sex and age, medical history (e.g., cancer survivor), and drug treatment pathways. Unfortunately, publicly-available datasets of cardiac signals do not contain such exhaustive information. However, these datasets do entail cardiac arrhythmia disease labels. Therefore, each pair of patients was assigned a ground-truth relevance score ($s = 1$ relevant, $s = 0$ irrelevant) according to whether they shared the same cardiac arrhythmia label. We then identified pairs of patients as being relevant if their Euclidean distance, $d < d_E$, was less than some threshold distance, d_E. If, however, we were interested in retrieving dissimilar patients, then we would simply need to redefine the ground-truth relevance score and identify pairs of patients with $d > d_E$ as being relevant (notice the swap in the sign).

For the precision metric, we were looking to quantify how many of the retrieved patients (those identified as relevant) were actually relevant. In contrast, for the NPV, we were looking to quantify how many of the retrieved patients (those identified as irrelevant) were actually irrelevant.

4 Experimental results

4.1 Patient cardiac prototypes are distinct and can distinguish between cardiac arrhythmias

We began by visualizing the learned patient cardiac prototypes. Doing so would provide insight into the kind of information being captured by PCPs. For example, because of the way we learned PCPs (see Methods),

we hypothesized that they will reflect a patient's cardiac arrhythmia. To test this hypothesis, we present their two-dimensional UMAP projection(McInnes et al., 2018) (Fig. 2).

We found that patient cardiac prototypes are distinct from one another. For example, qualitatively, we see that PCPs have not collapsed to a select few points (Fig. 2). This is favourable for the following reason. During inference on an unseen cardiac signal, prototypes are explicitly input into a hypernetwork that generates parameters enabling a cardiac arrhythmia classification (see Fig. 1 right, and Methods). Therefore, if prototypes were to collapse to a single point, a process which we refer to as mode-collapse, then the same parameters will be generated regardless of the patient data input, diminishing their expressiveness. Acknowledging the potential imperfections in UMAP visualizations, we also provide more quantitative evidence to support this claim in the following section. We also found that PCPs do indeed reflect cardiac arrhythmia information. This is evident by the visible separability of the PCPs projections based on the cardiac arrhythmia categories.

4.2 Patient cardiac prototypes capture information unique to patient's ECG data

We explored the extent to which patient cardiac prototypes capture information unique to an individual patient's data. To do so quantitatively, we calculated the distance (Euclidean) between PCPs and representations of cardiac signals, belonging to either the same patient (*PCP to Same Training Patient*) or different patient (*PCP to Different Training Patient*). We present the distribution of these distance values for the Chapman dataset (Fig. 3).

We found that PCPs do indeed capture information that is unique to a patient's ECG data. This is evident by the smaller distance values of the PCP to Same Training Patient group than the PCP to Different Training Patient group (Fig. 3). For example, these two distributions have an average distance of 4 and 9, respectively. Expressed differently, this finding suggests that PCPs are twice as similar to representations of cardiac signals from the same patient than to representations of cardiac signals from a different patient.

When making diagnoses for unseen cardiac signals (i.e., inference), we implemented a retrieval mechanism whereby representations of *unseen* cardiac signals are used to retrieve the single PCP to which they were closest (see Methods). For this retrieval mechanism to work, the distance between unseen representations and PCPs (*PCP to Validation Patient*) have to be reasonable and on the same order of magnitude (Fig. 3). We found that these distance values are indeed reasonable. This is supported by the observation that they are on the same order of magnitude as the distance values of the PCP to Different Training Patients group. It is also worthwhile to note that the minimal overlap between the distributions PCP to Same Training Patient and PCP to Validation Patients reaffirms that the patients in the training and validation sets of the dataset do *not* overlap, which we had enforced by design for the purpose of rigorous evaluation.

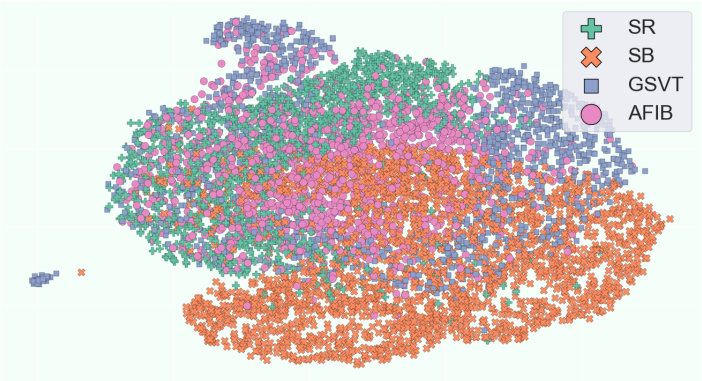

Figure 2: **Patient cardiac prototypes are distinct and can distinguish between cardiac arrhythmias.** The colours reflect four cardiac arrhythmia labels, Sinus Rhythm (SR), Sinus Bradycardia (SB), GSVT, and Atrial Fibrillation (AFIB).

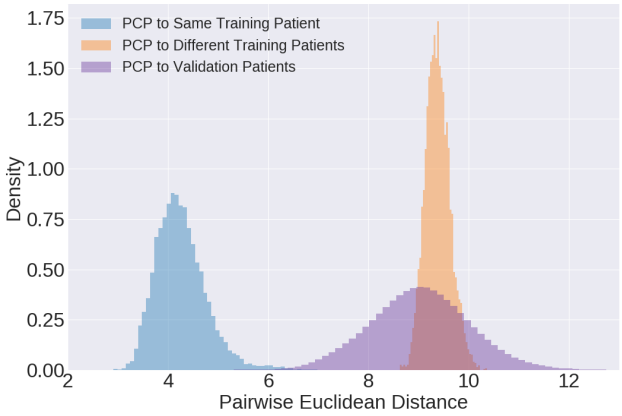

Figure 3: **Patient cardiac prototypes capture information unique to individual patient's ECG data.** We calculate the distance between PCPs and representations of instances in the training set associated with the same patient annotation, representations in the training set associated with a different patient annotation, and representations in the validation set.

4.3 Ablation studies

4.3.1 Effect of retrieval mechanism

When making a diagnosis for an unseen cardiac signal during inference, we retrieved the *single* PCP that was closest to its representation. This retrieval mechanism is a critical component of our framework (see Methods, Fig. 1). To appreciate this, note that the retrieval of an inappropriate PCP would have ramifications on the linear parameters that are generated by the hypernetwork, and, in turn, the cardiac arrhythmia diagnosis. We therefore explored four variants of this retrieval mechanism which differ in the extent to which they leverage information unique to a patient's ECG data (see Methods for details on variants).

We present the performance of these variants when retrieving PCPs learned exclusively on the training set of data (Fig. 4, top). We found that incorporating imprecise information about an individual patient's ECG data hindered the generalization performance of the deep learning system. For example, the *mean* variant, in which all PCPs are simply averaged before retrieval, achieved AUC ≈ 0.65 irrespective of the embedding dimension, which is significantly lower than that achieved by other variants of the retrieval mechanism. However, incorporating some patient information resulted in a significant improvement in performance. For example, the *similarity-weighted mean* variant, in which we retrieve a weighted linear combination of PCPs, achieved AUC ≈ 0.75 compared to 0.66 for *mean* at $E = 64$. This suggests that the PCP-derived similarity coefficients in the latter approach were beneficial, thus lending support to the utility of PCPs in the diagnosis pipeline.

We continued to test the limits of the retrieval mechanism and limited it to retrieving a *single* PCP. We found that individual PCPs were relevant for diagnosis during inference. This can be seen by the strong generalization performance achieved by the *nearest* variant. For example, at $E = 64$, *similarity-weighted mean* and *nearest* achieved AUC ≈ 0.75 and 0.89, respectively. We also found that incorporating additional information from a subset of patients, as reflected by *nearest 10* further improved performance, albeit in a more marginal way.

4.3.2 Effect of patient cardiac prototypes

The previously-introduced variants of the retrieval mechanism were designed to probe the degree to which information from an individual patient's ECG data (a) is captured by the prototype and (b) assists in the diagnostic performance of the system. We also explored whether it was necessary to incorporate prototypes into the deep learning system. To do so, we trained a variant of our framework without the prototypes (and without the contrastive loss, see Methods). As such, an individual patient's prototype is now replaced by the

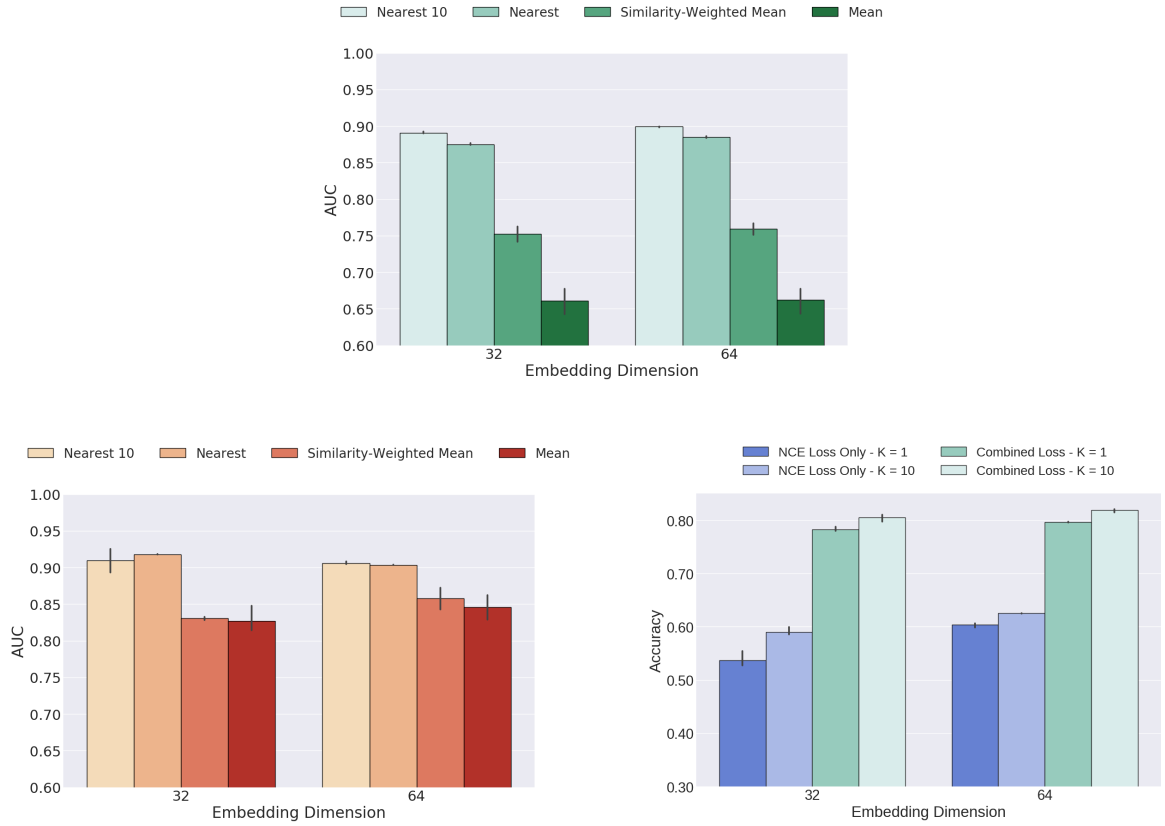

Figure 4: **Ablation studies examining the marginal impact of the components of our framework on performance.** Effect of the retrieval mechanism variants on the performance of the deep learning systems when using **(top)** PCPs and **(bottom left)** embeddings learned without a contrastive loss, also referred to as *mean representations*. (**bottom right**) Comparison of the performance of the K-nearest neighbours model with embeddings learned by exclusively optimizing the InfoNCE loss (NCE Loss Only) against that of our framework (Combined Loss). We show that such embeddings are less predictive of cardiac arrhythmia classes than the prototypes learned via our proposed framework.

average representation of that patient (*mean representation*). We present the results of these experiments while using the same retrieval mechanisms (Fig. 4, bottom left).

We found that mean representations capture information less unique to an individual patient's ECG data than PCPs. This is evident by the smaller difference in performance across the variants of the retrieval mechanism when using mean representations relative to PCPs (Fig. 4, bottom left). For example, the greatest difference in the AUC $\approx 0.92 - 0.83 = 0.09$ and AUC $\approx 0.89 - 0.66 = 0.23$, for the mean representations and the PCPs, respectively.

4.3.3 Effect of supervised contrastive learning

We also explored the extent to which supervision, in the form of cardiac arrhythmia labels, was necessary for the learning of prototypes with diagnostic value. To do so, we first trained a variant of our framework without such labels, and subsequently used the prototypes for cardiac arrhythmia classification (see Methods) with K-nearest neighbours (KNN). Here, K reflects the number of prototypes retrieved during inference. We compare the performance of this approach (NCE Loss Only) to when using patient cardiac prototypes (Combined Loss) (Fig. 4, bottom right). We found that prototypes learned without cardiac arrhythmia label supervision perform more poorly than those learned with supervision. This is evident by the lower accuracy

achieved by the former (NCE Loss Only - $K = 10$) relative to the latter (Combined Loss - $K = 10$) with accuracy ≈ 0.60 and 0.80, respectively.

4.4 Patient cardiac prototypes allow for interpretable diagnoses

We claimed that our framework and the associated patient cardiac prototypes allow for a high degree of interpretability in diagnoses made by deep learning systems. To provide evidence in support of this claim, we present an unseen patient's 12-lead ECG and clinical data (e.g., age, sex, etc.) alongside the data associated with the nearest PCP retrieved during inference (Fig. 5).

We found that when making a correct diagnosis for an unseen patient in the validation set, our framework depended on a relevant PCP (Fig. 5a). This is evident by the similar patient data associated with the retrieved PCP and the unseen cardiac signal in the validation set. For example, both exhibit a ground-truth cardiac arrhythmia label of sinus bradycardia, reflect patients of a similar age (60's), and outline similar cardiac-specific statistics (e.g., ventricular rate and atrial rate ≈ 55).

We also found that when making an incorrect diagnosis for an unseen patient in the validation set, our framework depended on an irrelevant PCP (Fig. 5b). This is evident by the dissimilar patient data associated with the retrieved PCP and the unseen cardiac signal in the validation set. For example, they exhibit different ground-truth cardiac arrhythmia labels (supraventricular tachycardia vs. atrial fibrillation). These patients also differ in their age (63 vs. 87), sex, and other cardiac-specific statistics (e.g., ventricular rate = 164 and 91). These discrepancies shed light on *why* the deep learning system failed to make the correct diagnosis. Inspecting such errors at the population level may also allow researchers to capture trends in the errors generated by a deep learning system.

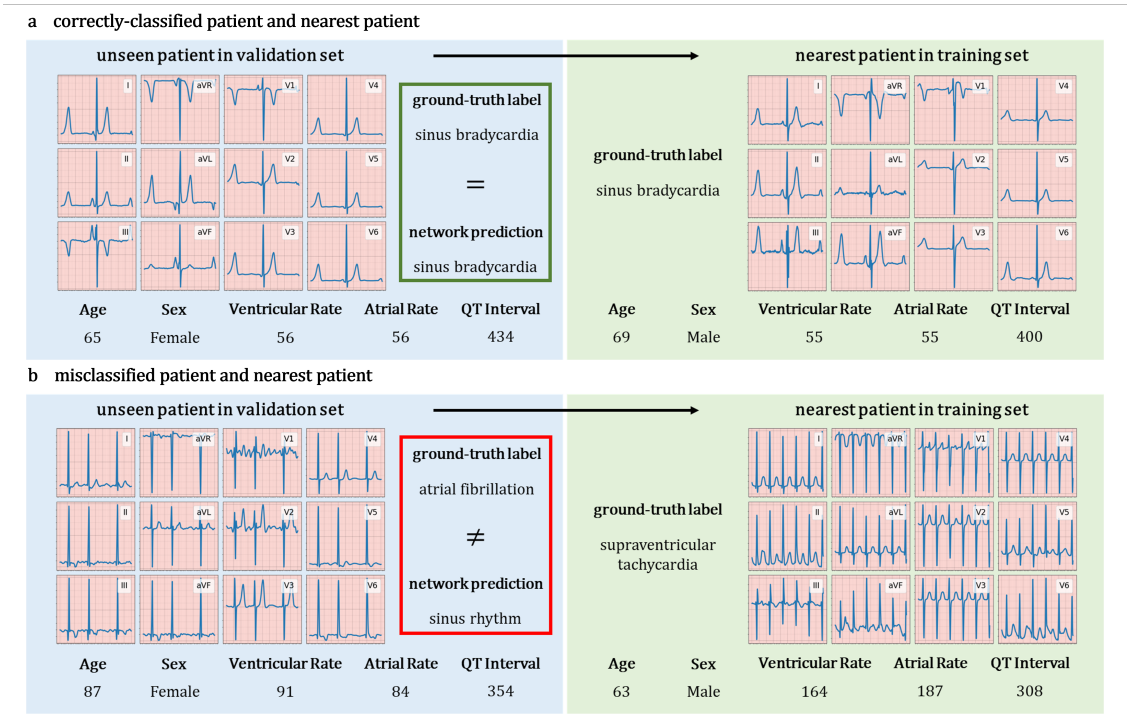

Figure 5: **Incorporating patient cardiac prototypes into the diagnosis pipeline allows for interpretable diagnoses.** We show the 12-lead ECG and patient data associated with a **(a)** correctly-classified patient and **(b)** misclassified patient from the validation set and that associated with the nearest PCP retrieved during inference. This process of tracing predictions back to the PCP retrieved during inference allows for improved network interpretability, and thus allows researchers to determine *why* an individual patient's diagnosis was made.

4.5 Patient cardiac prototypes outperform state-of-the-art dataset distillation methods

There are fewer patient cardiac prototypes than actual cardiac signals (because multiple cardiac signals map to a single prototype). We therefore hypothesized that PCPs could be exploited as a compact subset of, and substitute for, the labelled training data which, when trained on by a model, results in comparable performance.

To test this hypothesis, we *trained* a model on PCPs and *evaluated* it on unseen representations of cardiac signals. We compared PCPs to several state-of-the-art methods which identify compact subsets of data (known as core-sets) which we refer to as **Lucic** (Lucic et al., 2016), **Lightweight** (Bachem et al., 2018), and **Archetypal** (Mair & Brefeld, 2019) (see Appendix B.2). We deployed these methods on either raw cardiac signals or their representations learned via our framework (Table 1).

We found that core-sets of raw instances, generated by baseline core-set construction methods, did not provide a sufficient training signal to allow machine learning models to achieve strong generalization performance. For example, on the Chapman dataset, Lucic, Lightweight, and Archetypal achieved AUC = 56.8, 56.6 and 54.8, respectively. We attribute this performance to the low class separability of the input features. However, the baseline methods continued to perform poorly even when provided with the opportunity to construct core-sets from representations learned via our framework. Recall that these representations are more separable along the disease class dimension (see Fig. 2 and the associated performance in Fig. 4). For example, on the Chapman dataset, Lucic, Lightweight, and Archetypal achieved AUC = 57.8, 58.9 and 58.1, respectively. We found that PCPs outperformed these state-of-the-art core-set construction methods, instead achieving an AUC = 88.7.

4.6 Patient cardiac prototypes are effective dataset distillers

To determine whether PCPs are effective dataset distillers, we compared the performance of a model trained on 100% of the PCPs ($F = 1$) to one trained on the entire training set. To explore the extent to which further distillation was possible, we randomly chose a fraction, $F \in [0.05, 0.1, 0.2, 0.5]$, of the PCPs and trained a machine learning model on that subset, while continuing to evaluate on the same held-out set of data (Fig. 6). We also depict the performance of these frameworks when trained on all instances in the larger, original dataset (horizontal, dashed lines), which is several folds larger than the number of PCPs.

We found that PCPs are indeed effective dataset distillers (Fig. 6). For example, when training on 100% of the PCPs ($\Omega_{train} = 6,387$), an SVM model achieved similar performance (AUC ≈ 0.89) to one trained on all cardiac signals in the training set ($N = 76,614$). Expressed differently, similar performance was achieved despite a *12-fold* reduction in the number of training instances provided to the model. Such a finding suggests

Core-set	Chapman	CPSC	PTB-XL
Raw Instances			
Lucic(Lucic et al., 2016)	56.8 (0.8)	50.1 (0.1)	-
Lightweight(Bachem et al., 2018)	56.6 (0.4)	50.1 (0.1)	-
Archetypal(Mair & Brefeld, 2019)	54.8 (0.3)	50.1 (0.1)	-
Representations			
Lucic(Lucic et al., 2016)	57.8 (17.5)	50.6 (1.2)	51.6 (4.5)
Lightweight(Bachem et al., 2018)	58.9 (16.8)	50.5 (1.2)	52.4 (3.6)
Archetypal(Mair & Brefeld, 2019)	58.1 (16.8)	50.5 (1.2)	51.0 (5.0)
PCPs	**88.7** (0.5)	**52.8** (0.1)	**63.5** (0.7)

Table 1: **Patient cardiac prototypes outperform state-of-the-art core-set construction methods.** For the Chapman and PTB-XL datasets, we train a support vector machine. For CPSC, we train a random forest since it comprises a multi-label classification task. For the baseline methods, the core-set size was chosen to equal total number of PCPs. Mean and standard deviation (SD) are shown across five seeds. Note that since the raw instances of PTB-XL are 12-lead ECG signals, they could not be used with an SVM.

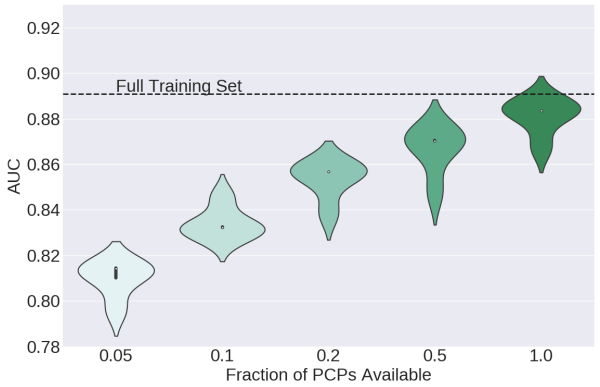

Figure 6: **Patient cardiac prototypes are effective dataset distillers.** Results are averaged across five random seeds. The horizontal dashed line depicts the performance of an SVM trained on all *instances* ($N = 76644$) in the training set. The PCPs are initially learned via a framework which optimizes the categorical cross-entropy loss and the noise contrastive estimation loss. We show that PCPs are effective dataset distillers; despite a 12-fold reduction in the number of training instances ($F = 1$, $\Omega_{train} = 6,387$), the SVM achieved similar performance (AUC ≈ 0.89) to one trained on all instances.

that the PCPs are able to capture the most pertinent information in the dataset and neglect that which is redundant for solving the task at hand.

We also observed that more extreme distillation does not significantly hinder the performance of PCPs. For example, an SVM model trained with only 5% of the available PCPs ($\Omega_{train} = 319$) achieved AUC ≈ 0.82. Expressed differently, this corresponds to a 7% drop in performance despite a *240-fold* reduction (relative to training on all instances) in the number of training instances provided to the model. Such a finding reaffirms the potential of PCPs as dataset distillers.

4.7 Patient cardiac prototypes reliably retrieve similar patient data

By virtue of capturing information that is unique to a patient's ECG data, patient cardiac prototypes have the potential to retrieve similar patient data in large clinical databases. Here, we provide both quantitative and qualitative evidence in support of that claim.

Quantitative evaluation We used PCPs to search for and retrieve similar patient data (see Methods), and calculated the corresponding precision and negative predictive value (NPV) as a function of the similarity between PCPs and unseen representations of cardiac signals (Fig. 7, left with Euclidean distance, Fig. 7, right with cosine similarity).

We found that patient cardiac prototypes reliably retrieved patients with a similar cardiac arrhythmia label. This is evident by the high precision achieved by PCPs at low Euclidean distance (high cosine similarity) values (Fig. 7). For example, $> 90\%$ of the pairs of patients that were deemed very similar to one another by our framework (i.e., $d_E < 6.2$) exhibited the same exact cardiac arrhythmia label. We also found that the precision decays as patients are deemed less similar to one another. For example, as $d_E \rightarrow 8.5$, the Precision $\rightarrow 0.3$. This finding, which is promising and expected from a reasonable similarity metric, is also exhibited by the precision curve in Fig. 7 (right). Note that, in this case, the precision curve is flipped along the y-axis since the most similar patients are those with the largest cosine similarity.

Based on these findings, we can identify a threshold distance between pairs of patients beyond which we are guaranteed a particular precision. This is useful for end-users looking for an empirical upper bound on the error. For example, in Fig. 7 (left), given a user-defined acceptable level of precision (e.g., 0.90), we identify the threshold distance (e.g., $d_E \approx 6.2$) below which patients have a high probability of being similar. This

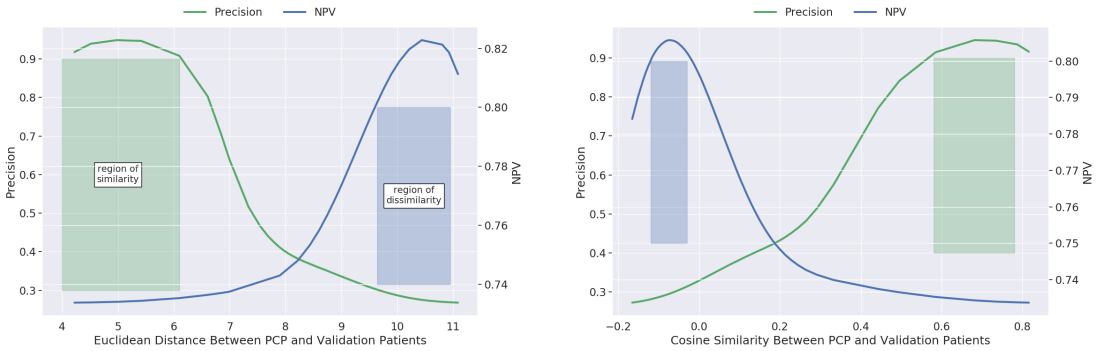

Figure 7: **Patient cardiac prototypes reliably retrieve similar patient data.** The similarity (or dissimilarity) metric used is the (**left**) Euclidean distance or **(right)** cosine similarity between PCPs and patient data in the validation set. The relevance of the retrieved patient data is based on their corresponding cardiac arrhythmia class. (**left**) For a given level of precision (e.g., 0.90), we can identify an appropriate threshold distance between patient data (e.g., $d < d_E = 6.2$) and thus delineate a region of similarity. Similarly, for a given level of NPV (e.g., 0.80), we can identify an appropriate threshold distance between patient data (e.g., $d > d_E = 9.7$) and thus delineate a region of dissimilarity. Our framework is agnostic to the type of similarity metric (e.g., Euclidean distance, cosine similarity) that is exploited for retrieval.

naturally lends itself to a *region of similarity*, where patients identified as being similar are highly likely of actually being so.

Qualitative evaluation To provide a qualitative understanding of the retrieval capabilities of our framework, we produced a matrix reflecting the distance (or similarity) between each pair of patients before retrieving the most similar pair of patients. We present their associated data in Fig. 10.

We found that PCPs were able to sufficiently distinguish between unseen patient data and thus act as reasonable patient data similarity tools. This is evident by the large range of distance values for any chosen PCP (any matrix row, Fig. 10). In other words, PCPs were closer to some representations than to others, implying that a chosen PCP was not trivially equidistant to all other representations. However, distinguishing between patient data is not sufficient for a patient data similarity tool. We found that PCPs can also correctly retrieve relevant patient data. We found that the two patients identified as being most similar to one another, using our method, do indeed share many similarities (Fig. 10, bottom). For example, their respective 12-lead ECG data are both associated with the cardiac arrhythmia label of sinus rhythm. Furthermore, similarities are observed when comparing the cardiac-specific statistics such as ventricular rate (84 in both cases) and atrial rate (84 in both cases). We hypothesize that this behaviour arises due to the ability of PCPs to efficiently summarize the cardiac state of a patient, a finding which reaffirms the potential of PCPs as tools for patient data retrieval. We observed similar findings on the remaining datasets (see Appendix C).

5 Discussion

Clinical deep learning systems are becoming increasingly adept at automating medical diagnoses. However, the vast majority of these systems make opaque decisions, rendering it difficult to understand why and how a diagnosis for an individual patient is made. This reduces the trustworthiness of such systems and the likelihood of their adoption by clinical stakeholders.

In this study, we took inspiration from how primary care physicians tend to the unique patient under consideration and proposed a deep learning framework that learns embeddings which capture the information unique to a patient's ECG data. We demonstrated that these embeddings, which we referred to as patient cardiac prototypes, or PCPs, can be leveraged for multiple downstream tasks. We demonstrated that PCPs,

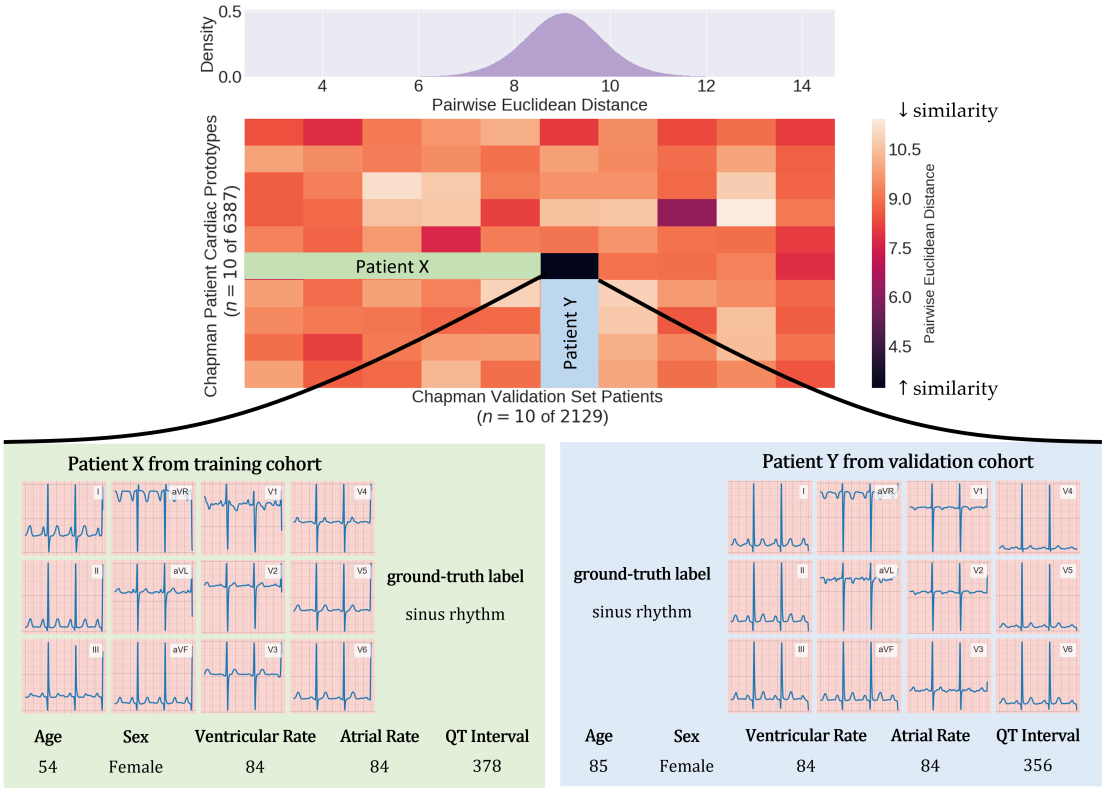

Figure 8: **Patient cardiac prototypes retrieve patients who have similar attributes.** We show a distribution of pairwise Euclidean distances between PCPs and representations in the validation set. We also present a subset of these pairwise distances in a matrix reflecting patient-patient distance values. We identify the most similar patient pair (↓ Euclidean distance) and retrieve their corresponding 12-lead electrocardiogram recordings, and where available, additional patient information.

when incorporated into the diagnosis pipeline, can allow for more interpretable medical diagnoses generated by deep learning systems. We also showed that PCPs are effective dataset distillers, where they can be used in lieu of the full datasets, which is orders of magnitude larger, to achieve comparable diagnostic performance. We also demonstrated that PCPs can reliably retrieve similar patients across clinical databases.

Our framework has the potential to improve the transparency of clinical deep learning systems, providing further insight into why and how a medical diagnosis was made for an individual patient. This transparency is likely to instil clinical stakeholders with trust, and engage them further on the path to the deployment of such systems within clinical ecosystems. Of note, learning prototypes and incorporating them into the diagnosis pipeline, as we have described, trivially extends to other disciplines of medicine in which a summary of data from a particular cohort is desired. Furthermore, as an effective approach to dataset distillation, our framework has the potential to reduce the resources and computational requirements for training clinical deep learning systems. Instead of training networks on full datasets, in their raw format, researchers can now distribute and learn from prototypes directly (which can be orders of magnitude smaller in size than full datasets). Additionally, as clinical data are generated at an ever-growing pace, prototypes allow researchers to search for and quickly retrieve patient data which are relevant for their use-case (e.g., clinical trial enrolment). For example, recent work demonstrated that prototypes can be learned with additional patient attributes (e.g., sex and age) as a means of retrieving and annotating previously-unlabelled cardiac signals (Kiyasseh et al., 2021a).

Our study does suffer from several limitations. Despite the multi-purpose use of patient cardiac prototypes, they remain relatively myopic; they do not account for diverse spatial and long-range temporal changes in patient data. When learning PCPs, we implicitly assumed that all representations of cardiac signals belonging to the same patient should be attracted to a single PCP. This is a valid assumption if such representations do indeed reflect a similar underlying physiological state. However, this assumption may not hold if patient data span multiple years or pertain to distinct modalities (e.g., imaging, electronic health records, etc.). In such a setting, patient cardiac prototypes would be a spatial and temporal average of a patient's health state. Although this summary embedding could be of use in certain scenarios, it would conceal subtle but useful changes in the patient's physiological state.

Moving forward, we turn our attention to multi-modal prototypes; those which are learned from multiple data modalities. This would allow researchers to obtain a more holistic summary of the state, cardiac or otherwise, of the patient. For example, given coronary angiogram data, electrocardiograms, and clinical reports, would it be possible to learn prototypes that further propel the diagnostic performance of clinical deep learning algorithms? Addressing this question could simultaneously improve the interpretability of the framework and its utility as a patient data similarity tool.

We also aim to exploit the quantification of the similarity of patient data to advance graph neural networks. Such networks typically require the design of an adjacency matrix, one that quantifies the presence and weight of edges (relationships) between nodes (patients). However, designing this matrix is non-trivial, particularly when dealing with physiological data, and can become a computational burden if dense. By interpreting these nodes as patients, our PCP-derived similarity values can be used to *initialize* the weights of edges between nodes and potentially inform the sparsity of node connections. As such, graph neural networks can be thought of as being trained with a PCP-derived prior, one that could accelerate, and inject valuable information into, the learning process.

References

A K Akobeng. Understanding randomised controlled trials. *Archives of Disease in Childhood*, 90(8):840–844, 2005. ISSN 0003-9888.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

Zachi I Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M McKie, Dorothy J Ladewig, Gaurav Satam, Patricia A Pellikka, Maurice Enriquez-Sarano, Peter A Noseworthy, Thomas M Munger, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine*, 25(1):70–74, 2019a.

Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019b.

Olivier Bachem, Mario Lucic, and Andreas Krause. Scalable k -means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 1119–1127, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520.

J. Bailey Carter. Value of the electrocardiogram in clinical practice. *Journal of the American Medical Association*, 143(7):644–652, 06 1950. ISSN 0002-9955.

Nancy Cartwright. Are RCTs the gold standard? *BioSocieties*, 2(1):11–20, 2007.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Conner D Galloway, Alexander V Valys, Jacqueline B Shreibati, Daniel L Treiman, Frank L Petterson, Vivek P Gundotra, David E Albert, Zachi I Attia, Rickey E Carter, Samuel J Asirvatham, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiology*, 4(5):428–436, 2019.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Margaret A Hamburg and Francis S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

Dani Kiyasseh, Tingting Zhu, and David Clifton. CROCS: Clustering and retrieval of cardiac signals based on patient disease class, sex, and age. *Advances in Neural Information Processing Systems*, 34, 2021a.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021b.

Wei-Yin Ko, Konstantinos C Siontis, Zachi I Attia, Rickey E Carter, Suraj Kapa, Steve R Ommen, Steven J Demuth, Michael J Ackerman, Bernard J Gersh, Adelaide M Arruda-Olson, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *Journal of the American College of Cardiology*, 75(7):722–733, 2020.

Mario Lucic, Olivier Bachem, and Andreas Krause. Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. volume 51 of *Proceedings of Machine Learning Research*, pp. 1–9, Cadiz, Spain, 09–11 May 2016. PMLR. URL `http://proceedings.mlr.press/v51/lucic16.html`.

Sebastian Mair and Ulf Brefeld. Coresets for archetypal analysis. In *Advances in Neural Information Processing Systems*, pp. 7247–7255, 2019.

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

Shraddha Pai and Gary D Bader. Patient similarity networks for precision medicine. *Journal of Molecular Biology*, 430(18):2924–2938, 2018.

Shraddha Pai, Shirley Hui, Ruth Isserlin, Muhammad A Shah, Hussam Kaka, and Gary D Bader. netdx: Interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology*, 15(3), 2019.

Anis Sharafoddini, Joel A Dubin, and Joon Lee. Patient similarity in prediction models based on health data: a scoping review. *JMIR Medical Informatics*, 5(1):e7, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2020.

Solomon Strouse, Louis N. Katz, and Herbert F. Binswanger. The clinical value of the electrocardiogram: an analysis of 100 private cases. *Journal of the American Medical Association*, 113(7):576–579, 08 1939. ISSN 0002-9955.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset, 2020. URL `https://physionet.org/content/ptb-xl/1.0.1/`.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):1–8, 2020.

Appendix

A Datasets

A.1 Data pre-processing

For all of the datasets, frames consisted of 2500 samples and consecutive frames had no overlap with one another. Data splits were always performed at the patient-level.

CPSC (Alday et al., 2020). Each ECG recording varied in duration from 6 seconds to 60 seconds with a sampling rate of 500Hz. Each ECG frame in our setup consisted of 2500 samples (5 seconds). We assign multiple labels to each ECG recording as provided by the original authors. These labels are: AF, I-AVB, LBBB, Normal, PAC, PVC, RBBB, STD, and STE. The ECG frames were normalized in amplitude between the values of 0 and 1.

Chapman (Zheng et al., 2020). Each ECG recording was originally 10 seconds with a sampling rate of 500Hz. We downsample the recording to 250Hz and therefore each ECG frame in our setup consisted of 2500 samples. We follow the labelling setup suggested by Zheng et al. (2020) which resulted in four classes: Atrial Fibrillation, GSVT, Sudden Bradychardia, Sinus Rhythm. The ECG frames were normalized in amplitude between the values of 0 and 1.

PTB-XL Wagner et al. (2020). Each ECG recording was originally 10 seconds with a sampling rate of 500Hz. We extract 5-second non-overlapping segments of each recording generating frames of length 2500 samples. We follow the diagnostic class labelling setup suggested by Wagner et al. (2020) which resulted in five classes: Conduction Disturbance (CD), Hypertrophy (HYP), Myocardial Infarction (MI), Normal (NORM), and Ischemic ST-T Changes (STTC). We alter the original setup in two main ways. Firstly, we only consider ECG segments with one label assigned to them. Secondly, we convert the task into a binary classification problem of NORM vs. (CD, HYP, MI, STTC) from above. The ECG frames were normalized in amplitude between the values of 0 and 1.

A.2 Data samples

In this section, we outline the number of instances used during training, validation, and testing for the CPSC, Chapman, and PTB-XL datasets.

Table 2: Number of instances (number of patients) used during training. These represent sample sizes for all 12 leads.

Dataset	Train	Validation	Test
CPSC	157,188 (4,402)	37,296 (1,100)	47,460 (1,375)
Chapman	76,614 (6,387)	25,524 (2,129)	25,558 (2,130)
PTB-XL	286,632 (10,807)	36,816 (1,411)	37,008 (1,383)

B Further implementation details

B.1 Network architecture

In this section, we outline the architecture of the neural network used for all experiments.

Table 3: Network architecture used for all experiments. K, C_{in}, and C_{out} represent the kernel size, number of input channels, and number of output channels, respectively. A stride of 3 was used for all convolutional layers. E represents the dimension of the final representation.

Layer Number	Layer Components	Kernel Dimension
1	Conv 1D BatchNorm ReLU MaxPool(2) Dropout(0.1)	7 x 1 x 4 (K x C_{in} x C_{out})
2	Conv 1D BatchNorm ReLU MaxPool(2) Dropout(0.1)	7 x 4 x 16
3	Conv 1D BatchNorm ReLU MaxPool(2) Dropout(0.1)	7 x 16 x 32
4	Linear ReLU	320 x E
5	Linear	E x C (classes)

Table 4: Batchsize and learning rates used for training with different datasets. The Adam optimizer was used for all experiments.

Dataset	Batchsize	Learning Rate
CPSC	256	10^{-4}
Chapman	256	10^{-4}
PTB-XL	256	10^{-3}

B.2 Baseline methods

In constructing a core-set, the baseline methods (Lucic (Lucic et al., 2016), Lightweight (Bachem et al., 2018), and Archetypal (Mair & Brefeld, 2019)) typically followed a similar strategy. These methods generated a categorical proposal distribution over all instances in the dataset before sampling k instances and assigning them weights. For a fair comparison to patient cardiac prototyps (PCPs), we chose $k = P$ where P is the number of PCPs.

C Additional results

In this section, we provide further evidence in support of the use of PCPs for patient retrieval. Specifically, we show how PCPs can be used to retrieve both similar and dissimilar patients *across* distinct datasets.

C.1 Effect of InfoNCE Loss on Dataset Distillation

In this section, we explore the effect of exclusively optimizing the InfoNCE loss on dataset distillation. Specifically, we train our network to optimize the InfoNCE loss (without the supervised cross-entropy loss) and learn prototypes in an end-to-end manner. To evaluate the dataset distillation capabilities of these learned prototypes, we exploit them to train a support vector machine (SVM) to classify cardiac arrhythmia classes. The intuition is that if these prototypes happen to be effective dataset distillers, then we would models trained exclusively on them to perform equally well (on a held-out dataset) as models trained on the entire dataset.

In Fig. 9, we present the accuracy of the SVM models trained on these prototypes as a function of the fraction of prototypes available for training. We show that prototypes learned in such a manner are *ineffective* dataset distillers; the performance of SVM models trained exclusively on such prototypes is worse than that achieved with a model trained on all representations. For example, at $F = 1$, these two models achieve Accuracy ≈ 0.40 and 0.60, respectively. We hypothesize that this poor performance (across all fractions) is due to the lack of class-discriminative behaviour exhibited by the learned prototypes. In other words, there exists a weak mapping from prototypes to cardiac arrhythmia classes. Such a finding suggests that, from the perspective of dataset distillation, our proposed combined loss (see Methods section of main manuscript) is preferable to the InfoNCE loss.

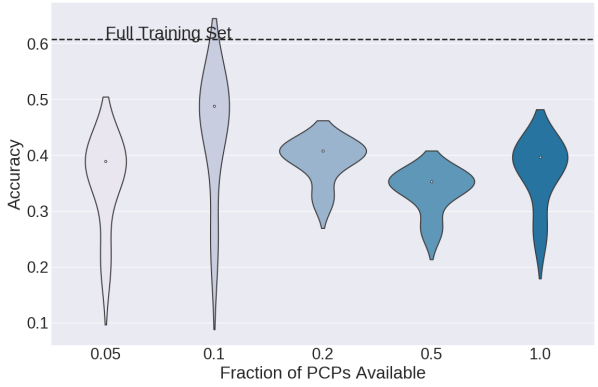

Figure 9: **AUC of SVM model trained on a fraction, F, of the prototypes and evaluated on a held-out set of data.** Results are averaged across five random seeds. The horizontal dashed line depicts the performance of a model trained on all of the training representations. The prototypes are initially learned via a framework which optimizes only the InfoNCE loss. We show that prototypes learned in such a manner are ineffective dataset distillers; the performance of SVM models trained exclusively on such prototypes is worse than that achieved with a model trained on all representations.

C.2 Retrieving similar patients across datasets

In Fig. 10, we illustrate the pair of patients (across distinct datasets) identified as being most similar to one another based on our PCP framework. We also present the 12-lead electrocardiogram data for this pair of patients in order to help validate our patient retrieval mechanism. We find that PCPs can reliably retrieve patients *across* distinct datasets that are similar to one another. This is evident by the similar morphology exhibited by the pair of 12-lead ECG data. For example, both patients exhibit normal electrical heart activity, which is also known as normal sinus rhythm.

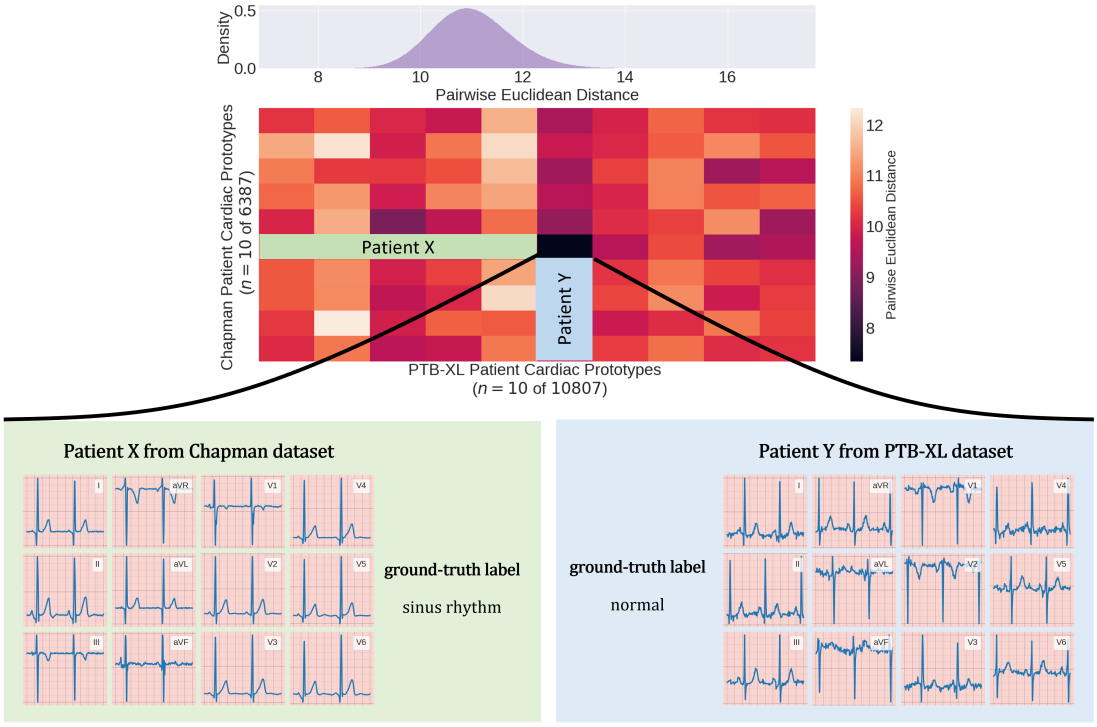

Figure 10: **Exploiting PCPs to discover similar patients across the Chapman and PTB-XL datasets.** We show a distribution of pairwise Euclidean distances between PCPs and representations in the validation set. We also present a subset of these pairwise distances in a matrix reflecting patient-patient distance values. We identify the most similar patient pair (↓ Euclidean distance) and retrieve their corresponding 12-lead electrocardiogram recordings. We show that PCPs can reliably identify similar patients. This is evident by the high degree of similarity exhibited by the pair of patients. For example, both patients exhibit a cardiac arrhythmia label of sinus rhythm.

C.3 Retrieving dissimilar patients across datasets

In addition to showing that PCPs can be used to retrieve similar patients, we claim that they can also be exploited for the retrieval of *dissimilar* patients. In this section, we provide qualitative evidence in support of this claim. In Fig. 11 (top), we present a pair of patients within the same dataset identified as being most dissimilar from one another. In Fig. 11 (bottom), we present a pair of patients *across* distinct datasets identified as being most dissimilar from one another. In both cases, we find that PCPs can reliably retrieve dissimilar patients. This is evident by the observation that the morphology of the pair of the 12-lead ECG data differs. For example, in Fig. 11 (top), the two dissimilar patients exhibit a normal rhythm (sinus rhythm) and a potentially fatal one (atrial fibrillation).

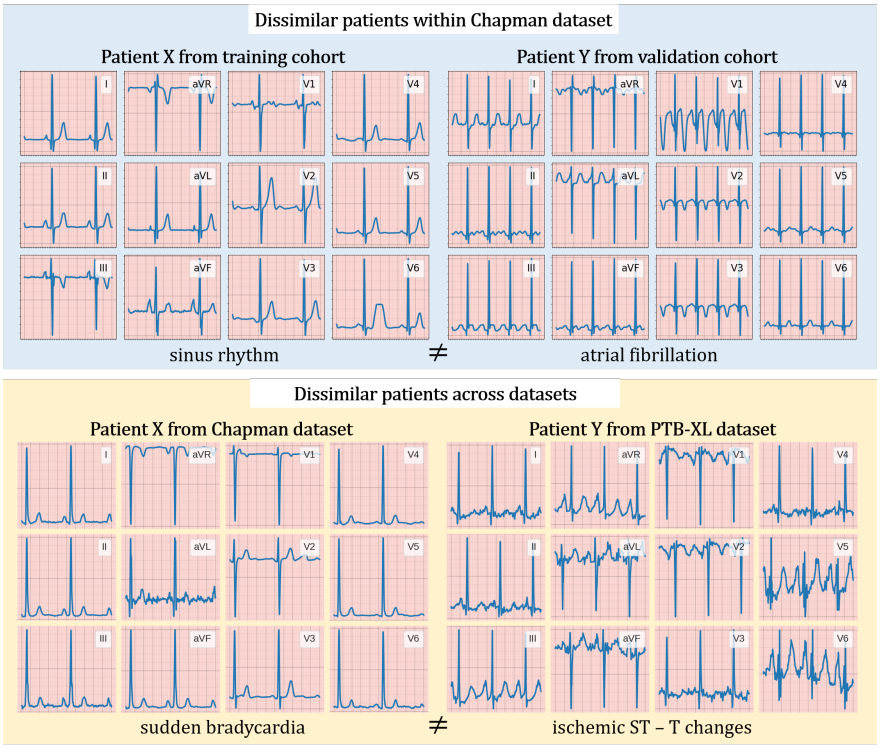

Figure 11: **12-Lead ECG segments corresponding to a pair of patients identified as being dissimilar from one another based on the PCPs.** Dissimilarity is defined as a high Euclidean distance between patient cardiac prototypes (PCPs) and representations of instances in the validation set. We show that PCPs can reliably retrieve dissimilar patients. This is evident by the observation that the ECG segments between patients exhibit different morphology and correspond to the different cardiac arrhythmia labels (sudden bradycardia vs. ischemic ST-T changes).