Uncovering Heart Rate Response Patterns to Threat Pictures through Deep Latent Representation Learning with a Variational Autoencoder

*Stephan Moratti[1, 2] & Sergio Felipe Calvo García[1, 2]

[1]Department of Experimental Psychology, Complutense University of Madrid, Spain

[2]Emotional Processing Laboratory (EPL), Center for Cognitive and Computational Neuroscience C3N, Complutense University of Madrid, Spain

*Corresponding author: Stephan Moratti, PhD
Department of Experimental Psychology,
Emotional Processing Laboratory, C3N,
Complutense University of Madrid,
Campus Somosaguas,
28223 Pozuelo de Alarcón (Madrid),
Spain,
Email: smoratti@ucm.es

Key words: Heart Rate, Threat, Orienting, Defense, Deep Learning, Variational Autoencoder

Abstract

Group-level averaging of psychophysiological data often obscures meaningful individual differences, masking response patterns that may explain variability in behavior and central nervous system activity. Identifying such patterns is particularly relevant in heart rate (HR) responses to threat, where subtle variations may reflect distinct coping mechanisms such as orienting or defense. Machine learning techniques that learn latent representations, particularly variational autoencoders (VAEs), offer powerful tools for revealing such hidden structures.

This methodological report introduces a simple VAE-based approach for characterizing HR responses to threat pictures in 165 participants. To validate the method, simulations first demonstrated that the model accurately separated noisy sine and cosine waveforms. The VAE was then applied to empirical HR responses, mapping them into a two-dimensional latent space for subsequent cluster analysis, which was compared to clustering based directly on raw HR waveforms.

The VAE revealed three distinct response profiles: (1) strong decelerators (fear bradycardia), (2) weak decelerators with late acceleration, and (3) immediate accelerators without a decelerative phase. In contrast, clustering raw HR waveforms identified only two groups. Clusters derived from the latent space were more coherent and exhibited greater within-group consistency. Finally, applying the pre-trained autoencoder to a small fear-conditioning dataset enabled characterization of distinct HR response patterns despite limited sample size. These findings show that even a basic autoencoder enhances the categorization of psychophysiological response patterns, offering a framework for linking individual autonomic variability to broader models of affective and defensive behavior.

Introduction

Humans and animals flexibly react to threats depending on threat imminence, which can be defined as the spatial or psychological proximity of the threatening stimulus (Gladwin et al., 2016; Löw et al., 2015; Mobbs et al., 2020; Roelofs, 2017). Thereby, two fundamental response strategies have been identified: orienting and defense (Graham & Clifton, 1966; Pavlov, 1927; Sokolov, 1963). Critically, these strategies involve distinct physiological and behavioral response patterns that are highly adaptive depending on threat imminence, reflecting either sympathetic or parasympathetic dominance. Distant threat cues typically elicit a parasympathetically dominated orienting response, which inhibits ongoing behavior, increases attention toward the potential threat, and facilitates its sensory processing (Echegaray & Moratti, 2021; Lojowska et al., 2015; Martín et al., 2025; Moratti et al., 2004; Roelofs et al., 2010). This provides the basis for model-based, flexible action planning. In contrast, imminent counter-strike situations are best encountered by rapid, hard-wired, model-free responses (Mobbs et al., 2020). These evoke sympathetically dominated defensive responses, characterized by action preparation and "sensory rejection" (Martín et al., 2025; Roelofs, 2017; Sokolov, 1963).

Studies using fear conditioning and inherently threatening visual stimuli have shown that humans differ considerably in the extent to which orienting versus defensive response patterns are evoked by the same threat cue. For example, the same fear-conditioned stimuli elicit heart rate (HR) acceleration in some participants and HR deceleration (fear bradycardia) in others (Battaglia et al., 2024; Hamm & Vaitl, 1996; Hodes et al., 1985; Moratti et al., 2006; Moratti & Keil, 2005; Sevenster et al., 2015). Critically, in fear conditioning, HR accelerators engage their defense system, as evidenced by potentiated startle responses to the fear-conditioned cue; a pattern not observed in HR decelerators (Hamm & Vaitl, 1996; Sevenster et al., 2015). Similarly,

inherently threat-related pictures, such as mutilation or attack scenes, typically evoke fear bradycardia in most participants, which has been interpreted as an orienting response to motivationally significant cues (Bradley, 2009; Bradley et al., 2012). However, HR acceleration is also observed in a subset of participants who report stronger fear of mutilation or who avoid approaching threat pictures more quickly than HR decelerators (Klorman et al., 1977; Martín et al., 2025).

Together, these findings demonstrate that subjectively perceived threat imminence modulates HR patterns that index either parasympathetically dominated orienting or sympathetically dominated defense. Interestingly, HR-indexed orienting has been associated with increased visual processing in the visual cortex and reduced excitatory cortical motor-circuit preparation, whereas HR-indexed defense shows the opposite pattern (Echegaray & Moratti, 2021; Martín et al., 2025). This aligns with theoretical accounts of sensory intake during orienting and sensory rejection during defense (Graham & Clifton, 1966; Lacey & Lacey, 1970; Sokolov, 1963).

However, the classification of participants into HR decelerators and accelerators, and the associated strategic response type toward threat cues, also depends on the methodology used to identify the corresponding HR response patterns. Although all reported findings on subclassifications of HR accelerators and decelerators rely on some form of clustering algorithm (e.g., hierarchical clustering in Moratti et al., 2006; Moratti & Keil, 2005); k-means clustering of predefined time windows in Hamm & Vaitl, 1996; Hodes et al., 1985; Sevenster et al., 2015; a simple zero-threshold approach in Echegaray & Moratti, 2021; or k-means on all time bins in Martín et al., 2025), a purely data-driven method that captures the full richness of HR response waveforms at the individual level would be ideal.

Recent advances in machine learning for time-series classification are promising in this regard. For example, autoencoders can learn to reduce a time series into a low-dimensional latent space and then reconstruct the series from that space. These approaches have typically been used to detect artifacts or anomalies, based on a pre-trained autoencoder trained on "ideal" data (Hu et al., 2024; *Mathworks/Anomaly-Detection-Using-Variational-Autoencoder-VAE-*, 2020/2024; Yang & Paparrizos, 2025; Yu et al., 2023). Here, we propose using a simple variational autoencoder to encode HR response waveforms to threat stimuli into a two-dimensional latent space and then decode these response patterns from that space. This allows each participant to be located within a two-dimensional Cartesian space. Subsequently, cluster analysis can be applied to this latent space to identify groups of participants with similar HR waveform shapes. Another advantage is that, once an autoencoder has been trained on a sufficiently large sample, it can be applied to new, smaller samples to classify participant groups showing different HR response patterns.

Here, we present a methodological approach that combines an autoencoder with a Bayesian Gaussian Mixture Model (BGMM) clustering procedure to demonstrate the usefulness of autoencoders for identifying groups of HR responders to threat stimuli. First, the approach is introduced using simulated data (sine and cosine waveforms with added noise) to show that the autoencoder, together with BGMM clustering, can successfully identify different types of time series based on ground-truth data. Second, the autoencoder is applied to real HR change response waveforms obtained from three independent samples of two previous studies and of one ongoing study (N = 165, Echegaray & Moratti, 2021; Martín et al., 2025; Calvo García et al., in preparation). Finally, we compare the performance of the autoencoder in clustering participants with similar HR change waveforms against our previous approach of applying clustering

directly to all time bins (Martín et al., 2025). Because one-dimensional convolutional layers in an autoencoder are expected to capture the richness of individual HR change waveforms, we hypothesize that the autoencoder will outperform clustering applied directly to the raw data. Importantly, all data, statistical analysis and code for performing the analysis presented in this methodological report are available and can be re-used for future research (https://osf.io/k58uy/).

Methods

Participants

All participants were sampled from two published studies (Echegaray & Moratti, 2021; Martín et al., 2025) and one ongoing study (Calvo Garcia et al., in preparation; total N = 165; mean age = 23.2 years, range = 18 years – 55 years, 111 females, 16 left-handed). All volunteers had normal or corrected-to-normal vision and provided written informed consent for participation in the corresponding experiments. Participants received course credit for their participation. Exclusion criteria included a history of epilepsy, family history of epilepsy, or self-reported psychiatric or cardiac pathology. Screening for epilepsy-related conditions was necessary because all three experiments used flickering visual stimuli to evoke steady-state visual fields or potentials (measured with MEG and EEG, respectively). However, MEG and EEG data are not reported here. All three studies were approved by the Ethics Committee of the Complutense University in accordance with the Declaration of Helsinki. Data were stored in compliance with the Spanish and European data protection laws (Ley Orgánica 15/1999 LOPD and Real Decreto 994/1999).

Procedure and Stimuli

After having received instructions about the experimental tasks, all participants were shown an example of mutilation and attack scenes (that were not later presented during the experiment) to decide whether they wished to participate. All participants then signed an informed consent form. Following completion of the experiments, participants were fully debriefed about the purposes of the study.

In all three experiments, threat-related pictures (mutilation and attack scenes) and neutral pictures (neutral faces and household objects) were taken from the International Affective Picture System (IAPS; Lang et al., 2005). In the study by Echegaray & Moratti (Echegaray & Moratti, 2021), 20 threat and 20 neutral pictures were selected, with mean normative valence ratings of 2.76 (SD = 0.19) and 4.87 (SD = 0.05), respectively. Mean normative arousal ratings were 7.33 (SD = 0.13) for threat scenes and 2.80 (SD = 0.13) for neutral pictures. In the study by Martín et al. (Martín et al., 2025), 40 threat and 40 neutral pictures were used, with mean normative valence ratings of 2.02 (SD = 0.46) and 4.75 (SD = 0.25), respectively. Mean normative arousal ratings were 6.71 (SD = 0.37) for threat scenes and 2.65 (SD = 0.36) for neutral pictures. For each participant, 20 threat and 20 neutral pictures were randomly selected from the original pool of 40 per category. Pictures were presented in two experimental blocks with different pseudo-randomized sequences (with no more than three pictures of the same category presented consecutively), resulting in a total of 40 threat and 40 neutral pictures. In the ongoing study (Calvo Garcia et al., in preparation), the same 40 threat and 40 neutral IAPS pictures as in the Martín et al. (2025) study were used, but without selecting a subset of 20 per category. Thus, participants viewed all 40 threat and all 40 neutral pictures in pseudo-randomized order across two experimental blocks (again with no more than three pictures

of the same category presented consecutively) and without repetitions. Mean valence and arousal ratings correspond to those reported in the Martín et al. (2025) study.

Across all experiments, all pictures were matched for luminance and contrast using the SHINE toolbox (version 0.0.4; Ben, 2019) in MATLAB (MathWorks™, R2021b). Pictures were presented centrally on a screen (refresh rate: 60 Hz), subtending a visual angle of 10° horizontally and 7.5° vertically. Participants were instructed to maintain fixation on a central cross (visual angle: 1.6° horizontally and vertically), which remained visible throughout the entire experiment (during both picture presentation and inter-trial intervals, ITIs). To evoke steady-state visual evoked fields (ssVEFs) or potentials (ssVEPs), pictures flickered for 4 s at 10 Hz in the studies by Echegaray & Moratti (2021) and Martín et al. (2025), whereas in the ongoing study the flicker rate was 15 Hz. The ITI following each picture presentation was randomly chosen between 8 s and 12 s in all three studies. In Echegaray & Moratti's report (2021), the task consisted of passive viewing of emotional pictures. By contrast, in the Martín et al. (2025) report and the ongoing study, picture size started to gradually increase after the 4 s presentation, and participants were required to terminate picture presentation by pressing a button to indicate how close they allowed the pictures to approach (results not reported here). After the second block of picture presentations, all participants in all three studies rated valence and arousal using the Self-Assessment Manikin scale (SAM; (Lang et al., 2005)). However, the SAM ratings of eight participants were lost due to technical problems resulting in SAM ratings of 157 participants. The experiments were controlled using the Pychtoolbox (Kleiner et al., 2007) in MATLAB (Mathworks™).

Data acquisition and processing of the Electrocardiogram (ECG) data

As MEG and EEG data from the three studies have been reported elsewhere (Echegaray & Moratti, 2021; Martín et al., 2025) and will be reported for the ongoing study in the future, only ECG acquisition and processing are described here. In all studies, ECG was recorded using Ag/AgCl electrodes filled with electrolytic gel. In Echegaray & Moratti (2021), electrodes were placed at the right mid-clavicle and lower left rib, sampled at 600 Hz with a 0.1–200 Hz online band-pass filter using the MEG-integrated EEG amplifier (VectorView©, Elekta Neuromag Oy, Helsinki, Finland, 2005). Martín et al. (2025) used the same montage but with a 1000 Hz sampling rate, a 0.1–100 Hz band-pass filter, and an additional 50 Hz notch filter. In the ongoing study (Garcia Calvo at el., in preparation), recordings were obtained with two active, pre-amplified electrodes placed at the right and left mid-clavicles, using a BrainAmp amplifier (BrainProducts©) with a 1000 Hz sampling rate and a 0.1–100 Hz band-pass filter plus a 50 Hz notch filter. Ground electrodes differed across studies: the left earlobe in the MEG study and AFz in the two EEG studies.

Before heartbeat detection (R-peak identification), ECG data were down-sampled to 250 Hz and filtered offline with a FIR band-pass filter (0.1–40 Hz, 60 dB stopband attenuation). In the MEG study (Echegaray & Moratti, 2021), R-peaks were detected using a simple Schmitt trigger. In the study by Martín et al. (2025), a pre-trained long short-term memory (LSTM) network (available in the MATLAB Deep Learning Toolbox; see Martín et al., 2025 for details) was used to identify the QRS complex and corresponding R-peaks. In the ongoing study, R-peaks were detected using the open-source MATLAB toolbox R-DECO (version 1.0.0; https://physionet.org/content/r-

deco/1.0.0/), which applies an envelope-based method combined with an adapted Pan-Tompkins algorithm (Moeyersons et al., 2019).

Across all studies, R-peak detections were visually inspected for artifacts such as noise or ectopic beats, and contaminated trials were omitted from further analysis. For each participant and picture category (threat and neutral), inter-beat intervals (IBIs) derived from R-peaks during a 2 s pre-stimulus and 4 s post-stimulus interval were transformed into beats per minute (bpm) in 0.5 s steps using weighted averages (Reyes del Paso & Vila, 1998), implemented with code from the open-source MATLAB toolbox KARDIA (Perakakis et al., 2010); https://sourceforge.net/projects/mykardia/). Finally, HR change waveforms were baseline-corrected using the 2 s pre-stimulus interval. To improve the autoencoder learning performance, HR data was scaled (-1 to 1) and smoothed using a Gaussian one-dimensional filter ($\sigma = 1$).

Generation of simulated data

To obtain ground truth data for testing the autoencoder (see below), 160 simulated datasets were generated. Specifically, 80 sine and 80 cosine waveforms (amplitude range: –1 to 1) were created using nine time bins, matching the post-stimulus HR data that also consisted of nine time bins from 0 s to 4 s post-stimulus. Random noise drawn from a normal distribution (mean = 0.1, SD = 0.3) was then added to each sine and cosine waveform.

Construction of the variational autoencoder (VAE)

A VAE was implemented in Keras/ TensorFlow (https://keras.io; https://www.tensorflow.org/) to learn two dimensional latent representations of the simulated data and heart rate change waveforms. The encoder consisted of three one-

dimensional convolutional layers with kernel size 3, leaky rectified linear unit activation functions, and progressively increasing filter sizes (4, 16, and 32). Convolutional outputs were down-sampled using a max-pooling layer (pool size = 3, padding = "same") and flattened. Two fully connected layers generated the mean (z_{mean}) and log-variance (z_{logvar}) of a two-dimensional latent Gaussian distribution. Latent variables were sampled using reparameterization:

$$z = z_{mean} + exp(0.5 \cdot z_{logvar}) \cdot epsilon;$$

where *epsilon* is drawn from a standard normal distribution. This allows gradients to propagate through the stochastic sampling step allowing differentiable training of the network. The decoder received these two-dimensional latent vectors as input. These were mapped through a fully connected layer (leaky rectified linear unit activation function), reshaped to match the convolutional input dimensions, and up-sampled (factor = 3). The reconstructed signals were generated through two one-dimensional convolutional layers (filters of 16 and 4; kernel sizes = 3; padding = "same"; leaky rectified linear unit activation function), with the final output layer using linear activation to allow continuous-valued reconstruction. The VAE decoding layer was chosen to be sparser to avoid overfitting given that the time series only consisted of nine time bins.

The VAE was separately trained (epochs = 2000, batch size = 2, validation split = 0.2) for the (i) simulated time series and (ii) the HR change data to threat pictures using a loss function combining the reconstruction error and a Kullback-Leiber (KL) divergence term (that quantifies the distance to the Gaussian Normal distribution). The KL divergence terms regularized the two-dimensional latent space $q(z|x)$ by pulling the encoded latent distributions towards a standard normal distribution $p(z)$ = *N(0,1)*.

$$Reconstruction\ loss = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 ;$$

where x_i and $\hat{x}_i$ are the original and reconstructed time series for each time bin, respectively. *N* is the number of time bins.

$$\mathrm{KL}(\ q(z|x)\ ||\ p(z)\ ) = -\frac{1}{2}\sum(1 + \log(\sigma^2) - \mu^2 - \sigma^2);$$

where μ and σ^2 are the mean and variance of the encoder output. The KL term was summed over latent dimensions and averaged across batches. The final loss function was defined as the sum of the reconstruction loss and a β-weighted KL divergence term. During training, the β value was gradually increased to implement a KL annealing schedule. This ensured that, at the beginning of training (low β values), the reconstruction loss dominated, allowing the network to learn to reconstruct the signals accurately. As β increased, the KL term became more influential, guiding the network to also encode the signals in a smooth and continuous two-dimensional latent space.

Clustering of simulated and HR change waveforms using a Bayesian Gaussian Mixture Model (BGMM)

To identify clusters of sine and cosine time series (simulated data) and participants with similar heart rate waveforms (real data), we applied a Bayesian Gaussian Mixture Model (BGMM) as implemented in the python module *scikit-learn*. The BGMM is a clustering method that groups data based on the assumption that it comes from a mixture of several Gaussian distributions. Unlike the standard Gaussian Mixture Model (GMM), the BGMM automatically determines how many clusters are needed by placing a Bayesian prior on the mixture weights. This allows it to "turn off" unnecessary

components, making it more flexible when the true number of clusters is unknown (Bishop, 2006).

For simulated data, the BGMM was applied to the two-dimensional latent space learned by the autoencoder. If this space adequately captured the sine and cosine waveforms, the BGMM was expected to identify two corresponding clusters. Latent vectors were z-transformed ($\mu = 0$, $\sigma = 1$), and the BGMM was initialized with 10 clusters and 500 iterations. To account for sensitivity to the covariance prior, values between 0.01 and 5.1 (step size 0.1) were tested, and the solution with the highest Silhouette score was selected. The same procedure was applied to the encoded latent space of HR change waveforms evoked by threatening pictures. With respect to HR data, we refer to this method as the AUT approach (autoencoder). For comparison, BGMM clustering was also performed directly on the nine-dimensional HR time series (0–4 s post-stimulus time bins), which were reduced to two components for visualization using principal component analysis (PCA) as implemented in the *sklearn.decomposition* module. With respect to HR data, we refer to this method as the ATP approach (all time points).

Comparing clustering performance

To evaluate the performance of the two clustering approaches (AUT and ATP) for the HR data, we compared the Silhouette and Calinski–Harabasz scores. Both are widely used indicators of cluster quality: the Silhouette score reflects the similarity of an observation to its own cluster (Rousseeuw, 1987), whereas the Calinski–Harabasz index quantifies overall cluster separation versus cohesion (Caliński & Harabasz, 1974). In addition, we assessed the similarity of HR change waveforms within clusters using cosine similarity (CS) scores (implemented in the *sklearn.metrics.pairwise* module), defined as the normalized dot product between two time series. For each participant, we computed

the CS of their HR waveform relative to all other participants within the same cluster. To enable statistical comparison between the AUT and ATP approaches - and to address issues related to unequal cluster numbers as well as dependence versus independence of observations - CS scores were averaged across the clusters obtained for each approach. Because the distributions of CS scores were skewed (see Results, Figure 7), group comparisons were performed using the Wilcoxon signed-rank test and complemented by a Bayesian paired-samples t-test with a scaled Cauchy prior (r = 0.7).

Statistical comparison between HR responses to threating and neutral picture content

As in previous reports (e. g. Martín et al., 2025), each time bin of the HR responses to threat and neutral pictures was compared using paired parametric t-tests, and additionally with Bayesian paired t-tests using a scaled Cauchy prior (r = 0.7), since Bayesian comparisons do not require correction for multiple comparisons. The p-values from the paired parametric t-tests were corrected using the false discovery rate (FDR) procedure by Benjamini & Yekutieli (Benjamini & Yekutieli, 2001) to account for dependence. First, HR change waveforms were compared between neutral and threat picture categories across the entire sample (N = 165) to replicate previously reported fear bradycardia in response to threatening or unpleasant complex emotional scenes (Bradley et al., 2001). The same comparisons were then performed within the clusters obtained using the AUT and ATP approaches.

Application to a fear conditioning data set of small sample size

To evaluate whether the pre-trained autoencoder, together with the pre-trained BGMM, can contribute to identifying distinct HR response patterns to fear-relevant stimuli in a small sample and generalize to simple visual stimuli that had been fear

conditioned, HR responses to a fear-conditioned CS+ and a CS− control stimulus from a previous study (Santos-Mayo & Moratti, 2025) were analyzed. In that study, 35 participants volunteered (23 females, 32 right-handed, mean age 24.97 years, range 19-47 years). Due to technical issues only 33 datasets (21 females, 32 right-handed, mean age 25.4 years, range 20-47 years) were available for the current analysis. In the original study, HR responses were considered only up to 3 s after stimulus onset. However, to ensure compatibility with the pre-trained autoencoder used here, HR responses from 0 s (stimulus onset) to 4 s post-stimulus, with a 2 s pre-stimulus baseline correction, were analyzed. Details regarding R-peak detection can be found in Santos-Mayo and Moratti (2025).

The fear-relevant CS+ HR response waveforms were submitted to the pre-trained autoencoder to obtain their representations in the two-dimensional latent space. Subsequently, the means, covariances, and weights of the pre-trained BGMM were applied to these latent representations to form the corresponding clusters. Finally, the mean HR change from baseline for the CS+ and CS− was calculated for each resulting cluster and compared using the same statistical approach described above. As this is a method contribution Table 1 below lists the corresponding Python scripts for each analysis protocol that can be found on the Open Science Framework server (https://osf.io/k58uy/):

Analysis	Corresponding Python script
VAE and BGMM simulated data	AutoencoderVAE_1dConv_sim.py
VAE and BGMM real HR data (AUT)	AutoencoderVAE_1dConv_HR.py
BGMM direct on HR data (ATP)	BaysGMM_clustering_allTimePoints.py
Comparison of cluster performance	Comparisons.py
Application of pre-trained VAE and BGMM	Apply2CondData.py

Table 1: The analysis steps and the corresponding Python scripts are listed. VAE = variational autoencoder, BGMM = Bayesian Gaussian Mixture Model procedure, AUT = autoencoder approach, ATP = all time points approach, HR = heart rate

Results

SAM ratings

Threat pictures were rated as less pleasant (M = 2.14, SD = 0.81) than neutral pictures (M = 5.34, SD = 0.77; $t(156) = 38.41$, $p < 0.001$, Cohen's $\delta = 3.06$). Threating pictures were also rated as more arousing (M = 6.85, SD = 1.36) than neutral pictures (M = 3.08, SD = 1.36; $t(157) = 30.11$, $p < 0.001$, Cohen's $\delta = 2.40$). In sum, participant's SAM ratings were in line with the normative IAPS SAM ratings (see methods).

Autoencoder performance on simulated data

Figure 1 shows the training loss function of the autoencoder. Further, two representative simulated waveforms and their reconstructions from a two-dimensional latent space (decoding) by the autoencoder are depicted. Figure 1A illustrates that 2000 training epochs were enough and further training would not have improved the loss function anymore.

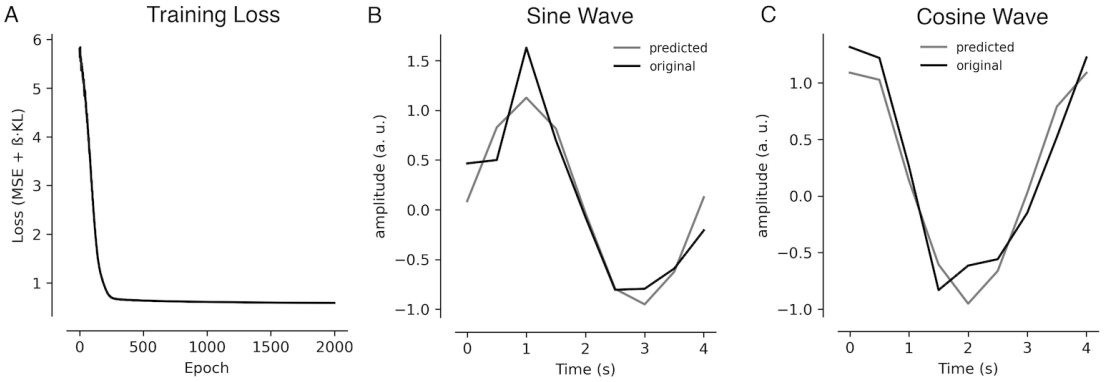

Figure 1: (A) The training loss function is shown. (B) A representative sinus wave (plus noise) and the reconstructed waveform from two-dimensional latent space by the autoencoder is depicted. (C) The same is shown for a representative cosine wave (plus noise). MSE = mean squared error, β = β weight, KL = Kullback-Leiber divergence.

The autoencoder separated the sine and cosine waves into two very well definable clusters within the two-dimensional latent space (see Figure 2). The BGMM clustering procedure applied to the latent space indicated that two active clusters (weights > 1e-2) best describe the positions of the sine and cosine waveforms in the latent space. The two clusters were obtained using a covariance prior of 0.11. This covariance prior was determined by running the BGMM procedure with priors from 0.01 to 5.01 in 0.1 steps and then by selecting the prior that resulted in the highest Silhouette score (0.52). The two clusters as determined by the BGMM procedure also resulted in a Calinski Harabasz score of 167. 96 indicating coherent clusters.

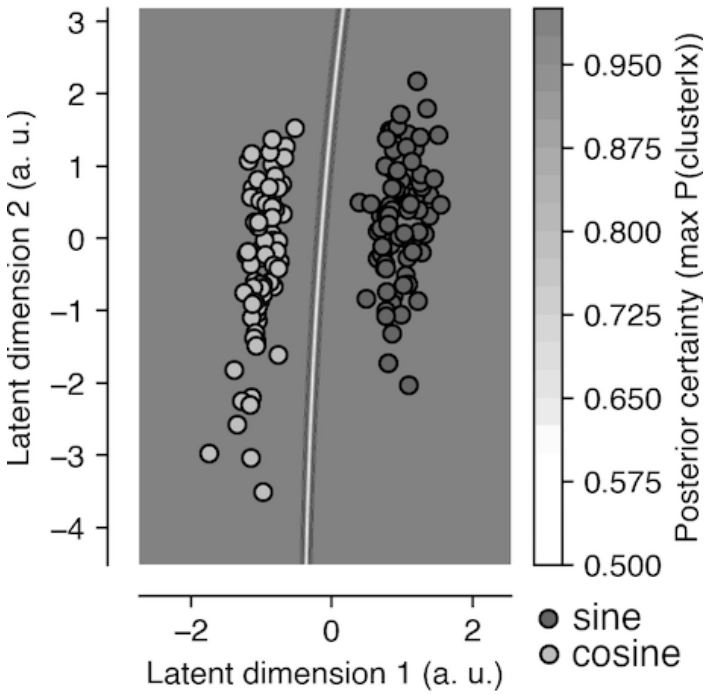

Figure 2: Locations of the 80 sine and 80 cosine waveforms in the two-dimensional space after encoding. A Bayesian Gaussian Mixture Model procedure with a covariance prior of 0.11 indicated best clustering using two clusters (Silhouette score = 0.52, Calinski Harabasz score = 167. 96). The grey shaded heat map indicates the posterior certainty derived from the Bayesian Gaussian Mixture Model clustering procedure. P (cluster | x) = maximal probability given the data.

Overall HR responses across all participants

Considering the overall HR response across all participants (N = 165), threat related pictures evoked increased fear bradycardia (HR deceleration) than neutral picture contents from 1.5 s to 4 s after stimulus onset (maximum t value at 4.0 s: $t(164) = -2.76$, $p_{corrected} = 0.024$, $BF_{10} = 3.35$, median $\delta = -0.21$, 95% CI: [-0.36, -0.06]; minimum t value at 2 s: $t(164) = -4.04$, $p_{corrected} = 0.002$, $BF_{10} = 185.81$, median $\delta = -0.31$, 95% CI: [-0.46, -0.15]; see Figure 3).

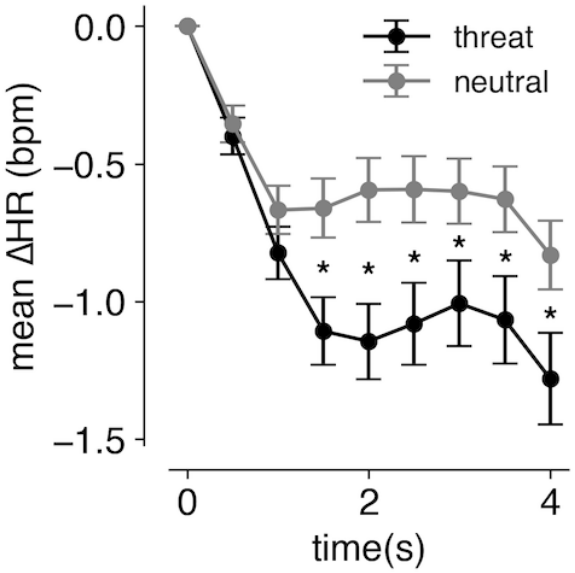

Figure 3: Mean HR changes compared to baseline for threat and neutral picture contents. The error bars represent the standard errors. * $p < 0.05$ (corrected for multiple comparisons) and $BF_{10} >= 3$.

Autoencoder and direct BGMM clustering of HR response data

In this section we will compare between the clustering performance of a BGMM procedure in the two-dimensional latent space after encoding the HR data using an autoencoder (AUT approach) and applying the BGMM directly on the HR data in a nine-dimensional space corresponding to the nine time bins (all time points ATP approach, see methods). Corresponding to the AUT approach Figure 4 shows the training loss function of the autoencoder. Again 2000 training epochs were far enough, and more epochs would not have improved the loss function further. Additionally, two representative HR response waveforms and their corresponding reconstructions from a two-dimensional latent space (decoding) by the autoencoder are shown.

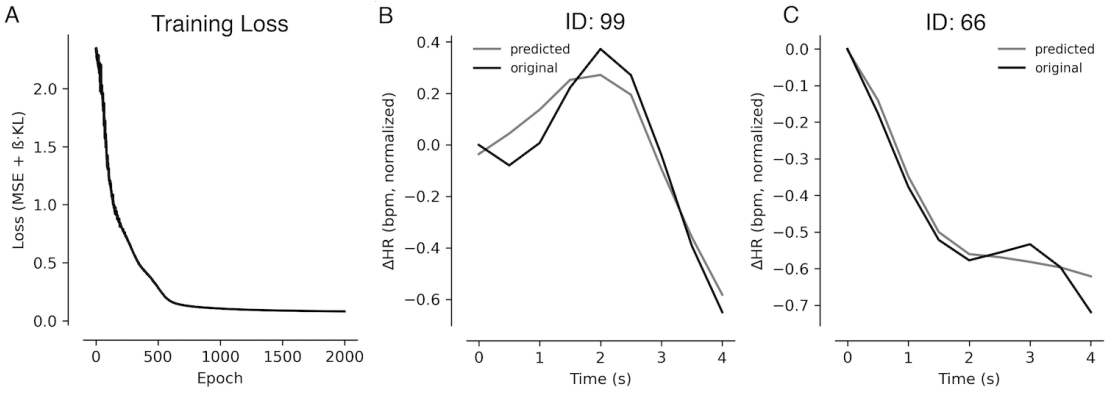

Figure 4: (A) The training loss function is shown. (B) A representative HR waveform (participant ID: 99) and the reconstructed waveform from the two-dimensional latent space by the autoencoder is depicted. (C) The same is shown for another representative participant (ID: 66). MSE = mean squared error, β = β weight, KL = Kullback-Leiber divergence, HR = heart rate, bpm = beats per minute, ID = identity code

The same BGMM procedure as in the simulation study was applied to the locations in the two-dimensional latent space of the autoencoder. Then, the same procedure was repeated directly using the HR change waveforms (ATP approach as done in our previous report that used Kmeans clustering, see Martín-Gil et al., 2025). For both approaches (AUT and ATP) the best covariance priors were determined by applying priors between 0.01 to 5.1 in 0.1 steps and selecting the prior that resulted in the greatest Silhouette score for the clusters. The locations in the two-dimensional latent space of the autoencoder were best fitted by 3 clusters (AUT approach Figure 5A, covariance prior = 2.11, Silhouette score = 0.56, Calinski-Harabasz score = 235.13). In contrast, the BGMM explained best the locations of the participants in the nine-dimensional time bin space by 2 clusters (ATP approach Figure 5B, covariance prior = 0.11, Silhouette score = 0.55, Calinski-Harabasz score = 197.71). Although the Silhouette scores of both approaches were similar, the Calinski-Harabasz score indicated superior clustering in the latent space of the autoencoder.

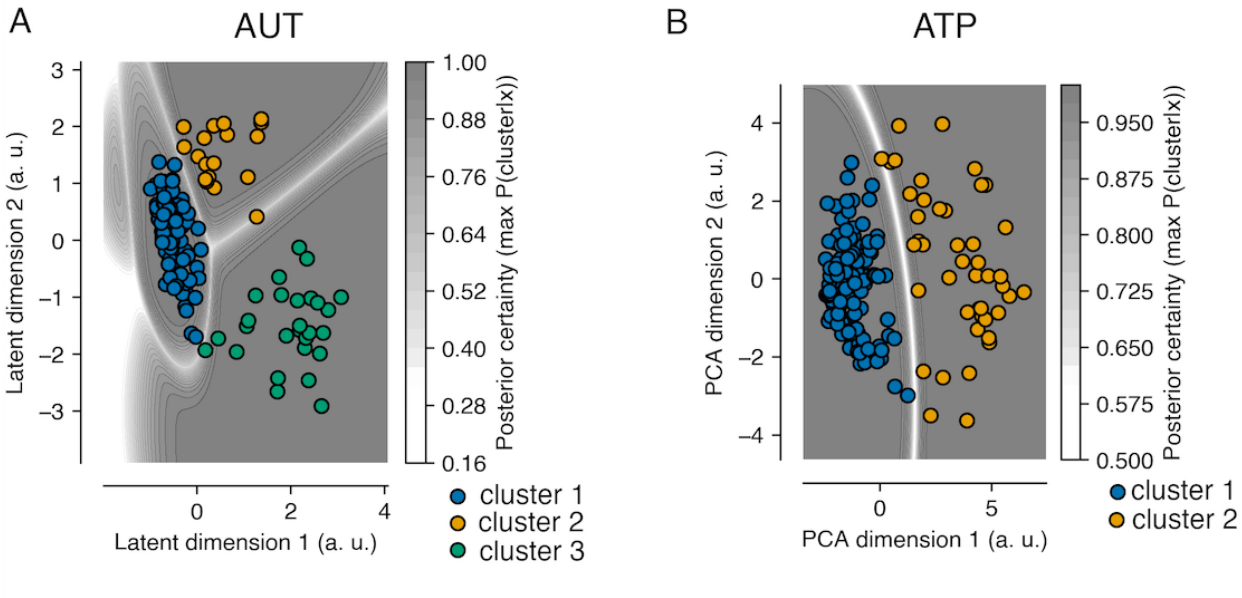

Figure 5: (A) The locations (colored dots) of the encoded HR responses in two-dimensional latent space are shown. A Bayesian Gaussian Mixture Model procedure with a covariance prior of 2.11 indicated best clustering using three clusters (Silhouette score = 0.56, Calinski-Harabasz score = 235.13). The grey heat map indicates the posterior certainty. (B) The same is shown for the two clusters in nine-dimensional (time bin) space projected by a PCA into a two-dimensional space. Here, a Bayesian Gaussian Mixture procedure with a covariance prior of 0.11 indicated best clustering using two clusters (Silhouette score = 0.55, Calinski-Harabasz score = 197.71). AUT = autoencoder approach, ATP = all time points approach. max P(cluster | x) = maximal probability of the cluster given the data.

The AUT approach resulted in three clusters (cluster 1: N = 120, cluster 2: N = 27, cluster 3: N = 18). The mean HR change responses for the first cluster were characterized by increased HR deceleration (fear bradycardia) for threat pictures in comparison to neutral images (Figure 6A; increased deceleration from 1 s to 4 s after picture onset: minimum t-value at 2 s: t(119) = -7.09, $p_{corrected}$ = 0.002, BF_{10} = 8.43 · 10^7, median δ = -0.63, 96% CI: [-0.83, -0.44]; maximum t-value at 1 s: t(119) = -3.12, $p_{corrected}$ = 0.007, BF_{10} = 9.74, median δ = -0.28, 95% CI: [-0.46, -0.10]). The second cluster consisted of participants who accelerated their HR to threat pictures in comparison to neutral images. This cluster was characterized by an increased HR for threat images in comparison to neutral pictures at late latencies from 3 s to 4 s after stimulus onset (Figure 6B; minimum t-value at 3 s: t(26) = 3.4, $p_{corrected}$ = 0015, BF_{10} = 18.14, median δ = 0.61,

95% CI: [0.20 1.02]; maximum t-value at 4 s: $t(26) = 4.12$, $p_{corrected} = 0.007$, $BF_{10} = 88.49$, median $\delta = 0.73$, 95% CI: [0.31, 1.17]). However, we also observed evidence for a tendency for an early increased HR deceleration at 0.5 s after stimulus onset ($t(26) = -2.71$, $p_{corrected} = 0.064$, $BF_{10} = 4.08$, median $\delta = -0.48$, 95% CI: [-0.87, -0.09]). The third AUT cluster indicated HR acceleration for the threat picture content compared to neutral images right after picture onset from 0.5 s to 2.5 s (Figure 6C, minimum t-value at 2.5 s: $t(17) = 2.83$, $p_{corrected} = 0.05$, $BF_{10} = 4.70$, median $\delta = 0.59$, 95% CI [0.10 1.10]; maximum t-value: $t(17) = 5.01$, $p_{corrected} = 0.002$, $BF_{10} = 262.88$, median $\delta = 1.07$, 95% CI [0.48 1.70]).

The ATP method resulted in two clusters (cluster 1: N = 121, cluster 2: N = 44). The mean HR change responses for the first cluster generated by the ATP approach were characterized by increased HR deceleration (fear bradycardia) for threat pictures in comparison to neutral images (Figure 6D; increased deceleration from 1 s to 4 s after picture onset: minimum t-value at 2 s: $t(120) = -7.34$, $p_{corrected} < 0.001$; $BF_{10} = 3.07 \cdot 10^8$, median $\delta = -0.66$, 95% CI: [-0.85, -0.46], maximum t-value at 1 s: $t(120) = -3.57$, $p_{corrected} = 0.002$, $BF_{10} = 38.33$, median $\delta = -0.32$, 95% CI: [-0.50, -0.14]). The second cluster of the ATP approach contained participants that were characterized by HR acceleration for threat pictures in comparison to neutral images (Figure 6E; increased HR from 1 s to 4 s after picture onset: minimum t-value at 1 s: $t(43) = 2.79$, $p_{corrected} = 0.024$, $BF_{10} = 4.89$, median $\delta = 0.40$, 95% CI: [-0.10, -0.70]; maximum t-value at 3 s: $t(43) = 4.26$, $p = 0.001$, $BF_{10} = 379.77$, median $\delta = 0.64$, 95% CI: [0.31, 0.96]).

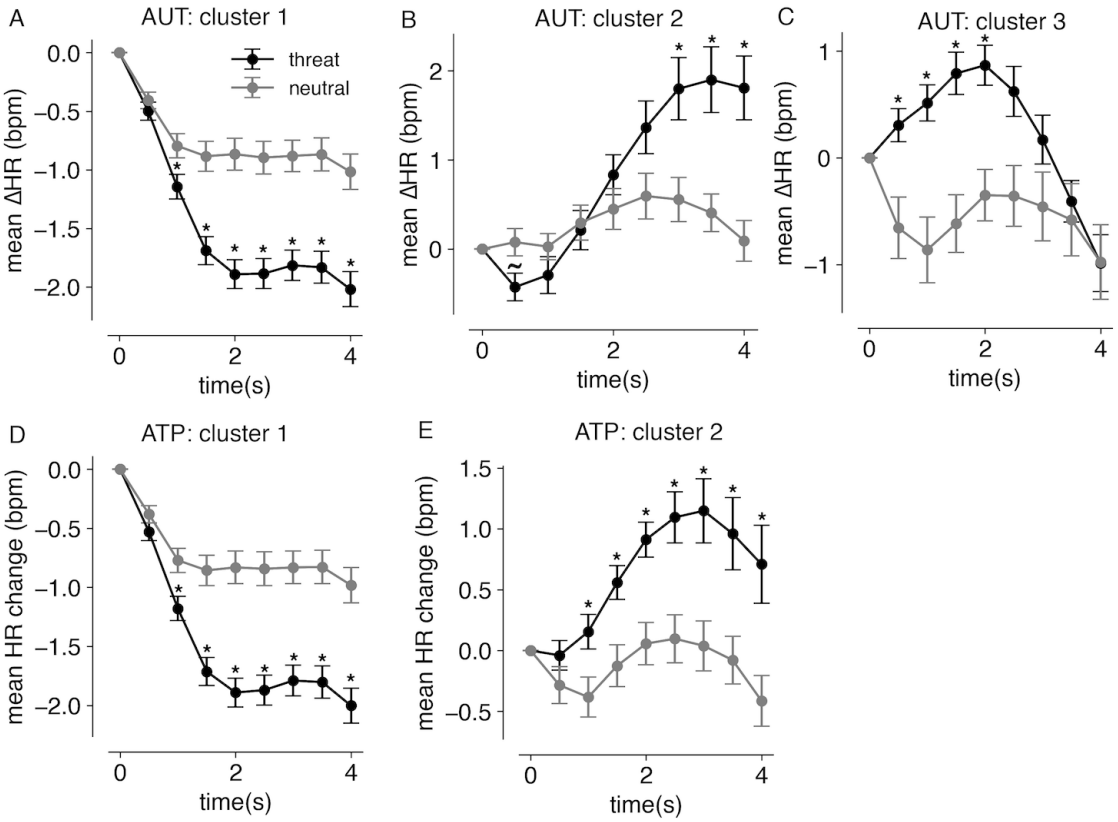

Figure 6: (A) Mean heart rate (HR) changes elicited by threat and neutral pictures for the first cluster produced by the autoencoder approach are shown. (B & C) show the same information for clusters 2 and 3. (D) Mean HR changes for threat and neutral picture content for the first cluster generated by directly clustering using all time points are shown. (E) The same information is shown for the second cluster. Error bars represent s. e. m. AUT = autoencoder, ATP = all time points. * $p_{corrected}$ <= 0.05 and BF_{10} >= 3, ~ $p_{corrected}$ = 0.06 and BF_{10} >= 3

Although the emergence of two different accelerators groups and the Calinski-Harabasz score indicated superior clustering for the AUT approach, HR change waveforms within clusters should be more similar with respect to their shapes for the AUT than for the ATP approach if the AUT approach generated more consistent HR responder groups. The mean cosine similarity scores across clusters between HR change time series for threat related pictures increased significantly for each participant when changing the approaches from ATP to AUT (Figure 7; Wilcoxon = 9233, z = 3.881, p < 0.001; BF_{10} = 522.74, median δ = 0.33, 95% CI: [0.18 0.49]).

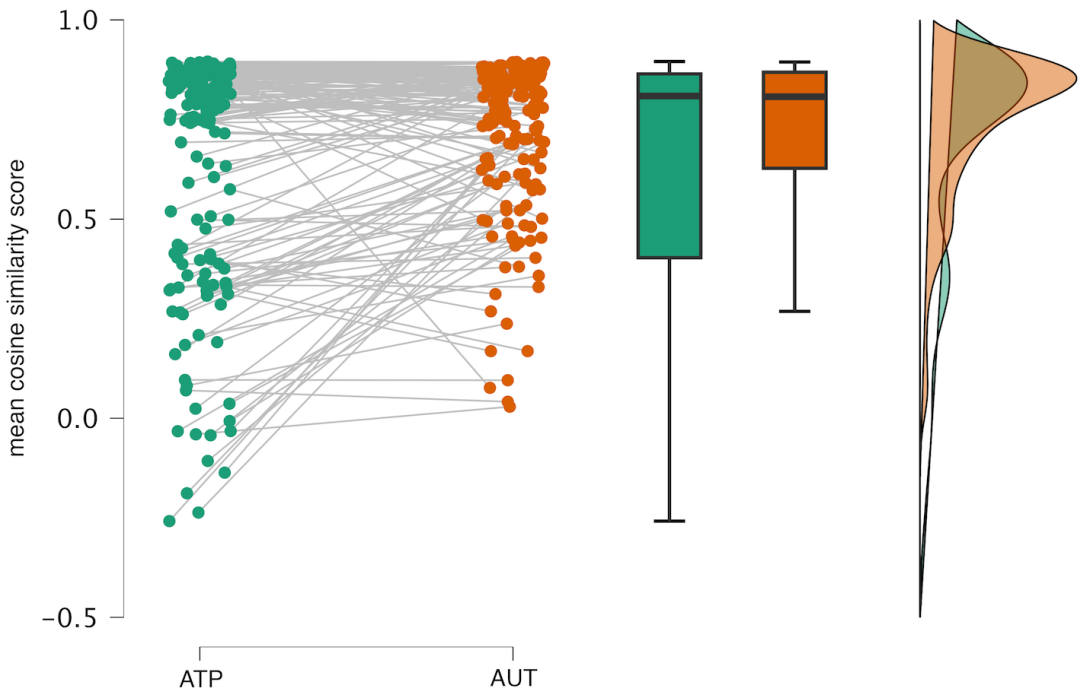

Figure 7: The left panel shows the raincloud plot of mean cosine similarity scores across clusters for heart rate change waveforms obtained by the ATP (green) and AUT (orange) approaches. Each point corresponds to the mean similarity of a participant's HR waveform to all other participants' HR time series in a cluster (averaged across all clusters, see methods). The middle panel and right panel depict the corresponding box plots and density plots, respectively.

Application of the pre-trained autoencoder to a fear conditioning data set

Applying the pre-trained autoencoder and the pre-trained means, covariances, and weights of the BGMM (see above) to a small fear-conditioning dataset (Santos-Mayo & Moratti, 2025) resulted in a three-cluster structure based on the locations in the two-dimensional latent space of the fear-relevant CS+ related HR change patterns (Silhouette score = 0.43, Calinski-Harabasz score = 31.37; see Figure 8). Participants were assigned to clusters characterized by HR deceleration (N = 20), initial deceleration followed by a late-latency acceleration (N = 10), and a small subgroup showing immediate HR acceleration (N = 4) (see Figures 8B–8D for CS+ related HR responses of three representative participants from each cluster).

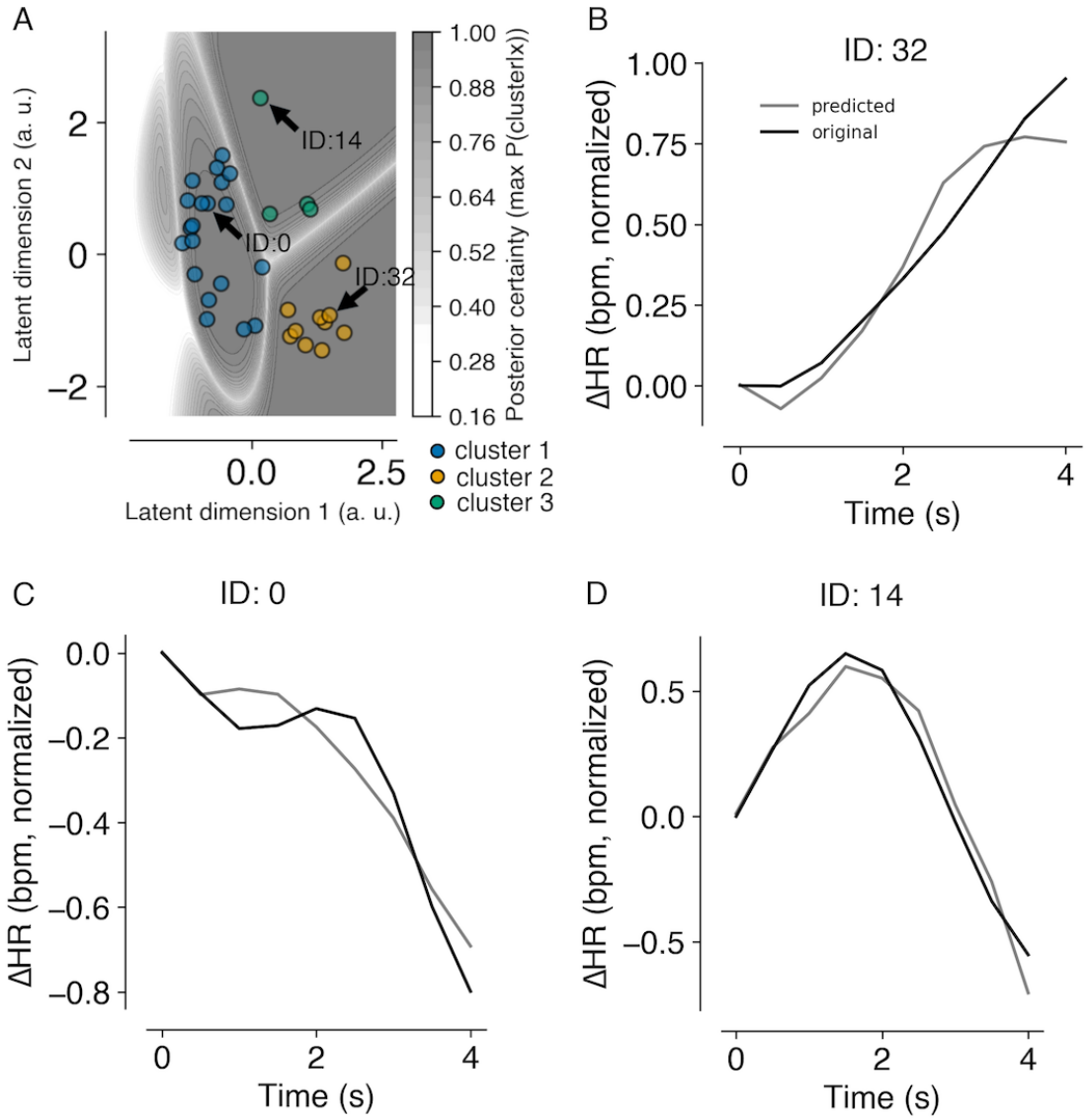

Figure 8: (A) The locations (colored dots) of the encoded HR responses in two-dimensional latent space are shown. The latent representations were clustered using the model weights of the pre-trained Bayesian Gaussian Mixture procedure trained on the previous training HR data set (N = 165). The grey heat map indicates the posterior certainty. (B), (C), and (D) show the original and predicted (by the autoencoder) HR change waveforms of three representative participants pertaining to one of the three clusters. The HR change responses were evoked by the CS+ during the learning trials. max P(cluster | x) = maximal probability of the cluster given the data.

For the HR decelerators (N = 20) and late-latency accelerators (N = 10), statistical comparisons between CS+ and CS− HR responses will be reported. However, for the immediate HR accelerators (N = 4), no statistical analyses are presented due to the very small subsample size (see Figure S2 for individual HR change responses of these four participants).

Heart rate decelerators (cluster 1) showed strong evidence of reduced HR between 1.5 s and 2 s after stimulus onset for the CS+ compared to the CS− (after 1.5 s: t(19) = -3.77, $p_{corrected}$ = 0.028; BF_{10} = 29.31, median δ = -0.76, 95% CI: [-1.28, -0.27]; after 2 s: t(19) = -3.15, $p_{corrected}$ = 0.058; BF_{10} = 8.72, median δ = -0.63, 95% CI: [-1.12, -0.16] Figure 9A). In contrast, participants belonging to cluster 2 showed strong evidence of an initial HR deceleration followed by a late-latency HR acceleration for the CS+ compared to the CS− (after 0.5 s: t(9) = -4.08, $p_{corrected}$ = 0.026; BF_{10} = 17.54, median δ = -1.09, 95% CI: [-1.98, -0.29]; after 3.5 s: t(9) = 3.91, $p_{corrected}$ = 0.026; BF_{10} = 14.27, median δ = 1.04, 95% CI: [0.26, 1.91]; after 4 s: t(9) = 4.44, $p_{corrected}$ = 0.026; BF_{10} = 27.29, median δ = 1.19, 95% CI: [0.35, 2.13], Figure 9B). Figure 9C shows the mean HR change responses to the CS+ and CS− for participants in cluster 3 (see Figure S2 for individual responses).

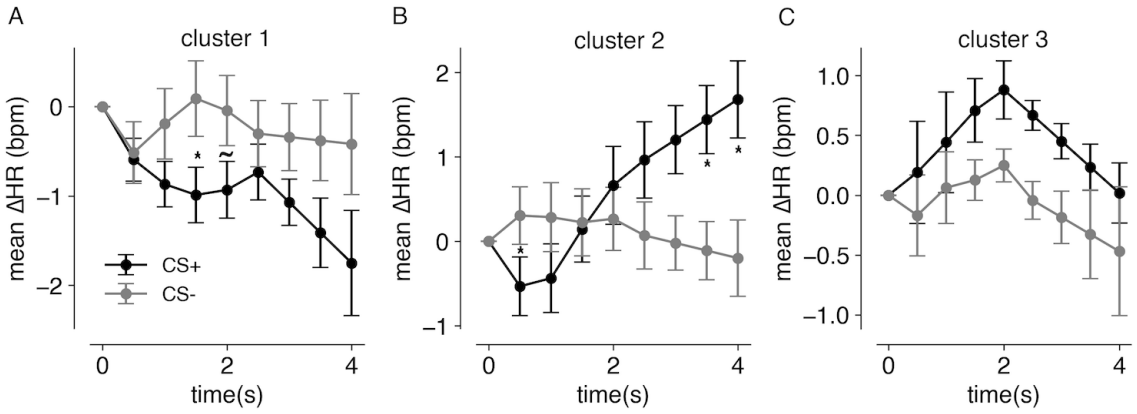

Figure 9: (A) Mean heart rate (HR) changes elicited by the CS+ and CS- for the first cluster produced by the pre-trained autoencoder are shown. (B & C) show the same information for clusters 2 and 3. Error bars represent s. e. m. * $p_{corrected}$ <= 0.05 and BF_{10} >= 3, ~ $p_{corrected}$ = 0.058 and BF_{10} >= 3.

Discussion

Here, we show that a simple variational autoencoder (VAE) can efficiently capture interindividual variability in heart rate (HR) responses to threat-related pictures by projecting HR time series into a two-dimensional latent space. A Bayesian Gaussian

Mixture Model (BGMM) clustering procedure identified three clusters: one large HR decelerator group and two HR accelerator groups. One accelerator cluster was characterized by a small HR deceleration after stimulus onset followed by delayed HR acceleration to threat images, whereas the other included participants who exhibited an immediate HR increase without any decelerative component.

Clustering in the two-dimensional latent space representation (AUT approach) learned by the VAE yielded more coherent clusters and greater within-cluster similarity of HR change waveforms across participants, compared with BGMM clustering applied directly to the HR data (ATP approach). The ATP approach produced only two clusters: one accelerator group and one large HR decelerator group that was nearly identical to that found with the AUT approach. The predominance of participants showing fear bradycardia (HR deceleration) when viewing threat-related pictures reflects a well-established statistical finding in the literature (e.g. Battaglia et al., 2024; Bradley et al., 2012 for reviews) when averaging across the whole sample without accounting for individual differences. When testing the whole sample (N = 165) without clustering, the often-reported fear bradycardia could be replicated. However, both cluster methods (AUT and ATP) resulted in a more differentiated HR response patterns to threat related visual scenes.

In our previous studies (Echegaray & Moratti, 2021; Martín et al., 2025), using a simple threshold criterion or a k-means clustering on all data points on predefined time windows, only two clusters—HR decelerators and HR accelerators—emerged, without any differentiation between immediate and delayed HR accelerators. This is consistent with applying BGMM clustering to all post-stimulus time bins in the combined sample of the three studies here (ATP approach). However, representing the HR response in a two-dimensional space by VAE learning, the AUT approach provides a more fine-grained

characterization of HR response patterns at the individual level as reflected by increased cluster performance scores and HR time series shape similarities between participants within clusters.

This resonates with findings from engineering fields such as speech recognition, biomedical imaging, and fault detection, where deep latent representation learning has been shown to substantially improve classification compared to traditional feature engineering, by capturing high-dimensional dynamics through multiple one-dimensional convolutional layers (Abdelaziz Dahou Djilali et al., 2023; Arefeen et al., 2023; Baevski et al., 2020; Hu et al., 2024; Kleesiek et al., 2021; Santos-Mayo et al., 2025; Siddiqui et al., 2025; Stephen et al., 2023; Yang & Paparrizos, 2025; Yu et al., 2023). The present study demonstrates that VAEs can uncover meaningful patterns of HR responses to threat-related pictures. Thus, clustering approaches applied to learned latent representations provide a promising strategy for disentangling heterogeneous physiological responses in psychophysiological research. Critically, representing high-dimensional dynamics in a latent space with a VAE revealed meaningful psychophysiological response patterns to threat. The long-standing assumption of a homogeneous fear bradycardia to threat, unpleasant, or fear-conditioned stimuli has already been challenged by identifying subgroups of different HR responder types (Echegaray & Moratti, 2021; Hamm & Vaitl, 1996; Klorman et al., 1977; Martín et al., 2025; Moratti et al., 2006; Moratti & Keil, 2005; Sevenster et al., 2015)

Because HR deceleration and acceleration index parasympathetic and sympathetic dominance, respectively (Graham & Clifton, 1966; Turpin & Siddle, 1978), and are associated with orienting and defensive responses (Gladwin et al., 2016; Hashemi et al., 2019; Mobbs et al., 2020; Roelofs, 2017), the response patterns identified here offer important insights. Most participants responded with parasympathetically driven

orienting (HR deceleration) to threat pictures, consistent with previous findings. This is not surprising as viewing threat pictures represent a low threat imminence situation and orienting should dominate (Mobbs et al., 2020). However, the VAE revealed smaller subgroups with distinct autonomic dynamics. One cluster exhibited a modest HR deceleration followed by a later acceleration, suggesting that initial orienting (or possibly a detection response) was followed by a defense stage, reflecting a shift from parasympathetic to sympathetic dominance. Another group showed immediate HR acceleration, indicating the absence of an orienting strategy and the initiation of a rapid defense response.

The present methodological report also demonstrates that a pre-trained autoencoder, together with a pre-trained Bayesian Gaussian Mixture Model (BGMM), can be used to identify subgroups of participants exhibiting distinct HR response patterns to learned CS+ fear relevance, even in a relatively small sample. In the original study (Santos-Mayo & Moratti, 2025), the application of *k*-means clustering within a predefined time window yielded one predominant group characterized by orienting responses, indexed by fear bradycardia to the CS+, and a smaller subgroup showing a defensive response associated with HR acceleration. In contrast, applying the pre-trained VAE and BGMM—trained on a larger dataset—allowed the identification of three distinct patterns: HR deceleration, HR deceleration followed by acceleration, and immediate HR acceleration.

These patterns closely resemble the HR response types obtained during VAE training on complex visual threat scenes. Thus, the pre-trained VAE not only extracted more meaningful HR response profiles associated with orienting, orienting followed by defense, and immediate defense, but also generalized successfully from complex visual scenes to simple visual stimuli (sine patches) that had acquired fear relevance. However,

the cluster of participants showing immediate HR acceleration included only four individuals. Two of these participants displayed a clear early HR acceleration, whereas one showed a brief HR deceleration preceding the accelerative phase, and the fourth exhibited no HR modulation in response to the CS+. Consistent with this observation, the two participants with less clear early accelerative responses were located near the border of cluster 3. Although small sample sizes inherently produce greater variability in response patterns, examining the HR response waveforms individually—together with their corresponding locations in the latent space of a pretrained autoencoder—allows for meaningful conclusions about the underlying HR response type.

A possible concern is that the emergence of similar HR response types—like those found in the test data using simple fear conditioned visual cues—might reflect an artifact of the pretrained latent space based on threat pictures. That is, the VAE could be forcing the fear-conditioning HR responses to cluster in this way. However, the VAE is optimized not only to minimize reconstruction error but also to regularize its latent distribution toward a smooth Gaussian prior, which promotes continuity rather than discrete clustering. Thus, any cluster structure that emerges reflects genuine organization in the data, not an architectural bias of the VAE. In addition, the pretrained BGMM does not modify the latent representations; it simply identifies structure within their existing configuration. Consequently, the cluster assignments are determined by the data-driven latent representations rather than imposed upon them. Finally, the mean HR change responses analyzed after clustering are derived from the original HR time series, not from the reconstructed waveforms. In summary, the identification of three HR responder groups to the learned fear-relevant CS+—paralleling those observed for complex threat scenes—appears to reflect intrinsic structure in the fear-learning data rather than a modeling artifact.

However, the present results must be considered in the light of a relatively small sample size than normally used in machine learning. The primary goal of this methodological work was to demonstrate that VAEs can be a useful tool in psychophysiological research. By making all data and code fully accessible, we hope to encourage a broader community of psychophysiology researchers to apply deep latent representation learning and extend these findings to much larger samples.

Data availability statement: All data and code can be downloaded at: https://osf.io/k58uy/

Acknowledgements: This work was funded by the Ministerio de Ciencia, Innovación y Universidades of the Spanish government and the Agencia Española de Investigación (AEI) under the grant number: PID2021-126074NB-I00.

Usage of AI: For the manuscript ChatGPT was used to correct the English and assisted in developing the Python code for the analysis.

References

Abdelaziz Dahou Djilali, Y., Narayan, S., Boussaid, H., Almazrouei, E., & Debbah, M. (2023). Lip2Vec: Efficient and Robust Visual Speech Recognition via Latent-to-Latent Visual to Audio Representation Mapping. *arXiv E-Prints*, arXiv:2308.06112. https://doi.org/10.48550/arXiv.2308.06112

Arefeen, Y., Xu, J., Zhang, M., Dong, Z., Wang, F., White, J., Bilgic, B., & Adalsteinsson, E. (2023). Latent signal models: Learning compact representations of signal evolution for improved time-resolved, multi-contrast MRI. *Magnetic Resonance in Medicine*, *90*(2), 483–501. https://doi.org/10.1002/mrm.29657

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (No. arXiv:2006.11477). arXiv. https://doi.org/10.48550/arXiv.2006.11477

Battaglia, S., Nazzi, C., Lonsdorf, T. B., & Thayer, J. F. (2024). Neuropsychobiology of fear-induced bradycardia in humans: Progress and pitfalls. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-024-02600-x

Ben, R. D. (2019). *Luminance control of colorful images*. https://doi.org/10.17605/OSF.IO/AUZJY

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st ed.). Springer. https://link.springer.com/book/9780387310732

Bradley, M. M. (2009). Natural selective attention: Orienting and emotion. *Psychophysiology*, *46*(1), 1–11. https://doi.org/10.1111/j.1469-8986.2008.00702.x

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion*, *1*(3), 276–298. https://doi.org/10.1037/1528-3542.1.3.276

Bradley, M. M., Keil, A., & Lang, P. J. (2012). Orienting and Emotional Perception: Facilitation, Attenuation, and Interference. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00493

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Calvo Garcia, S. F., Díaz Sánchez, M., Molina Blanco, S., Pampín del Río, S., Sánchez del Coral, G., Moratti, S. (in preparation). Relax and orient! Progressive muscle relaxation biases towards orienting and reduced defense responses during threat processing.

Echegaray, J., & Moratti, S. (2021). Threat imminence modulates neural gain in attention and motor relevant brain circuits in humans. *Psychophysiology*, *58*(8). https://doi.org/10.1111/psyp.13849

Gladwin, T. E., Hashemi, M. M., Van Ast, V., & Roelofs, K. (2016). Ready and waiting: Freezing as active action preparation under threat. *Neuroscience Letters*, *619*, 182–188. https://doi.org/10.1016/j.neulet.2016.03.027

Graham, F. K., & Clifton, R. K. (1966). Heart-rate change as a component of the orienting response. *Psychological Bulletin*, *65*(5), 305–320. https://doi.org/10.1037/h0023258

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology*, *33*(6), 698–710. https://doi.org/10.1111/j.1469-8986.1996.tb02366.x

Hashemi, M. M., Gladwin, T. E., de Valk, N. M., Zhang, W., Kaldewaij, R., van Ast, V., Koch, S. B. J., Klumpers, F., & Roelofs, K. (2019). Neural Dynamics of Shooting Decisions and the Switch from Freeze to Fight. *Scientific Reports*, *9*(1), 4240. https://doi.org/10.1038/s41598-019-40917-8

Hodes, R. L., Cook III, E. W., & Lang, P. J. (1985). Individual Differences in Autonomic Response: Conditioned Association or Conditioned Fear? *Psychophysiology*, *22*(5), 545–560. https://doi.org/10.1111/j.1469-8986.1985.tb01649.x

Hu, H.-X., Cao, C., Hu, Q., Zhang, Y., & Lin, Z.-Z. (2024). A Real-Time Bearing Fault Diagnosis Model Based on Siamese Convolutional Autoencoder in Industrial

Internet of Things. *IEEE Internet of Things Journal*, *11*(3), 3820–3831. https://doi.org/10.1109/JIOT.2023.3307127

Kleesiek, J., Kersjes, B., Ueltzhöffer, K., Murray, J. M., Rother, C., Köthe, U., & Schlemmer, H.-P. (2021). Discovering Digital Tumor Signatures—Using Latent Code Representations to Manipulate and Classify Liver Lesions. *Cancers*, *13*(13), 3108. https://doi.org/10.3390/cancers13133108

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox 3? *Perception, 36*(14), 1–16.

Klorman, R., Weissberg, R. P., & Wiesenfeld, A. R. (1977). Individual Differences in Fear and Autonomic Reactions to Affective Stimulation. *Psychophysiology*, *14*(1), 45–51. https://doi.org/10.1111/j.1469-8986.1977.tb01154.x

Lacey, J. I., & Lacey, B. C. (1970). Some autonomic-central nervous system interrelationships. In *Physiological Correlates of Emotion*. Academmic Press.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-6*. University of Florida.

Lojowska, M., Gladwin, T. E., Hermans, E. J., & Roelofs, K. (2015). Freezing promotes perception of coarse visual features. *Journal of Experimental Psychology: General*, *144*(6), 1080–1088. https://doi.org/10.1037/xge0000117

Löw, A., Weymar, M., & Hamm, A. O. (2015). When Threat Is Near, Get Out of Here: Dynamics of Defensive Behavior During Freezing and Active Avoidance. *Psychological Science*, *26*(11), 1706–1716. https://doi.org/10.1177/0956797615597332

Martín, C. G., Blanco, S. M., Sánchez, M. D., García, S. F. C., del Corral, G. S., del Río, S. P., & Moratti, S. (2025). How Orienting and Defence Drives Oscillatory

Responses in Human Visual and Motor Cortical Circuits During Viewing of Threat Pictures: Evidence From ssVEPs and Beta-Band Desynchronization. *European Journal of Neuroscience*, *61*(12), e70157. https://doi.org/10.1111/ejn.70157

Mathworks/Anomaly-detection-using-Variational-Autoencoder-VAE-. (2024). [Computer software]. MathWorks. https://github.com/mathworks/Anomaly-detection-using-Variational-Autoencoder-VAE- (Original work published 2020)

Mobbs, D., Headley, D. B., Ding, W., & Dayan, P. (2020). Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends in Cognitive Sciences*, *24*(3), 228–241. https://doi.org/10.1016/j.tics.2019.12.016

Moeyersons, J., Amoni, M., Van Huffel, S., Willems, R., & Varon, C. (2019). R-DECO: An open-source Matlab based graphical user interface for the detection and correction of R-peaks. *PeerJ. Computer Science*, *5*, e226. https://doi.org/10.7717/peerj-cs.226

Moratti, S., & Keil, A. (2005). Cortical activation during Pavlovian fear conditioning depends on heart rate response patterns: An MEG study. *Cognitive Brain Research*, *25*(2), 459–471. https://doi.org/10.1016/j.cogbrainres.2005.07.006

Moratti, S., Keil, A., & Miller, G. A. (2006). Fear but not awareness predicts enhanced sensory processing in fear conditioning. *Psychophysiology*, *43*(2), 216–226. https://doi.org/10.1111/j.1464-8986.2006.00386.x

Moratti, S., Keil, A., & Stolarova, M. (2004). Motivated attention in emotional picture processing is reflected by activity modulation in cortical attention networks. *NeuroImage*, *21*(3), 954–964. https://doi.org/10.1016/j.neuroimage.2003.10.030

Pavlov, I. P. (1927). *Conditioned reflexes*. Dover.

Perakakis, P., Joffily, M., Taylor, M., Guerra, P., & Vila, J. (2010). KARDIA: A Matlab software for the analysis of cardiac interbeat intervals. *Computer Methods and Programs in Biomedicine*, *98*(1), 83–89. https://doi.org/10.1016/j.cmpb.2009.10.002

Reyes del Paso, G. A., & Vila, J. (1998). The continuing problem of incorrect heart rate estimation in psychophysiological studies: An off-line solution for cardiotachometer users. *Biological Psychology*, *48*(3), 269–279. https://doi.org/10.1016/S0301-0511(98)00039-8

Roelofs, K. (2017). Freeze for action: Neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1718), 20160206. https://doi.org/10.1098/rstb.2016.0206

Roelofs, K., Hagenaars, M. A., & Stins, J. (2010). Facing Freeze: Social Threat Induces Bodily Freeze in Humans. *Psychological Science*, *21*(11), 1575–1581. https://doi.org/10.1177/0956797610384746

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sevenster, D., Hamm, A., Beckers, T., & Kindt, M. (2015). Heart rate pattern and resting heart rate variability mediate individual differences in contextual anxiety and conditioned responses. *International Journal of Psychophysiology*, *98*(3), 567–576. https://doi.org/10.1016/j.ijpsycho.2015.09.004

Santos-Mayo, A., Gilbert, F., Mirifar, A., Tebbe, A. L., Fang, R., Ding, M., Keil, A. (2025). Concept2Brain: An AI model for predicting subject-level neurophysiological responses to text and pictures. *bioRxiv, https://www.biorxiv.org/content/10.1101/2025.08.04.668476v1*

Santos-Mayo, A., Moratti S. (2025). How fear conditioning affects the visuocortical processing of context cues in humans. Evidence from steady state visual evoked responses. *Cortex, 183,* 21 -37. https://doi.org/10.1016/j.cortex.2024.11.005

Siddiqui, A. A., Tirunagari, S., Zia, T., & Windridge, D. (2025). A latent diffusion approach to visual attribution in medical imaging. *Scientific Reports*, *15*, 962. https://doi.org/10.1038/s41598-024-81646-x

Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Pergamon.

Stephen, E. P., Li, Y., Metzger, S., Oganian, Y., & Chang, E. F. (2023). Latent neural dynamics encode temporal context in speech. *Hearing Research*, *437*, 108838. https://doi.org/10.1016/j.heares.2023.108838

Turpin, G., & Siddle, D. A. T. (1978). Cardiac and forearm plethysmographic responses to high intensity auditory stimulation. *Biological Psychology*, *6*(4), 267–281. https://doi.org/10.1016/0301-0511(78)90029-7

Yang, F., & Paparrizos, J. (2025). SPARTAN: Data-Adaptive Symbolic Time-Series Approximation. *Proc. ACM Manag. Data*, *3*(3), 220:1-220:30. https://doi.org/10.1145/3725357

Yu, T., Li, S., & Lu, J. (2023). Quantum stacked autoencoder fault diagnosis model for bearing faults. *Insight - Non-Destructive Testing and Condition Monitoring*, *65*(11), 631–638. https://doi.org/10.1784/insi.2023.65.11.631

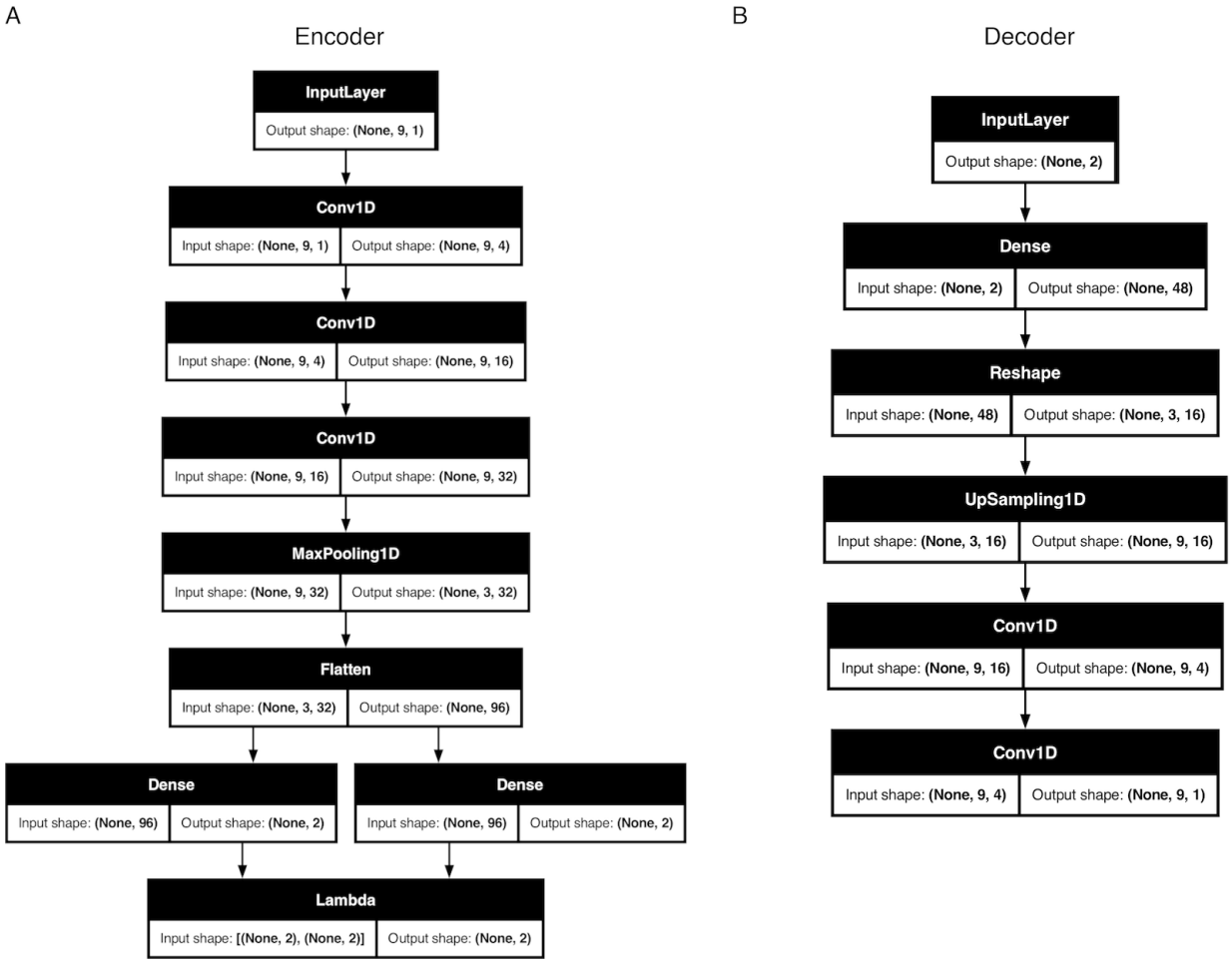

Figure S1: The layers of the autoencoder are shown. (A) The encoder layers are depicted. The two dense layers map the mean and the variance of the Gaussian distribution. (B) The decoder layers are shown.

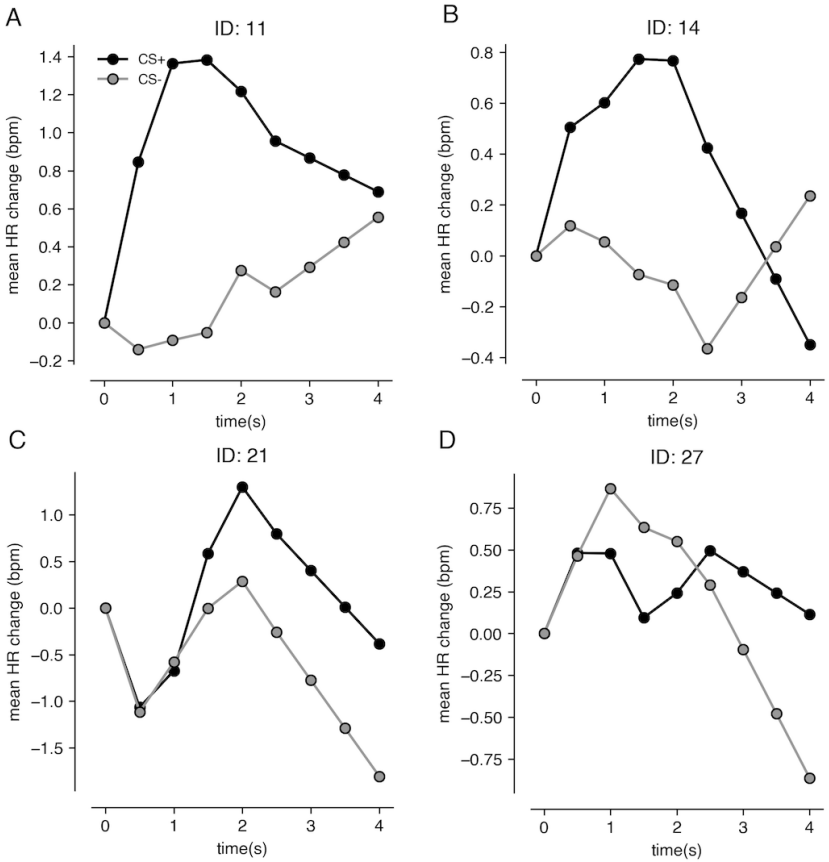

Figure S2: (A-D) Heart rate change responses to the CS+ and CS- during the acquisition phase for 4 individual participants of cluster 3 are shown. ID = identity number.