

Which representation clusters? A chance-corrected benchmark of ECG representations for unsupervised arrhythmia structure

Dongho Chun^a, Axel Faes^{b,*}

^a*Data Science Institute (DSI), Hasselt University, Hasselt, Belgium*

^b*Biomedical Data Sciences, Data Science Institute (DSI) and Biomedical Research Institute (BIOMED), Hasselt University, Hasselt, Belgium*

Abstract

Unsupervised analysis of the electrocardiogram (ECG) promises label-free discovery of cardiac structure, yet whether clustering succeeds depends less on the algorithm than on how the signal is represented. We benchmark four representation families for unsupervised separation of sinus rhythm (SR) from atrial fibrillation (AF) on PTB-XL: whole-recording raw flattening, a convolutional reconstruction autoencoder, compact hand-crafted rhythm and morphology features, and two frozen foundation-model embeddings (HuBERT-ECG and ECGFounder). Evaluation is chance-corrected (adjusted Rand index, ARI), patient-grouped, repeated over ten seeds for the clusterers that depend on one, and bounded by a supervised ceiling. ECG-Founder clusters best (ARI up to 0.885), but 16 interpretable hand-crafted features are competitive (0.711) and far ahead of the raw and autoencoder pipelines (ARI ≈ 0), which match a majority-class assignment despite supervised ceilings well above chance. The same ordering holds on a MIT-BIH beat-morphology task. Two cautions emerge: the silhouette score rewards the degenerate one-cluster solution (silhouette 0.81 at ARI 0.000) and cannot validate ECG clustering alone, and the feature scaler materially changes the outcome. The ranking is largely expected, since SR and AF differ by definition in rhythm regularity; the contribution is the chance-corrected controls that locate where naive pipelines fail and the finding that the cheap interpretable representation is the more deployable one. Under heavy class imbalance and federation across ten heterogeneous sites, the hand-crafted features

*Corresponding author.

Email address: `axel.faes@uhasselt.be` (Axel Faes)

are the most robust, overtaking the foundation embedding under imbalance while the embedding’s centralized advantage erodes under federation. The representation, not the algorithm, governs label-free ECG structure.

Keywords: electrocardiography, unsupervised clustering, representation learning, foundation models, adjusted Rand index, PTB-XL

1. Introduction

Most machine-learning work on the electrocardiogram (ECG) is supervised, but the clinically interesting questions are often unsupervised: which patients group together, what subtypes exist, and whether structure can be found without labels that are expensive or unavailable. A natural first attempt is to cluster recordings directly. Whether that attempt succeeds is usually attributed to the clustering algorithm, yet the choice that dominates is the representation fed to it.

This study began as a re-examination of a master’s project that clustered whole 12-lead recordings. That pipeline flattened each strip into a long vector, compressed it with a dense autoencoder, and clustered the code. It reported high cluster purity and a silhouette score near 0.89, but the purity simply equalled the majority-class proportion in every setting, and the silhouette was measured on a degenerate solution in which almost all points fell into a single cluster. The honest conclusion of that work, that the difficulty lies in processing the raw signal rather than in the clustering step, is correct and is the starting point here. We make it precise.

A representation can carry information about a label and still fail to cluster by it. Linear separability and clusterability are distinct properties: a representation may place the discriminative direction far from its dominant axes of variation, so that a supervised probe recovers the label while an unsupervised partition does not. This distinction is established in the self-supervised learning literature [1], and we treat it as background rather than as a finding. What has not been measured, for ECG, is how the common representation families compare on this axis under proper chance-corrected metrics.

The most directly related work benchmarks several ECG foundation models against one another, including a cluster-separability analysis on PTB-XL [2]. By its own statement it does not include a representation-wise comparison against hand-crafted features, and it omits the MIT-BIH database. Those two gaps are exactly where a careful, honest benchmark adds value,

because hand-crafted features remain the interpretable and inexpensive default in clinical signal processing, and MIT-BIH is the canonical morphology benchmark.

We therefore benchmark four representation families for unsupervised SR-versus-AF separation (sinus rhythm versus atrial fibrillation), with the controls the original pipeline lacked. The resulting ranking is not the contribution. Because SR and AF differ by definition in the regularity of the rhythm, that compact hand-crafted features and a large foundation model both recover the split is expected; what the benchmark adds is what the controls reveal and how the representations behave under deployment stress. Our contributions are modest and empirical. (i) A chance-corrected, patient-grouped, significance-tested clustering benchmark of raw, reconstruction-autoencoder, hand-crafted, and two frozen foundation-model representations, each bounded by a supervised ceiling and shown stable across the assumed number of clusters. (ii) Two reproducible methodological cautions that explain the original project’s misleading numbers: internal validity indices reward the degenerate collapse, and feature conditioning changes the result by a wide margin. (iii) The deployment finding, under two independent stressors: with heavy class imbalance and with federated non-independent client splits, the 16 hand-crafted features are the most robust, overtaking the foundation embedding under imbalance and holding far more of their clustering ARI under site heterogeneity. (iv) A generalization check on a MIT-BIH beat-morphology task, the dataset the closest prior benchmark omits. We do not propose a new method, we scope every foundation-model statement to the two models we test, and we credit the separability-versus-clusterability phenomenon to prior work.

2. Related work

Unsupervised ECG and representations.. Hand-crafted features built from RR-interval irregularity and beat morphology encode the defining signatures of common arrhythmias and underpin decades of automatic analysis [3, 4]. Tensor and wavelet decompositions have been combined with spectral clustering for unsupervised 12-lead classification [5]. Reconstruction autoencoders optimise signal energy rather than class structure, which makes them strong compressors but weak organisers of class geometry. Contrastive self-supervised methods built on physiologically motivated augmentations and patient- or lead-level invariances, including CLOCS [6], 3KG [7], and contrastive predictive coding for 12-lead ECG [8], established the pretraining

recipes that later foundation models scaled up. Self-supervised and foundation models for ECG, such as HuBERT-ECG [9] and ECGFounder [10], learn general representations from millions of recordings and are now evaluated as frozen feature extractors. The wavelet-tensor and spectral-clustering pipeline of [5] addresses a related unsupervised goal with a different (tensor) representation; we focus on the four families above, which span the energy-only to definition-encoding spectrum, and leave a tensor arm to future work.

Clustering evaluation. Cluster purity is not chance-corrected: assigning every point to the majority class attains a purity equal to that class’s prevalence. External, chance-corrected indices, the adjusted Rand index [11] and adjusted mutual information [12], are near zero for such trivial solutions and are the appropriate headline metrics. Internal indices such as the silhouette [13] measure compactness in the embedding being clustered and can be maximised by degenerate solutions, so they require cluster-size sanity checks.

Closest benchmark. The ECG foundation-model benchmark of [2] reports cluster-separability (ARI) of foundation-model embeddings on PTB-XL and finds that some models cluster by clinical label while others cluster by dataset. It compares foundation models to one another and states that prior work omits a representation-wise evaluation; it does not add a hand-crafted baseline, and it does not include MIT-BIH. We fill both gaps and add a supervised ceiling so that a low clustering score can be attributed to geometry rather than to missing information.

3. Methods

3.1. Data and cohorts

PTB-XL. We use PTB-XL [14], a public 12-lead database (21 799 records, 18 869 patients; 10 s strips). Rhythm labels are taken from the SCP-ECG (Standard Communications Protocol for Computer-Assisted Electrocardiography) statements. The SR-versus-AF cohort uses all records carrying an AF statement (AFIB or AFLT) as positives and sinus-rhythm records without any AF statement as controls. PTB-XL encodes rhythm statements at a likelihood of zero meaning “present”, so a presence-based rule is required; a confidence threshold would discard every SR record. We draw a balanced, patient-disjoint cohort of 2 542 records (1 282 SR, 1 260 AF), so the majority-class baseline is 0.504. Records are read from the 100 Hz distribution of the database with `wfdb.rdsamp` and cast once to a `float32` array of 1000 samples by 12 leads. No filtering, resampling or rescaling is applied at load

time, so every representation starts from the same array and declares its own preprocessing. The lead set is a per-representation choice rather than one setting for the whole study, and is stated with each representation below. All evaluation is patient-grouped: no patient appears in more than one fold or in both classes.

MIT-BIH. For generalization we use the MIT-BIH Arrhythmia Database [15, 16]. Each annotated beat is segmented into a window of $[-300, +250]$ ms around the R-peak across both leads. We form a balanced normal-versus-premature-ventricular (N-vs-V) task of 3 000 beats (1 500 per class) from 47 records, with the record as the grouping unit. This is a morphology task and a deliberate contrast to the PTB-XL rhythm task.

3.2. Representations

We compare four families on PTB-XL (Table 2). Each entry states the leads it consumes, the array it receives, and the principle it encodes.

- **Raw flattening** reproduces the original pipeline: each lead is moving-average downsampled and all leads are concatenated into one vector (~ 2000 dimensions). It carries amplitude and energy but no cross-recording temporal alignment. Input: all twelve leads, `float32`, 1000 by 12. A non-overlapping six-sample mean per lead leaves 166 samples per lead and a 1992-dimensional vector. The principle is that the sample amplitudes are themselves the features, with no clinical model imposed on the signal, and all twelve leads enter because this arm reproduces the originating pipeline, which concatenated every lead.
- **Convolutional autoencoder** (conv-AE): a 1-D convolutional autoencoder with a 32-unit bottleneck, trained to convergence with a validation split. This is a properly specified reconstruction model, in contrast to the original collapsed dense autoencoder. Input: all twelve leads, `float32`, resampled to 12 by 2500 and z-scored per lead. The principle is that a reconstruction objective compresses whatever carries signal energy, which need not be what separates the classes.
- **Hand-crafted** (16 features): RR-interval statistics and irregularity (the standard deviation of the RR intervals, SDNN; the root mean square of successive RR differences, RMSSD; the proportion of successive RR differences above 50 ms, pNN50; the coefficient of variation, RR entropy and mean heart rate), P-wave presence and amplitude,

QRS width and amplitude, and beat-template consistency [3, 4], extracted with NeuroKit2 [17]. Input: lead II alone, `float32`, 1000 samples, reduced to a 16-dimensional `float32` vector. R-peak detection and wave delineation run on that one lead. Lead II is the standard rhythm lead, the one on which RR regularity and P-wave presence are read, single-lead recordings are the established setting for automated AF detection [18], and a single-lead definition is what allows the same feature logic to be carried to the two-lead MIT-BIH beats, where a twelve-lead definition would not apply. The principle is that the AF signature is RR irregularity together with P-wave absence, so the representation measures those two quantities directly.

- **Foundation embeddings** (frozen): mean-pooled hidden states of HuBERT-ECG (768-d) [9] and the 1024-d pooled features of ECG-Founder [10], both used as fixed feature extractors with no fine-tuning. Input: all twelve leads, `float32`. HuBERT-ECG receives the first 5 s at 100 Hz, flattened lead-major into a 6000-sample vector; ECGFounder receives the full strip resampled to 12 by 5000 for its 500 Hz input and z-scored once over the whole array. Both lead sets are fixed by the published models, which are defined on the complete twelve-lead strip. The principle is that large-scale pretraining has already fixed an ECG feature space, and the question is whether that space places the rhythm classes in separate regions.

On MIT-BIH the rhythm features do not transfer to single beats, so we compare a raw beat window, a 21-dimensional beat-morphology and RR-context feature set, and the conv-AE. The 12-lead foundation models are not applied to 2-lead beats; this scoping limit is itself informative, because such models are tied to the 12-lead strip format.

The two lead sets are deliberate and answer to different constraints. The hand-crafted arm is single-lead by clinical convention and so that the same features transfer to the two-lead beats; the raw arm is twelve-lead because that is the pipeline it reproduces. The asymmetry gives the raw arm more input than the hand-crafted arm, not less, so the lead set cannot account for the raw arm’s chance-level clustering.

3.3. Clustering and evaluation

Each representation is clustered with k -means, Ward agglomerative, spectral, a Gaussian mixture model (GMM) and hierarchical density-based spatial clustering of applications with noise (HDBSCAN), with the number of

clusters set to the number of classes [19]. The implementations are those of scikit-learn, run with ten k -means restarts per seed, Ward linkage on Euclidean distance, a nearest-neighbour affinity with k -means label assignment for spectral clustering, diagonal covariance with a covariance floor of 10^{-4} for the Gaussian mixture, since a full covariance is singular for the high-dimensional representations, and a minimum cluster size of $n/50$ rounded down and floored at 10 for HDBSCAN [20], which is 50 records on the PTB-XL cohort. Tolerances and iteration caps are the library defaults, and no early stopping is imposed on the clusterers. Ward agglomerative and HDBSCAN are deterministic given the input matrix, so each is fitted once rather than once per seed. The conv-AE is trained with Adam at a learning rate of 10^{-3} , batch size 64, for at most 50 epochs on a 90/10 train/validation split, stopping when the validation reconstruction error has not improved for eight epochs and restoring the best-validation weights. The two foundation models are used in evaluation mode with gradients disabled and their published weights unchanged: HuBERT-ECG is mean-pooled over time from its last hidden layer at batch size 64, and ECGFounder returns its globally average-pooled 1024-dimensional feature layer at batch size 32. Conditioning is fitted on the full representation matrix, and any hand-crafted feature a record leaves undefined is filled with that feature’s cohort median before conditioning. The headline metrics are the chance-corrected adjusted Rand index (ARI) [11] and adjusted mutual information (AMI) [12]; normalized mutual information (NMI), V-measure, Fowlkes-Mallows, purity, silhouette [13], and the largest-cluster fraction are reported alongside. Every metric is computed against a majority and a random baseline. Table 1 gives each reported metric with its definition, its range, and how a value on it is read. There U is the partition by true class and V the partition by cluster, n is the number of records and n_{ij} the number of records in class i and cluster j , RI is the fraction of record pairs on which U and V agree, I and H are mutual information and entropy with $\bar{H}$ the arithmetic mean of $H(U)$ and $H(V)$, $E[\cdot]$ is the expectation under the permutation model that holds both sets of group sizes fixed, and a_k and b_k are the mean distance from record k to the rest of its own cluster and to the nearest other cluster. None of the eight carries a physical unit: the first five are dimensionless indices of agreement between the true partition and the recovered one, purity is the majority-class share inside the clusters, the silhouette is a dimensionless index of cluster compactness computed without labels, and the largest-cluster fraction is a proportion of the cohort. A supervised ceiling bounds how much of the label a probe recovers from the same features under the same patient-grouped five-fold protocol; we report both a linear probe (balanced logistic

regression) and a non-linear probe (random forest), so that a high ceiling with a low ARI identifies a geometric rather than an informational failure. Each configuration is repeated over ten seeds, and the headline contrasts are tested with a paired Wilcoxon signed-rank test across seeds. Because heavy-tailed features can dominate Euclidean distance, we sweep four conditioners (Yeo-Johnson power, standard, quantile, robust) and report power scaling as the default unless stated otherwise. To check that fixing the cluster count to the number of classes does not flatter the comparison, we also report adjusted mutual information, which is comparable across cluster counts, for K from 2 to 8. Because the natural SR-to-AF prevalence is imbalanced, we additionally evaluate on a 9:1 SR:AF subsample.

Table 1: Reported clustering metrics: definition, range and how a value is read. The notation is defined in the text.

Metric	Definition	Range	Reading
ARI	$\frac{\text{RI} - E[\text{RI}]}{\max \text{RI} - E[\text{RI}]}$	≤ 1	Chance-corrected pair agreement. 0 at chance, 1 for identical partitions.
AMI	$\frac{I - E[I]}{\bar{H} - E[I]}$	≤ 1	Chance-corrected shared information. 0 at chance, 1 for identical partitions.
NMI	$I/\bar{H}$	$[0, 1]$	Shared information without the chance correction; rises with the cluster count.
V-measure	$2hc/(h + c)$	$[0, 1]$	Harmonic mean of homogeneity h and completeness c ; equals NMI under arithmetic normalization.
Fowlkes-Mallows	$\sqrt{PR}$	$[0, 1]$	Geometric mean of pair precision P and pair recall R ; not chance-corrected.
purity	$\frac{1}{n} \sum_j \max_i n_{ij}$	$[0, 1]$	Majority-class share inside clusters. A single cluster already scores 0.504 here.
silhouette	$\frac{1}{n} \sum_k \frac{b_k - a_k}{\max(a_k, b_k)}$	$[-1, 1]$	Internal and label-free. 1 tight and separated, 0 overlapping, negative misassigned.
largest-cluster frac.	$\max_j v_j /n$	$[1/ V , 1]$	Degeneracy check. 1 means one cluster holds every record.

HDBSCAN labels records it does not assign to any dense region as noise, and the metrics treat that label in two ways. ARI, AMI, NMI, V-measure and Fowlkes-Mallows receive the prediction unchanged, so the noise records form

one further group of their own and are charged against the agreement with the true partition. Purity and the silhouette are computed on the assigned records only, because neither is defined for a set that is not a cluster, and the purity denominator stays the full cohort. Noise records are never dropped from the evaluation and never reassigned to a nearest cluster. This is why HDBSCAN sits at the bottom of every representation: at this minimum cluster size it assigns a minority of the cohort on the representations that cluster well, and on the raw, conv-AE and HuBERT-ECG matrices it labels the entire cohort as noise, which leaves one group and an ARI of exactly zero.

Uncertainty is quantified at two levels, which answer different questions. A percentile bootstrap over the ten per-seed values, 10 000 resamples, measures how far a result moves when only the random seed moves. These intervals are narrow: across the five PTB-XL representations the k -means seed interval spans less than 0.002 in ARI, and for two of them it has zero width because every seed returns the same partition. Ward agglomerative and HDBSCAN have no seed interval at all, since each is fitted once. A seed interval therefore carries no more information than the seed standard deviation it would replace. The interval reported with the results is the second one, a patient-grouped percentile bootstrap over the evaluation cohort with 5 000 resamples: the partitions are held fixed and the cohort is resampled with replacement at the patient level, so every record of a drawn patient travels with it. The statistic resampled is the quantity the tables report, the mean ARI over the ten seeds where the clusterer is seeded and the single fit where it is not, so each interval contains its own point estimate. All intervals are the 2.5th and 97.5th percentiles of the bootstrap distribution at a fixed bootstrap seed.

Each paired Wilcoxon signed-rank test is reported with two effect sizes. With ten paired seeds the two-sided p-value is pinned at its floor of 0.002 as soon as every seed moves the same way, so it separates none of the contrasts and cannot express a magnitude. Cliff’s delta gives the dominance of one arm over the other across all cross-seed pairs, on a scale from -1 to 1 , and the matched-pairs rank-biserial correlation rescales the signed-rank statistic itself onto the same scale. Each headline contrast additionally carries the paired difference in ARI with its own patient-grouped bootstrap interval.

3.4. Federated robustness

To probe deployment under data siloing we use standard federated k -means with FedAvg-style centroid aggregation [21]; we do not propose a new federated method. The 2 542 records are partitioned, patient-grouped,

across ten simulated sites, either independently and identically distributed (IID) or with class proportions drawn from a Dirichlet distribution so that sites differ sharply in arrhythmia prevalence (non-IID). The Dirichlet concentration is 0.3, applied per class over the ten sites, and any site left holding fewer records than clusters is dropped. Conditioning is fitted once on the full representation matrix before partitioning, so the federation concerns the clustering step alone. A run is fifteen aggregation rounds of three local Lloyd iterations per site, with the server replacing each centroid by the mean of the site centroids weighted by the number of records each site assigned to it, and the centroids initialised globally from two randomly drawn records. Both settings are repeated over ten seeds, with the site partition and the centroid initialisation redrawn at each seed. We report the global ARI relative to centralized clustering of the same representation.

4. Results

4.1. Representation, not algorithm, governs clusterability

Table 2 and Fig. 1 summarise the PTB-XL benchmark. ECGFounder clusters best (ARI up to 0.885 with k -means), the hand-crafted features are competitive (0.711 with agglomerative), HuBERT-ECG is weak (0.195), and the raw and conv-AE pipelines are indistinguishable from a majority-class assignment (ARI ≈ 0). The spread across clustering algorithms within a strong representation is smaller than the spread across representations, so the representation is the deciding factor.

Uncertainty enters at two levels and they are of different size. Moving only the seed leaves the k -means partitions almost unchanged, so the spread reported with the results is the patient-grouped cohort bootstrap given in the caption of Table 2, whose intervals separate every pair of representations except raw-flatten from conv-AE. A paired Wilcoxon signed-rank test on the k -means seeds confirms every adjacent contrast, ECGFounder over hand-crafted ($\Delta = 0.242$, paired bootstrap interval [0.202, 0.280]), hand-crafted over HuBERT-ECG ($\Delta = 0.470$, [0.424, 0.515]), and HuBERT-ECG over raw ($\Delta = 0.174$, [0.146, 0.203]), at $p = 0.002$, the floor for ten paired samples. That floor is reached by any contrast in which all ten seeds move the same way, so the p-value orders none of the three and each carries two effect sizes as well. Cliff’s delta and the matched-pairs rank-biserial correlation are both 1.00 on all three: every cross-seed pair and every signed rank favours the higher-ranked representation. The magnitudes are therefore read from the paired differences and their intervals above. The ordering does not depend on fixing the cluster count to the number of classes: with adjusted mutual

Table 2: PTB-XL SR-vs-AF: clustering ARI (power scaling) and the supervised ceiling (patient-grouped logistic-regression accuracy). The k -means, spectral and Gaussian-mixture columns are means over ten seeds; the agglomerative and HDBSCAN columns are single deterministic fits. Majority-class baseline 0.504; random ARI ≈ 0 . Patient-bootstrap intervals on the best clusterer of each representation: ECGFounder 0.885 [0.860, 0.909], hand-crafted 0.711 [0.677, 0.746], HuBERT-ECG 0.195 [0.166, 0.227], raw-flatten 0.013 [0.007, 0.019] and conv-AE 0.008 [0.002, 0.016]. Only the raw-flatten and conv-AE intervals overlap. Seed variation is the smaller of the two quantities: across the ten seeds the k -means solutions move by less than 0.002 in ARI and are identical for the hand-crafted and HuBERT-ECG matrices, the Gaussian mixture is the one seed-sensitive clusterer (seed standard deviation 0.013 on raw-flatten and 0.012 on ECGFounder), and the agglomerative and HDBSCAN columns have no seed interval. All adjacent k -means contrasts are significant at $p = 0.002$ (paired Wilcoxon, the ten-seed floor). The ordering holds for $K = 2$ to 8 under adjusted mutual information.

Representation	k -means	agglom.	spectral	GMM	HDBSCAN	sup. ceiling
raw-flatten (~ 2000 -d)	0.000	0.000	0.000	0.013	0.000	0.519
conv-AE (32-d)	0.004	0.008	0.001	0.005	0.000	0.678
hand-crafted (16-d)	0.644	0.711	0.672	0.544	0.060	0.942
HuBERT-ECG (768-d)	0.174	0.134	0.016	0.195	0.000	0.899
ECGFounder (1024-d)	0.885	0.709	0.860	0.851	0.137	0.983

information for K from 2 to 8, ECGFounder leads at every K , from 0.825 at $K = 2$ to a minimum of 0.441 at $K = 7$ and 0.450 at $K = 8$; the hand-crafted features fall monotonically from 0.557 to 0.329; HuBERT-ECG stays at or below 0.132 and the raw pipeline never exceeds 0.005.

The two foundation models do not behave as one category. ECGFounder clusters strongly while HuBERT-ECG does not, which matches the report of [2] that some ECG foundation embeddings organise by clinical label and others by dataset. We therefore name each model and do not generalise to foundation models as a class. HDBSCAN labels most points as noise and underperforms throughout, and spectral clustering collapses where the nearest-neighbour affinity graph is poorly connected; we report both rather than tune them away.

Each representation’s position follows from two things: how much of the label its space holds, which the supervised ceiling in the last column of Table 2 measures, and where in the space that label lies. The hand-crafted representation is 16-dimensional and each coordinate is a quantity by which the classes are defined, RR irregularity and P-wave presence; its ceiling is 0.942 and an unsupervised partition recovers 0.711. ECGFounder has the highest ceiling, 0.983, recovers the most, 0.885, and leads at every K from

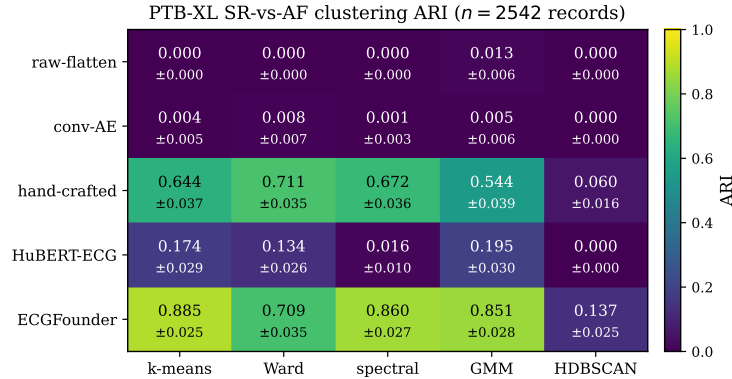


Figure 1: PTB-XL SR-vs-AF clustering ARI by representation and clusterer (power scaling). All 25 representation-by-clusterer cells on the cohort of 2542 records: ten seeds for k -means, spectral clustering and the Gaussian mixture, and one deterministic fit for Ward agglomerative and for HDBSCAN. Each cell gives the mean ARI over the half-width of its 95% patient-grouped bootstrap interval.

2 to 8, so the rhythm split is one of the larger sources of variation in its 1024 dimensions. HuBERT-ECG has almost the same ceiling as the hand-crafted features, 0.899, and recovers 0.195. The conv-AE ceiling of 0.678 is what a bottleneck optimised to reconstruct signal energy retains of a class it was never asked about, and it recovers 0.008. Raw flattening reaches 0.519 against a majority of 0.504, so a linear probe finds almost no label in the 1992-dimensional vector at all: its coordinates are amplitudes at fixed sample offsets, and no fixed offset falls on the same part of the cardiac cycle in two recordings. The raw and conv-AE rows both sit at a majority-class assignment under all five clusterers, and they do so for different reasons, one for want of label information in the space and one for want of any arrangement of it, so changing the clustering algorithm addresses neither. Which of the two accounts for HuBERT-ECG is what the next subsection measures.

4.2. Separability is not clusterability

The supervised ceilings (last column of Table 2; Fig. 2) expose the gap. The raw pipeline carries almost no label information (ceiling 0.519, barely above the 0.504 majority), consistent with its chance-level clustering. The conv-AE is more separable (0.678) yet still clusters at chance (best ARI 0.008, from Ward agglomerative): a properly trained reconstruction autoencoder encodes the class without organising it into clusters. HuBERT-ECG is the

sharpest case, with a ceiling of 0.899 but an ARI of only 0.195. ECGFounder and the hand-crafted features achieve both high separability and high clusterability. The discriminative information exists in the weaker embeddings; it simply does not lie along their dominant axes of variation, the established separability-versus-clusterability effect [1]. This is not an artefact of a linear probe: a random forest on the raw features reaches accuracy 0.714, well above its linear ceiling of 0.519 and far above its clustering ARI of 0.000, so even a non-linear decoder recovers a label that clustering does not. The same probe gives 0.951 on the hand-crafted features, 0.983 on ECGFounder and 0.825 on HuBERT-ECG. On ECGFounder the two probes agree to three decimals, the linear ceiling at 0.9831 and the random forest at 0.9835. For HuBERT-ECG the random forest sits below the linear ceiling of 0.899, so probe linearity does not account for that gap either.

That reading is testable and we test it out of fold. A linear discriminant is fitted on the training patients of a patient-grouped stratified five-fold split, the held-out patients are projected onto that single axis, and the projection is clustered with k -means without labels; the comparison arm is k -means on the same held-out records in their native representation. Fitting the discriminant on every label and then clustering the projection would recover the labels by construction, which is why the projected patients are held out. Over five folds and ten seeds, on 508 or 509 held-out records per fold, the mean ARI rises from 0.160 to 0.549 for HuBERT-ECG, from 0.003 to 0.121 for the conv-AE, from 0.646 to 0.758 for the hand-crafted features and from 0.893 to 0.923 for ECGFounder, and does not move for raw flattening, 0.000 to 0.001. These are held-out values, so the native column does not equal the full-cohort k -means ARIs of Table 2; the largest difference is 0.013, on HuBERT-ECG. HuBERT-ECG is where the two readings separate furthest: one direction estimated from other patients recovers more than three times what its own 768 dimensions yield unsupervised, so the class information is present, transfers across patients, and lies away from the directions that dominate the embedding. ECGFounder gains almost nothing, because its dominant axes already carry the split. The conv-AE rises off chance and stops far short of its ceiling of 0.678, so its class structure is both off-axis and diffuse. Raw flattening does not move at all, which separates its failure from HuBERT-ECG’s: no single direction estimated from other patients separates the classes in it, and whatever the non-linear probe recovers from it is not of that form.

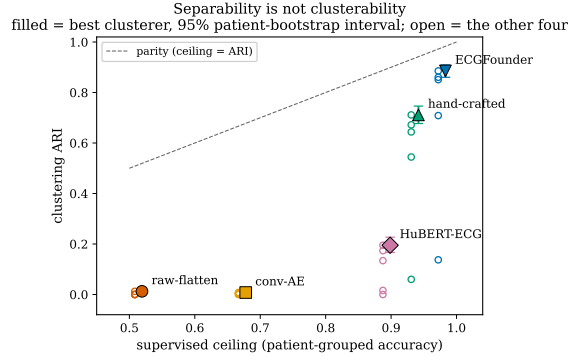


Figure 2: Separability is not clusterability. Supervised ceiling against clustering ARI per representation, on the cohort of 2542 records; points far below parity are separable but not clusterable. All five clusterers are plotted for each representation. The filled marker is that representation’s best clusterer and carries its 95% patient-grouped bootstrap interval; the four open markers are the remaining clusterers, offset slightly along the ceiling axis so that they do not cover the interval bar. The dashed line is parity between the ceiling and the ARI.

4.3. The silhouette trap

Internal indices mislead here in a specific, reproducible way. Raw flattening with agglomerative clustering attains a silhouette of 0.81 while every point but a handful falls into one cluster (largest-cluster fraction 1.00) and the ARI is 0.000. The hand-crafted representation, which actually recovers the classes (ARI 0.711, balanced clusters), has a *lower* silhouette of 0.28 (Fig. 3). A degenerate one-cluster solution maximises compactness while carrying no class information. This is the artefact behind the original project’s silhouette of 0.89, and it is why chance-corrected external metrics, read together with cluster-size sanity checks, are necessary.

4.4. Conditioning matters

The feature scaler changes the outcome materially. For the hand-crafted features the mean ARI over clusterers is 0.629 under standard scaling and 0.606 under power scaling, but falls to 0.360 under robust scaling and 0.229 under quantile scaling. The effect is not confined to hand-crafted features: ECGFounder’s mean ARI drops from 0.866 (standard) to 0.093 under robust scaling. Robust scaling, a common default for physiological signals, leaves heavy arrhythmic tails unbounded so that a few outliers dominate Euclidean distance and the partition collapses. Conditioning should be reported and swept, not assumed.

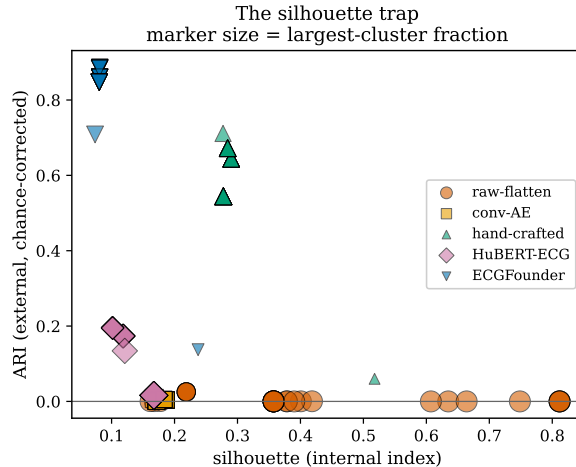


Figure 3: The silhouette trap. Internal silhouette score against chance-corrected ARI; marker size is the largest-cluster fraction. The degenerate one-cluster solutions (large markers, lower right) attain high silhouette at zero ARI. The cohort of 2542 records at power scaling, one point per seed for k -means, spectral clustering and the Gaussian mixture and one point for each deterministic Ward and HDBSCAN fit: 157 points over 22 of the 25 representation-by-clusterer cells. The three HDBSCAN fits that label the whole cohort as noise leave the silhouette undefined and are not plotted. Each point is a single clustering solution, so the panel carries no interval.

4.5. Generalization to MIT-BIH beat morphology

The ordering replicates on a different dataset and a different task type (Table 3). On the N-vs-V beat-morphology task the 21-dimensional hand-crafted features cluster well (ARI 0.783), whereas the raw beat window does not (ARI 0.018 with k -means and 0.049 with agglomerative clustering) despite beats being R-peak aligned. The conv-AE is again the clean separability-versus-clusterability case: a supervised ceiling of 0.927 but a best ARI of only 0.127, from the Gaussian mixture. That the same pattern appears for rhythm on PTB-XL and for morphology on MIT-BIH supports the representation-level reading.

4.6. Robustness under deployment stress: imbalance and federation

Two stressors closer to deployment separate the representations more sharply than the balanced benchmark. Under imbalance the cheap features overtake the foundation embedding; under simulated federation the two means are close and the cheap features retain a larger fraction of their centralized ARI and a higher worst-case partition. First, class imbalance. At

Table 3: MIT-BIH N-vs-V beat morphology: clustering ARI (power scaling). The k -means and Gaussian-mixture rows are means over ten seeds; the agglomerative row is a single deterministic fit. Supervised ceilings: beat-morphology 0.925, conv-AE 0.927, raw-beat 0.741. Foundation models are not applicable to 2-lead beats.

Clusterer	beat-morphology (21-d)	conv-AE (32-d)	raw-beat ($\sim$ 400-d)
k -means	0.750	0.110	0.018
agglomerative	0.783	0.071	0.049
GMM	0.679	0.127	0.018

the natural-like 9:1 SR:AF ratio (majority 0.90), every representation’s ARI falls, because isolating a 10% minority cluster with a partitional method is intrinsically hard, but the hand-crafted features retain the most structure over the ten seeds (ARI 0.265 [0.250, 0.280], AMI 0.245 [0.233, 0.257]) and now *overtake* ECGFounder (ARI 0.063 [0.061, 0.065], AMI 0.148 [0.143, 0.152]), with no overlap between the two ARI intervals. HuBERT-ECG reaches the same ARI as ECGFounder (0.061 [0.053, 0.069]) at a lower AMI (0.044 [0.037, 0.050]), and the raw and conv-AE pipelines stay at chance (0.000 and 0.004). The foundation embedding’s balanced-data lead does not survive imbalance.

Second, federation. Under IID federation across ten sites, every representation retains essentially its centralized ARI, which validates the federated procedure (Table 4; Fig. 4). Non-IID client heterogeneity degrades all representations, but unequally. The hand-crafted features are the most robust, retaining 92% of their centralized ARI (0.644 to 0.591). ECGFounder’s centralized advantage erodes from 0.885 to 0.612. The two means are close, 0.612 against 0.591, and they do not carry the same risk. Across the ten redrawn site partitions the hand-crafted bootstrap interval is [0.546, 0.631] and the worst single partition still returns 0.455, whereas ECGFounder’s interval is [0.382, 0.796] and three of its ten partitions fall to 0.132 or below, one of them to chance. A deployment claim rests on the worst site draw rather than on the average one, and on that reading the 16 hand-crafted features are the more predictable of the two under site heterogeneity. Across both stressors the compact interpretable representation is the more dependable choice for privacy-preserving, multi-site phenotyping.

4.7. Deployment cost

The deployability claim is about cost as well as accuracy, so Table 5 states what each representation costs to compute. The two are ordered

Table 4: Federated robustness: mean ARI over ten seeds and ten simulated sites. Retention is the ratio of non-IID to centralized ARI.

Representation	centralized	federated IID	federated non-IID	retention
ECGFounder	0.885	0.885	0.612	0.69
hand-crafted	0.644	0.643	0.591	0.92
HuBERT-ECG	0.174	0.172	0.106	0.61
raw-flatten	0.000	0.000	0.000	N/A

Table 5: Deployment cost per representation. Inference is the raw strip to the representation with the extractor already loaded, including that representation’s own preprocessing, timed over 100 records and reported as the median of three passes with the spread across those passes. Peak memory is the peak resident set of the process that ran the inference, each representation measured in its own process. Clustering is k -means over the full 2542-record cohort at that dimensionality, median of three fits. All figures are wall-clock on one CPU of a single machine (Apple silicon laptop, no GPU) and are for comparison between the rows rather than as absolute deployment budgets. Parameters are the encoder’s; raw flattening and the hand-crafted features have none. The conv-AE additionally costs 59.2s of one-off training, which is not part of its inference column, and loading the pretrained encoders costs 5.4s for HuBERT-ECG and 0.8s for ECGFounder.

Representation	dim	params (M)	infer. (ms/rec.)	mem. (MB)	clust. (s)
raw-flatten	1992	0	0.058 ± 0.001	261.6	0.298
conv-AE	32	0.079	1.441 ± 2.794	522.5	0.036
hand-crafted	16	0	28.296 ± 6.229	364.0	0.022
HuBERT-ECG	768	92.799	37.530 ± 0.687	1864.8	0.220
ECGFounder	1024	30.819	434.713 ± 18.994	1037.6	0.163

against each other: ECGFounder, which clusters best, is the most expensive per record, fifteen times the hand-crafted extraction, and both foundation models need a gigabyte or more of resident memory against 364 MB for the hand-crafted path, which carries no learned parameters at all. Clustering itself is negligible everywhere, under 0.3s for the whole cohort at any of the five dimensionalities, so the cost of a representation is the cost of extracting it.

5. Discussion

The practical reading is a decision rule, and it is a rule for the two tasks evaluated here: SR-versus-AF rhythm separation on PTB-XL and N-vs-V beat morphology on MIT-BIH. If a strong public foundation model can be run on balanced data in the target 12-lead format, ECGFounder gives the

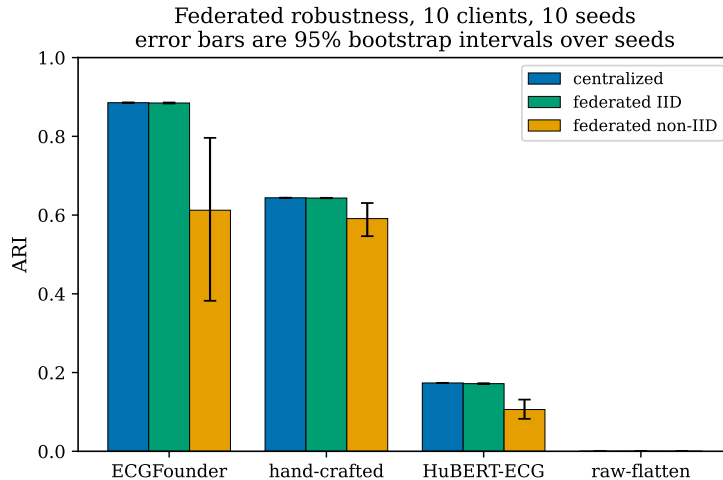


Figure 4: Federated robustness across ten sites. Non-IID heterogeneity erodes the foundation model’s lead while the hand-crafted representation is largely retained. The cohort of 2 542 records partitioned across ten simulated sites over ten seeds, with the site partition and the centroid initialisation redrawn at each seed. Error bars are 95% bootstrap intervals over the ten seeds, which under non-IID partitioning is the spread over the site draws: [0.382, 0.796] for ECGFounder against [0.546, 0.631] for the hand-crafted features.

best unsupervised separation here. If the data are imbalanced or siloed across sites, if interpretability or low dimensionality matters, or if the leads on hand are not the standard twelve, the 16 hand-crafted features are the better choice: they are within reach of ECGFounder on balanced data, overtake it under class imbalance, and hold far more of their clustering ARI under site heterogeneity, both measured on the PTB-XL cohort. Cost runs the same way as robustness. The hand-crafted extraction is fifteen times cheaper per record than ECGFounder, holds about a third of the resident memory and loads no learned parameters at all (Table 5), so the representation that holds up under both stressors is also the cheaper one to run. The lead constraint is concrete. The two foundation models we test are tied to the 12-lead strip format and cannot embed the 2-lead MIT-BIH beats at all, whereas the hand-crafted features transfer across lead sets. The raw and reconstruction-autoencoder pipelines, including the original project’s design, are not viable for label-free arrhythmia structure on either task: they are indistinguishable from predicting the majority class under all five clusterers.

The result also reframes the original master’s project honestly. Its conclusion that the bottleneck is the representation rather than the clustering

algorithm is correct, and we have quantified it. Its headline metrics were the artefacts of two avoidable choices, purity without a baseline and silhouette on a collapsed embedding. Both are corrected here, and both recur often enough in applied clustering to be worth stating plainly.

We claim no methodological novelty. The separability-versus-clusterability phenomenon is known [1], the federated procedure is standard [21], and the representations are existing tools. Six things separate this study from the closest ECG foundation-model clustering benchmark [2]. A hand-crafted axis, which turns a comparison between foundation models into a comparison between representation families. A supervised ceiling on every representation, linear and non-linear, which tells a space that does not hold the label apart from one that holds it in the wrong geometry. An out-of-fold test of that distinction, in which a discriminant fitted on other patients raises HuBERT-ECG from 0.160 to 0.549 in held-out ARI. A second dataset and a second task type, the MIT-BIH beat morphology that benchmark omits. Two cautions a practitioner can act on, the degenerate silhouette and the scaler dependence. And two deployment stressors with the cost of each representation beside them, which is where the balanced-data ranking stops being the useful answer. None of these is a new method, and the ordering of the representations is not itself the result.

The foundation-model axis is bounded to two models by selection rather than by survey. HuBERT-ECG and ECGFounder both publish weights that load into a frozen feature extractor without fine-tuning, which is what a benchmark of frozen embeddings requires, and they differ in pretraining objective and architecture, a self-supervised transformer [9] against a convolutional model pretrained on more than ten million recordings [10]. They also sit at opposite ends of the clustering range measured here, 0.885 against 0.195, so the foundation arm is not a single point. What the bound does to the conclusions is to scope them: every statement here about a foundation model is a statement about one of these two, the deployment comparison is between the hand-crafted features and ECGFounder specifically, and none of it extends to ECG foundation models as a class. A further model that clusters as well as ECGFounder on balanced data would leave the results reported here unchanged, and would have to be measured again under imbalance and under federation before the deployment reading could be extended to it.

6. Limitations

ECGFounder expects 500 Hz input, and because bulk streaming of the 500 Hz records was rate-limited at the source, the main benchmark embeds

the 100 Hz cohort upsampled to the expected length. A consistency check on a 300-record subset, rather than a full re-run, shows this does not materially change the result: ECGFounder clusters almost identically at native 500 Hz and at upsampled 100 Hz (ARI 0.896 versus 0.909, a difference of 0.013), and since upsampling can only discard information it cannot inflate the foundation model’s lead. We evaluate one model per foundation family; other foundation models, including some reported to cluster well [2], are not included, which is why we name the two we test rather than generalising. Both evaluated tasks are binary by design, one rhythm contrast and one morphology contrast, each chosen so that the class definition is unambiguous and a clustering failure can be charged to the representation rather than to label noise. No multi-class arrhythmia cohort is evaluated, and no result here is a statement about multi-class arrhythmia structure, where several rhythms share an irregularity signature and the largest class outnumbers every minority. A multi-class PTB-XL rhythm cohort over the same cached embeddings is the cheapest way to test whether the ordering survives that setting. The MIT-BIH arm omits the foundation models because they are tied to the 12-lead strip format. The federated study is a simulation on a single physical dataset, not a multi-institution deployment. Its only control on heterogeneity is the Dirichlet concentration, which moves class prevalence across sites and nothing else, and the spread that alone produces is already wide: over the ten redrawn non-IID partitions ECGFounder ranges from chance to 0.881, while the hand-crafted features stay between 0.455 and 0.643. Device vendor, lead placement, population case mix and local labelling practice also differ between hospitals and are absent from the simulation. Settling their effect needs recordings held at several institutions, clustered under a federated protocol with the site of origin retained as a variable, which was not available for this study.

7. Conclusion

For unsupervised arrhythmia structure in the ECG, the representation, not the clustering algorithm, decides whether clustering works, and chance-corrected external evaluation is required to see it, on the two tasks evaluated here: SR-versus-AF rhythm separation in PTB-XL and N-vs-V beat morphology in MIT-BIH. On balanced PTB-XL data ECGFounder, pretrained on more than ten million recordings [10], leads; under class imbalance 16 interpretable hand-crafted features overtake it, and under simulated federation across ten heterogeneous sites they are the more predictable of the two,

retaining 0.92 of their centralized ARI against 0.69 and never falling below 0.455 over the ten site draws. The raw and reconstruction-autoencoder pipelines sit at chance under all five clusterers. Internal indices and feature conditioning are not neutral and must be reported. Across the settings tested, the deciding choices were the representation and the evaluation protocol rather than the clustering algorithm; whether the same holds for multi-class arrhythmia cohorts, for other foundation models and for a real multi-institution federation is open.

Code and data availability

PTB-XL (version 1.0.3) [14] and the MIT-BIH Arrhythmia Database (version 1.0.0) [15] are public on PhysioNet. The benchmark code, with the exact cohort definitions, is available at <https://github.com/TheAxeC/ecg-representation-clustering>.

Ethics

This study uses only de-identified, publicly available databases and required no new human-subjects data collection.

CRedit author contributions

Dongho Chun: methodology, software, investigation, writing – original draft. **Axel Faes:** conceptualization, methodology, supervision, writing – review and editing.

Funding

This work was supported by the Flanders AI Research Program (FAIR), an Impulse Program of the Flemish Government (Flemish Ministry of Economy, Science and Innovation).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a generative AI assistant (Claude) to support software implementation and to edit and improve the language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] X. Weng, J. An, X. Ma, B. Qi, J. Luo, X. Yang, J. S. Dong, L. Huang, Clustering properties of self-supervised learning, arXiv preprint arXiv:2501.18452 (2025). doi:10.48550/arXiv.2501.18452.
- [2] F. Filice, E. De Rose, S. Bartucci, F. Calimeri, S. Perri, Looking beyond accuracy: A holistic benchmark of ECG foundation models, arXiv preprint arXiv:2601.21830 (2026). doi:10.48550/arXiv.2601.21830.
- [3] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, Heart rate variability: standards of measurement, physiological interpretation, and clinical use, *Circulation* 93 (5) (1996) 1043–1065. doi:10.1161/01.CIR.93.5.1043.
- [4] P. de Chazal, M. O’Dwyer, R. B. Reilly, Automatic classification of heartbeats using ECG morphology and heartbeat interval features, *IEEE Transactions on Biomedical Engineering* 51 (7) (2004) 1196–1206. doi:10.1109/TBME.2004.827359.
- [5] H. He, Y. Tan, J. Xing, Unsupervised classification of 12-lead ECG signals using wavelet tensor decomposition and two-dimensional gaussian spectral clustering, *Knowledge-Based Systems* 163 (2019) 392–403. doi:10.1016/j.knosys.2018.09.001.
- [6] D. Kiyasseh, T. Zhu, D. A. Clifton, CLOCS: Contrastive learning of cardiac signals across space, time, and patients, in: *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 5606–5615.
- [7] B. Gopal, R. Han, G. Raghupathi, A. Y. Ng, G. H. Tison, P. Rajpurkar, 3KG: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations, in: *Proceedings of Machine Learning for Health (ML4H)*, 2021, pp. 156–167.

- [8] T. Mehari, N. Strodthoff, Self-supervised representation learning from 12-lead ECG data, *Computers in Biology and Medicine* 141 (2022) 105114. doi:10.1016/j.compbiomed.2021.105114.
- [9] E. Coppola, M. Savardi, M. Massussi, M. Adamo, M. Metra, A. Signoroni, HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications, *medRxiv preprint* 2024.11.14.24317328 Pretrained weights: <https://huggingface.co/Edoardo-BS/hubert-ecg-base> (2024). doi:10.1101/2024.11.14.24317328.
- [10] J. Li, A. Aguirre, J. Moura, C. Liu, L. Zhong, C. Sun, G. Clifford, B. Westover, S. Hong, An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains, *arXiv preprint arXiv:2410.04133* Pretrained weights: <https://huggingface.co/PKUDigitalHealth/ECGFounder> (2024). doi:10.48550/arXiv.2410.04133.
- [11] L. Hubert, P. Arabie, Comparing partitions, *Journal of Classification* 2 (1) (1985) 193–218. doi:10.1007/BF01908075.
- [12] N. X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, *Journal of Machine Learning Research* 11 (2010) 2837–2854.
- [13] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.
- [14] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, *Scientific Data* 7 (1) (2020) 154. doi:10.1038/s41597-020-0495-6.
- [15] G. B. Moody, R. G. Mark, The impact of the MIT-BIH arrhythmia database, *IEEE Engineering in Medicine and Biology Magazine* 20 (3) (2001) 45–50. doi:10.1109/51.932724.
- [16] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* 101 (23) (2000) e215–e220. doi:10.1161/01.CIR.101.23.e215.

- [17] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S. H. A. Chen, NeuroKit2: A Python toolbox for neurophysiological signal processing, *Behavior Research Methods* 53 (4) (2021) 1689–1696. doi:10.3758/s13428-020-01516-y.
- [18] G. D. Clifford, C. Liu, B. Moody, L.-w. H. Lehman, I. Silva, Q. Li, A. E. W. Johnson, R. G. Mark, AF classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology challenge 2017, in: *Computing in Cardiology Conference (CinC)*, Vol. 44, 2017, pp. 1–4. doi:10.22489/CinC.2017.065-469.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [20] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: *Advances in Knowledge Discovery and Data Mining (PAKDD)*, 2013, pp. 160–172. doi:10.1007/978-3-642-37456-2_14.
- [21] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, Communication-efficient learning of deep networks from decentralized data, in: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.