A APPENDIX

A.1 EXPERIMENTAL METHODS

A.1.1 PARTICIPANTS

Both the participant and a legal guardian (where applicable) signed an informed consent form approved by their local university Institutional Review Board (IRB) and all procedures were conducted in line with the Declaration of Helsinki. Presentation ®(Neurobehavioural Systems Inc, www.neurobs.com) was used for stimulus presentation.

Site	AgeGroup	Gender	Diagnosis	Axis 1	Depression
Site 1	adult	F	0.0	35	35
			1.0	59	59
		M	0.0	11	11
			1.0	12	12
Site 2	child	F	0.0	72	95
			1.0	31	8
		M	0.0	101	119
			1.0	19	1
Site 3	child	F	0.0	82	128
			1.0	57	11
		M	0.0	98	150
			1.0	54	2
Site 4	child	F	0.0	41	58
			1.0	21	4
		M	0.0	45	63
			1.0	20	2

Table A1: Participant label distribution

A.1.2 EEG PREPROCESSING

EEG was recorded from a 34-channel ActiChamp system (BioSemi, Amsterdam, Netherlands), positioned according to the 10/20 system. The data was recorded at 250 Hz and referenced to electrode Cz. To monitor eye blinks, VEOG and HEOG electrodes were applied above, below and to the outer canthi of both eyes. The raw EEG data was digitized at a 24-bit resolution using a low-pass 5th order sinc filter with a half-power cutoff at 102.4 Hz.

Offline, the continuous EEG data were re-referenced to the mastoids and bandpass filtered using a 2nd order Butterworth high-pass filter with a 3 dB cut-off at 0.1 Hz and a 12th order Butterworth low-pass filter with a 3 dB cut-off at 30 Hz. Eye-movements were detected and removed from the continuous data using ICA (Gratton et al., 1983). In this process, the first ICA component was always targeted as the to-be-corrected blink artifact. We also examined the pipeline without ICA and found the general pattern of results to hold (see Table A6). EEG from three electrodes were chosen to be included in the analysis—Fz, Cz, Pz—because (1) the LPP is traditionally indexed from centro-parietal midline electrodes and (2) it was desirable for the project to test an EEG system that maximizes speed of application and ease of use. The full preprocessing pipeline was built with the MNE (Gramfort et al., 2014) software package in Python.

The final step in the pre-processing pipeline was to exclude anomalous trials (EEG segments) from each participants data. In order of application, a raw maximum amplitude threshold was applied at $\pm 0.005\,\mathrm{V}$, the data were normalized per-participant to zero mean and unit variance and a standard deviation threshold was applied at $|\sigma| >= 5$. Finally, standard deviation "transients" (i.e., deflections in the EEG time series between t and $t+1$ where $|\sigma| >= 3$) were identified and the associated trials were removed.Overall 758 trajectories were used to train the models.

A.2 MODELS

A.2.1 VARIATIONAL AUTOENCODER

A Variational Autoencoder (VAE) (Rezende et al., 2014; Kingma & Welling, 2014) is a generative model that aims to learn a joint distribution $p(\mathbf{x},\mathbf{z})$ of input data $\mathbf{x}$ and a set of latent variables $\mathbf{z}$ by

learning to maximize the Evidence Lower BOund (ELBO) on the data distribution. The neural network implementation consists of an inference network (equivalent to the AE encoder), that takes inputs $\mathbf{x}$ and parameterizes the posterior distribution $q(\mathbf{z}|\mathbf{x})$, and a generative network (equivalent to the AE decoder) that takes a sample from the inferred posterior distribution $\hat{\mathbf{z}} \sim \mathcal{N}(\mu(\mathbf{z}|\mathbf{x}),\sigma(\mathbf{z}|\mathbf{x}))$ and attempts to reconstruct the original image. The model is trained through a two-part loss objective:

$$\mathcal{L}_{VAE} = \mathbb{E}_{p(\mathbf{x})}[\, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}))\,]$$

where $p(\mathbf{x})$ is the probability of the input data, $q(\mathbf{z}|\mathbf{x})$ is the learnt posterior over the latent units given the data, $p(\mathbf{z})$ is the unit Gaussian prior with a diagonal covariance matrix $\mathcal{N}(\mathbf{0},\mathbb{I})$, and ϕ and θ are the parameters of the inference (encoder) and generative (decoder) networks respectively. The encoder parametrised a Gaussian distribution with a diagonal covariance matrix, while the decoder parametrised a Bernoulli distribution.

A.3 CLASSIFIERS

In each case, we extracted representations for the full dataset, and used them as input into the baseline classifiers, using the sklearn package implementation with default parameters, unless specified otherwise. In particular, given that the label data was often significantly unbalanced, we used the "balanced" class weight option where appropriate, which adjusted the weight for each dataset sample to be inversely proportional to its corresponding class frequency according to $w_i = N/n_y$, where w_i is the weight for sample i, n_y is the number of samples in the dataset with the same class label as y_i, and N is the total number of datapoints. We used 5-fold cross-validation to choose the best hyperparameters to avoid overfitting where appropriate. For LR we used 500 maximum iterations, and both L1 and L2 regularisation. For RF we used depth 2 (larger values did not improve performance).

A.4 UNSUPERVISED DISENTANGLEMENT RANKING

In order to quantify the quality of disentanglement achieved by the trained β-VAE models, we applied the recently proposed Unsupervised Disentanglement Ranking (UDR) score (Duan et al., 2019). UDR is the only known method for measuring the quality of disentanglement in VAEs without access to the ground truth attribute labels, which were not available for the EEG data. UDR works by performing pairwise comparisons between the representations learned by models trained using the same hyperparameter setting but with different seeds. For each trained β-VAE model we performed nine pairwise comparisons with all the other models trained with the same β value and calculated the corresponding UDR_{ij} score, where i and j index the two β-VAE models. Each UDR_{ij} score is calculated by computing the similarity matrix R_{ij}, where each entry is the Spearman correlation between the responses of individual latent units of the two models across the same data. The absolute value of the similarity matrix is then taken $|R_{ij}|$ and the final score for each pair of models is calculated according to:

$$\frac{1}{d_a + d_b}\left[\sum_b \frac{r_a^2 * I_{KL}(b)}{\sum_a R(a,b)} + \sum_a \frac{r_b^2 * I_{KL}(a)}{\sum_b R(a,b)}\right]$$

where a and b index into the latent units of models i and j respectively, $r_a = \max_a R(a,b)$ and $r_b = \max_b R(a,b)$. I_{KL} indicate the "informative" latent units within each model, operationalised as units with $KL > 0.01$ from the unit Gaussian prior, and d is the number of such latent units. The final score for model i is calculated by taking the median of UDR_{ij} across all j.

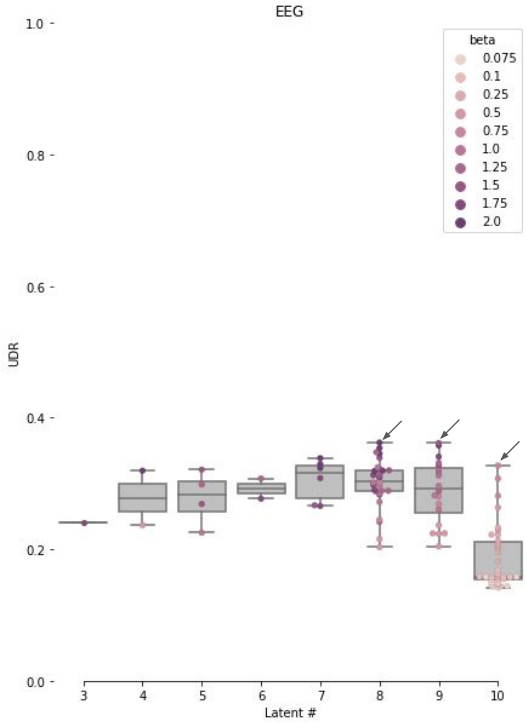

Figure A1: Distribution of UDR scores against the final number of "informative" latents across the hyperparameter search for β-VAE. A latent is "informative" if the distance of its posterior against the unit Gaussian prior is high $KL > 0.01$, otherwise the latent carries no information and is effectively switched off. Three models whose latent traversals are plotted in Figs. A8-A10 are indicated by arrows.

	Train ↓	Test →	Age		Gender		Site		Depression		Axis 1		Overall	
			ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL
Chance			0.50		0.50		0.25		0.50		0.50		0.45	
LR	ERP	LPP	0.74	**0.56**	0.52	0.51	0.38	**0.30**	**0.64**	**0.56**	0.52	0.52	0.56	**0.49**
LR	SMPL	LPP	0.69	0.54	0.52	**0.52**	0.36	**0.30**	0.61	0.54	0.52	0.52	0.54	**0.48**
RF	ERP	LPP	**0.80**	0.54	**0.59**	**0.53**	0.34	0.21	**0.64**	0.47	**0.63**	0.53	**0.60**	0.45
RF	SMPL	LPP	0.46	0.46	0.52	**0.52**	0.24	0.20	0.47	0.47	0.54	0.41	0.44	0.41
LDA	ERP	LPP	0.78	0.54	0.51	0.48	0.32	0.28	0.62	0.50	0.52	0.54	0.55	0.47
LDA	SMPL	LPP	0.46	0.46	0.51	0.51	0.34	0.27	0.47	0.47	0.58	**0.66**	0.47	0.47
SVM	ERP	LPP	0.76	0.54	**0.58**	0.51	**0.41**	0.27	**0.65**	**0.55**	0.57	0.51	**0.59**	**0.48**
SVM	SMPL	LPP	0.53	**0.55**	0.52	**0.53**	0.28	0.27	0.61	0.52	0.52	0.49	0.49	0.47

Table A2: Balanced classification accuracy calculated as the average of precision and recall. ERP - averaged trajectories, SMPL - single sampled EEG trajectories. LPP - canonical late positive potential baseline representation. LR - L2 regularised logistic regression, RF - random forest, LDA - linear discriminant analysis, SVM - support vector machine. Models trained on single EEG trajectories (SMPL) are tested on different single EEG trajectories in the SMPL/SMPL train/test case. Average balanced classification accuracy across all conditions: LR: 0.52, RF: 0.48, LDA: 0.49, SVM: 0.51

	Train ↓	Test →	Age		Gender		Site		Depression		Axis 1		Overall	
			ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL
Chance			0.50		0.50		0.25		0.50		0.50		0.45	
LR	ERP	β-VAE	0.85	0.62	0.66	0.53	0.51	0.34	0.68	0.59	0.60	0.55	0.66	**0.53**
LR	SMPL	β-VAE	0.76	0.65	0.58	0.53	0.40	**0.36**	0.67	0.58	0.56	0.56	0.59	**0.54**
RF	ERP	β-VAE	0.75	**0.56**	0.75	0.56	0.45	0.29	0.47	0.47	0.66	0.54	0.62	0.48
RF	SMPL	β-VAE	0.78	**0.72**	0.65	0.55	0.43	0.29	0.70	0.47	0.65	**0.66**	0.64	**0.54**
LDA	ERP	β-VAE	0.86	0.63	0.66	0.54	0.46	0.31	0.68	**0.63**	0.58	0.54	0.65	**0.53**
LDA	SMPL	β-VAE	0.82	0.61	0.63	0.54	0.50	0.31	0.69	0.52	0.56	0.53	0.64	0.50
SVM	ERP	β-VAE	**0.93**	0.61	**0.78**	0.54	**0.71**	0.33	**0.83**	0.59	**0.74**	0.57	**0.80**	**0.53**
SVM	SMPL	β-VAE	0.88	0.64	0.71	**0.56**	0.60	0.34	0.80	0.57	0.69	0.53	0.74	**0.53**

Table A3: Balanced classification accuracy calculated as the average of precision and recall. ERP - averaged trajectories, SMPL - single sampled EEG trajectories. LPP - canonical late positive potential baseline representation. LR - L2 regularised logistic regression, RF - random forest, LDA - linear discriminant analysis, SVM - support vector machine. Models trained on single EEG trajectories (SMPL) are tested on different single EEG trajectories in the SMPL/SMPL train/test case. Average balanced classification accuracy across all conditions: LR: 0.58, RF: 0.57, LDA: 0.58, SVM: 0.65

	Train ↓	Test →	Age		Gender		Site		Depression		Axis 1		Overall	
			ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL
Chance			0.50		0.50		0.25		0.50		0.50		0.45	
LR	ERP	AE	0.85	0.62	0.68	**0.57**	0.48	**0.37**	0.69	0.59	0.60	**0.56**	0.66	**0.54**
LR	SMPL	AE	0.76	**0.65**	0.63	**0.57**	0.46	**0.36**	0.66	0.58	0.59	0.54	0.62	**0.54**
RF	ERP	AE	0.72	0.46	0.71	**0.57**	0.44	0.29	0.47	0.47	0.66	0.41	0.60	0.44
RF	SMPL	AE	0.77	**0.64**	0.68	**0.57**	0.45	0.32	0.74	0.47	0.62	0.54	0.65	0.51
LDA	ERP	AE	0.87	0.63	0.67	**0.58**	0.46	0.31	0.67	**0.70**	0.59	0.50	0.65	**0.54**
LDA	SMPL	AE	0.86	0.61	0.67	0.56	0.46	0.33	0.70	0.62	0.58	0.54	0.65	**0.53**
SVM	ERP	AE	**0.92**	0.63	**0.80**	0.56	**0.69**	**0.36**	**0.82**	0.60	**0.76**	**0.55**	**0.80**	**0.54**
SVM	SMPL	AE	0.89	**0.65**	0.75	**0.57**	0.60	**0.36**	0.80	0.58	0.69	0.54	0.75	**0.54**

Table A4: Balanced classification accuracy calculated as the average of precision and recall. ERP - averaged trajectories, SMPL - single sampled EEG trajectories. LPP - canonical late positive potential baseline representation. LR - L2 regularised logistic regression, RF - random forest, LDA - linear discriminant analysis, SVM - support vector machine. Models trained on single EEG trajectories (SMPL) are tested on different single EEG trajectories in the SMPL/SMPL train/test case. Average balanced classification accuracy across all conditions: LR: 0.59, RF: 0.55, LDA: 0.60, SVM: 0.66

	Train ↓	Test →	Age		Gender		Site		Depression		Axis 1		Overall	
			ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL
Chance			0.50		0.50		0.25		0.50		0.50		0.45	
LR	ERP	LPP	0.74	0.56	0.52	0.51	0.38	0.30	0.64	0.56	0.52	0.52	0.56	0.49
LR	SMPL	LPP	0.69	0.54	0.52	0.52	0.36	0.30	0.61	0.54	0.52	0.52	0.54	0.48
LR	ERP	β-VAE	**0.85**	0.62	0.66	0.53	**0.51**	0.34	**0.68**	**0.59**	**0.60**	**0.55**	**0.66**	**0.53**
		AE	**0.85**	0.62	**0.68**	**0.57**	0.48	**0.37**	**0.69**	**0.59**	**0.60**	**0.56**	**0.66**	**0.54**
LR	SMPL	β-VAE	0.76	**0.65**	0.58	0.53	0.40	**0.36**	0.67	**0.58**	0.56	**0.56**	0.59	**0.54**
		AE	0.76	**0.65**	0.63	**0.57**	0.46	**0.36**	0.66	**0.58**	**0.59**	0.54	0.62	**0.54**
SCAN	ERP	β-VAE	0.78	**0.64**	0.58	0.50	0.39	0.30	**0.68**	**0.59**	0.54	0.52	0.59	0.51
		AE	0.48	0.50	0.48	0.50	0.21	0.25	0.50	0.49	0.50	0.50	0.43	0.45
SCAN	SMPL	β-VAE	0.73	0.59	0.57	0.53	0.38	0.30	0.67	0.56	0.55	0.50	0.58	0.5
		AE	0.41	0.45	0.43	0.46	0.16	0.20	0.46	0.49	0.53	0.49	0.40	0.42

Table A5: Balanced classification accuracy calculated as the average of precision and recall $Acc = \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)/2$. ERP - averaged trajectories, SMPL - single sampled EEG trials. LPP - canonical late positive potential baseline representation. LR - L2 regularised logistic regression (see Tbl. A2 for other baseline classifiers). Models trained on single EEG trajectories (SMPL) are tested on different single EEG trajectories in the SMPL/SMPL train/test case.

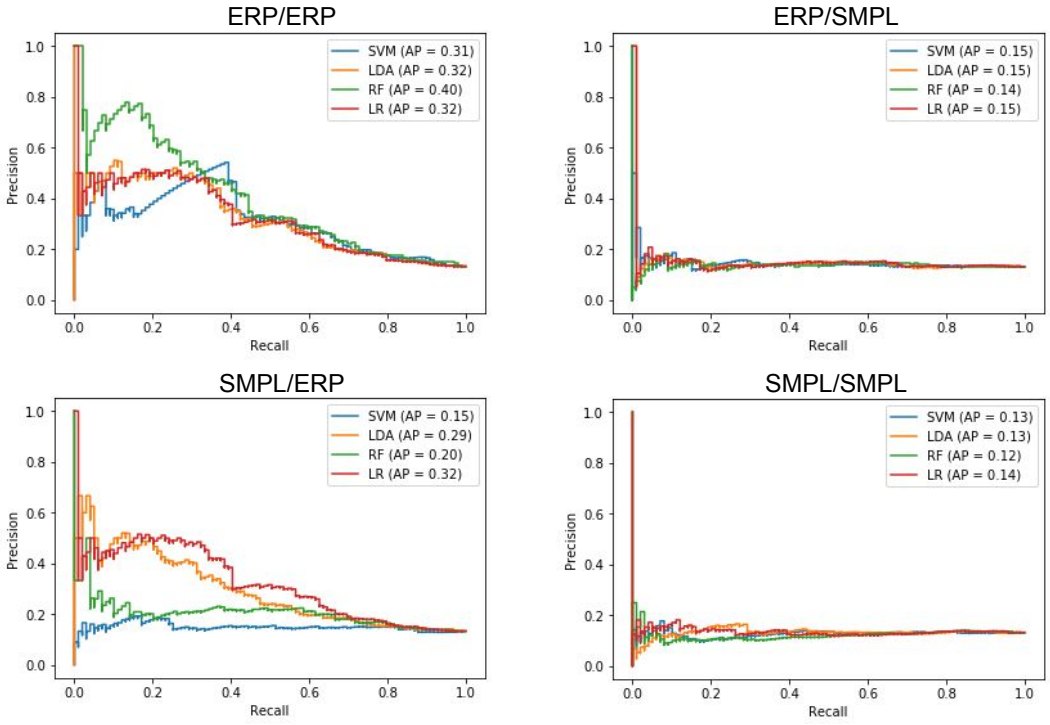

Figure A2: Precision-recall curves for balanced classification accuracy of depression diagnosis calculated as the average of precision and recall for canonical late positive potential (LPP) baseline representation. LR - L2 regularised logistic regression, RF - random forest, LDA - linear discriminant analysis, SVM - support vector machine.

	Train ↓	Test →	Age		Gender		Site		Depression		Axis 1		Overall	
			ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL
Chance			0.50		0.50		0.25		0.50		0.50		0.45	
LR	ERP	LPP	-0.03	-0.01	0.00	0.00	-0.03	0.00	-0.01	-0.06	0.00	-0.02	-0.01	-0.02
LR	SMPL	LPP	-0.02	0.00	0.00	-0.02	-0.02	-0.03	-0.01	-0.03	0.01	0.01	-0.01	-0.01
LR	ERP	β-VAE	-0.01	**0.03**	-0.02	**0.04**	-0.03	0.00	**0.03**	**0.02**	**0.03**	0.00	0.00	**0.02**
		AE	0.00	-0.01	-0.01	-0.01	**0.02**	-0.03	0.00	-0.04	**0.01**	-0.02	**0.01**	-0.02
LR	SMPL	β-VAE	**0.08**	-0.03	0.00	**0.01**	**0.08**	-0.02	**0.02**	**0.02**	**0.03**	-0.01	**0.04**	-0.01
		AE	**0.10**	-0.02	**0.01**	-0.02	**0.06**	-0.03	**0.03**	**0.01**	**0.01**	0.00	**0.04**	-0.01
SCAN	ERP	β-VAE	**0.05**	**0.03**	**0.01**	**0.01**	-0.04	**0.02**	-0.01	**0.01**	0.00	-0.05	0.00	0.00
		AE	-0.01	-0.04	**0.11**	**0.03**	**0.03**	-0.01	-0.03	-0.02	-0.01	-0.01	**0.02**	-0.01
SCAN	SMPL	β-VAE	**0.02**	**0.02**	-0.04	-0.04	-0.04	-0.03	**0.02**	**0.02**	0.00	**0.04**	-0.01	0.00
		AE	-0.12	-0.08	**0.11**	**0.04**	**0.01**	**0.02**	-0.09	-0.06	-0.10	-0.03	-0.04	-0.02

Table A6: Difference between balanced classification accuracy obtained from data without ICA-based artifact removal pre-processing and with such pre-processing (the latter shown in Table A5). Positive numbers (highlighted in bold) indicate that classification accuracy on data without ICA pre-processing is better than that from data with ICA pre-processing. ERP - averaged trajectories, SMPL - single sampled EEG trials. LPP - canonical late positive potential baseline representation. LR - L2 regularised logistic regression.

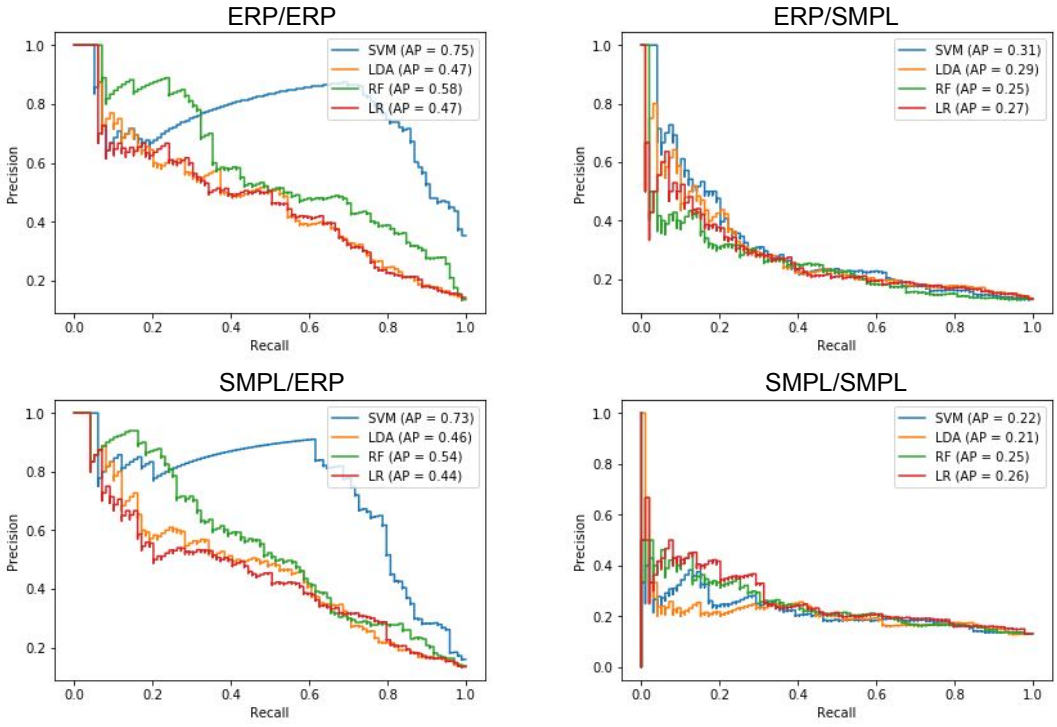

Figure A3: Precision-recall curves for balanced classification accuracy of depression diagnosis calculated as the average of precision and recall for β-VAE. LR - L2 regularised logistic regression, RF - random forest, LDA - linear discriminant analysis, SVM - support vector machine.

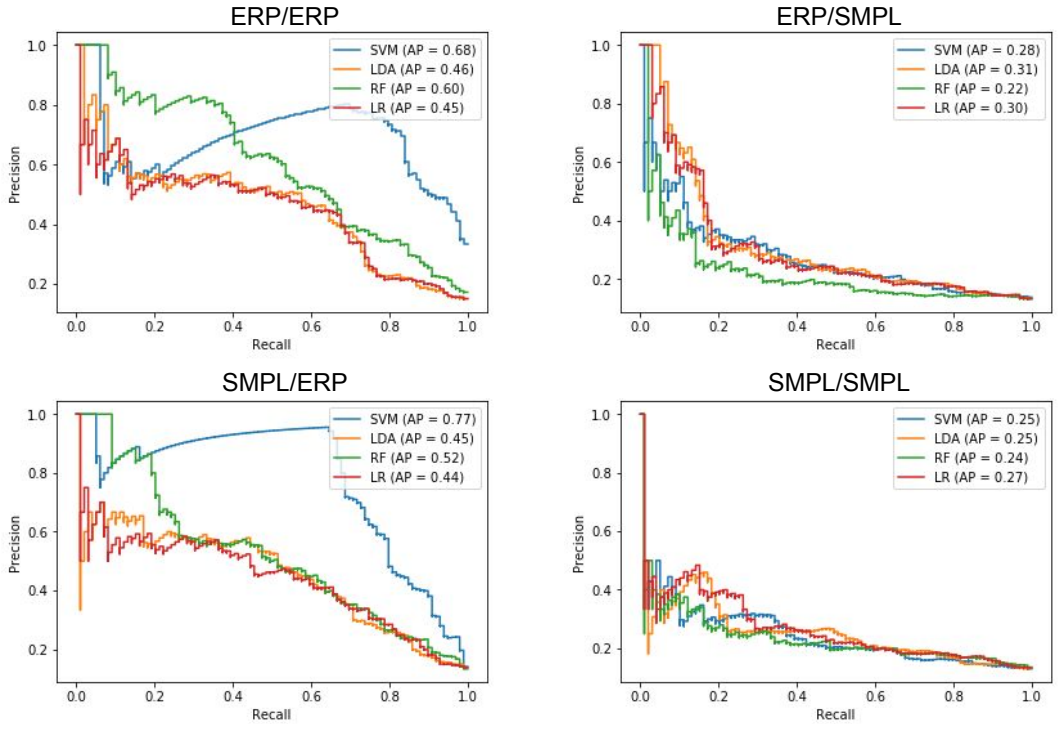

Figure A4: Precision-recall curves for balanced classification accuracy of depression diagnosis calculated as the average of precision and recall for AE. LR - L2 regularised logistic regression, RF - random forest, LDA - linear discriminant analysis, SVM - support vector machine.

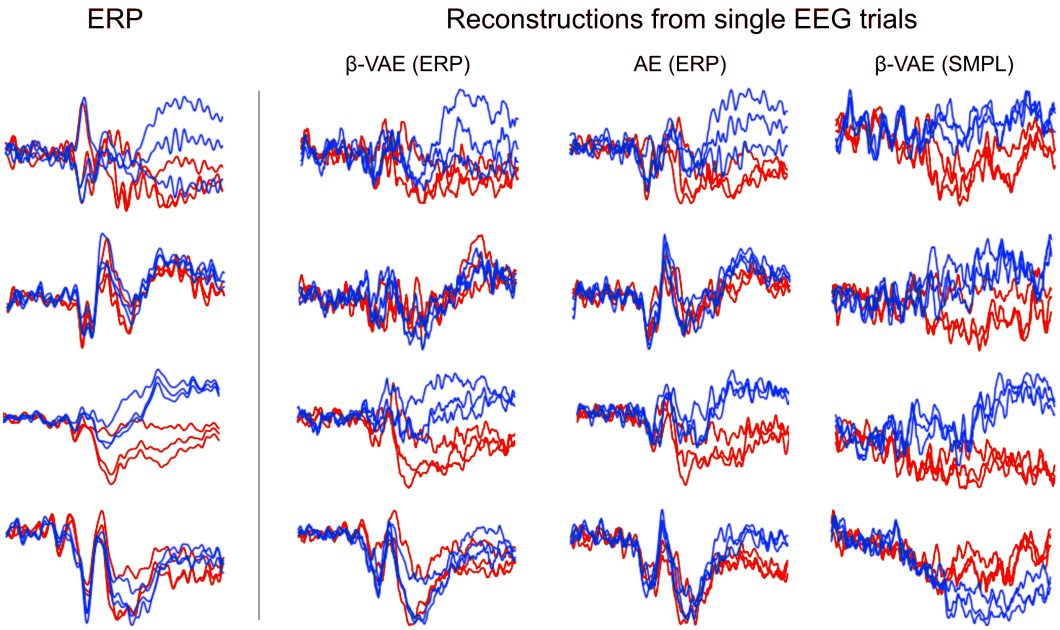

Figure A5: Reconstructions from single EEG trials using two disentangled β-VAEs pre-trained on either ERPs or single EEG sample trajectories (SMPL), as well as an AE pre-trained on ERPs (AE ERP). Models pre-trained on ERPs are able to reconstruct ERP-like trajectories even from single EEG samples, as demonstrated by the closer similarity of the β-VAEs (ERP) and AE (ERP) reconstructions to the ground truth ERPs (leftmost column). The ground truth ERPs were obtained by averaging on average 37 single EEG trials, including the single EEG trajectories that were used as inputs to the models.

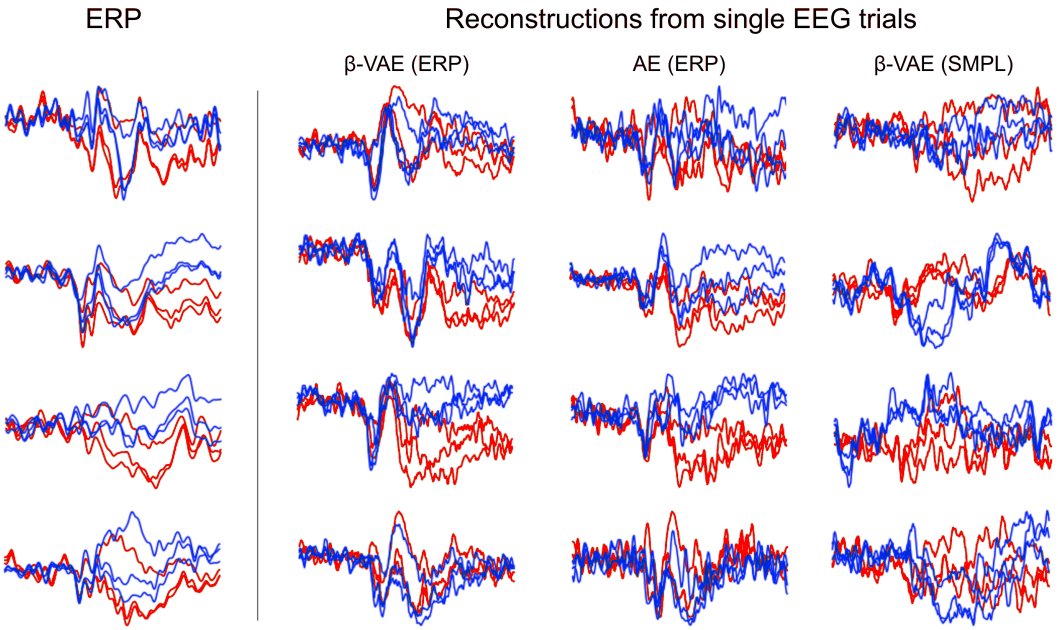

Figure A6: Reconstructions from single EEG trials using two disentangled β-VAEs pre-trained on either ERPs or single EEG sample trajectories (SMPL), as well as an AE pre-trained on ERPs (AE ERP). Models pre-trained on ERPs are able to reconstruct ERP-like trajectories even from single EEG samples, as demonstrated by the closer similarity of the β-VAEs (ERP) and AE (ERP) reconstructions to the ground truth ERPs (leftmost column). The ground truth ERPs were obtained by averaging on average 37 single EEG trials, including the single EEG trajectories that were used as inputs to the models.

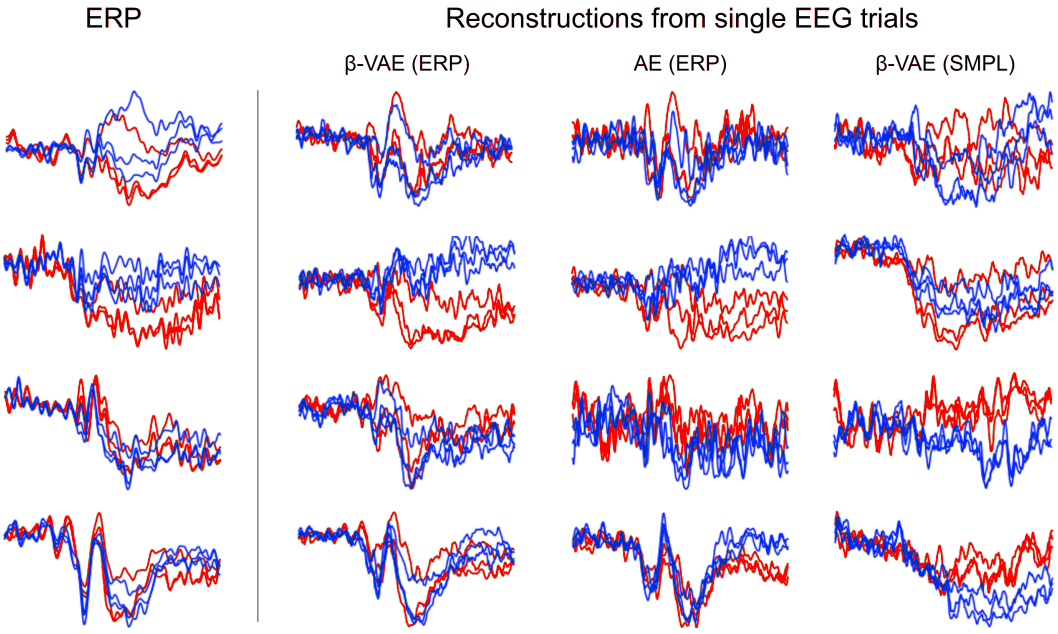

Figure A7: Reconstructions from single EEG trials using two disentangled β-VAEs pre-trained on either ERPs or single EEG sample trajectories (SMPL), as well as an AE pre-trained on ERPs (AE ERP). Models pre-trained on ERPs are able to reconstruct ERP-like trajectories even from single EEG samples, as demonstrated by the closer similarity of the β-VAEs (ERP) and AE (ERP) reconstructions to the ground truth ERPs (leftmost column). The ground truth ERPs were obtained by averaging on average 37 single EEG trials, including the single EEG trajectories that were used as inputs to the models.

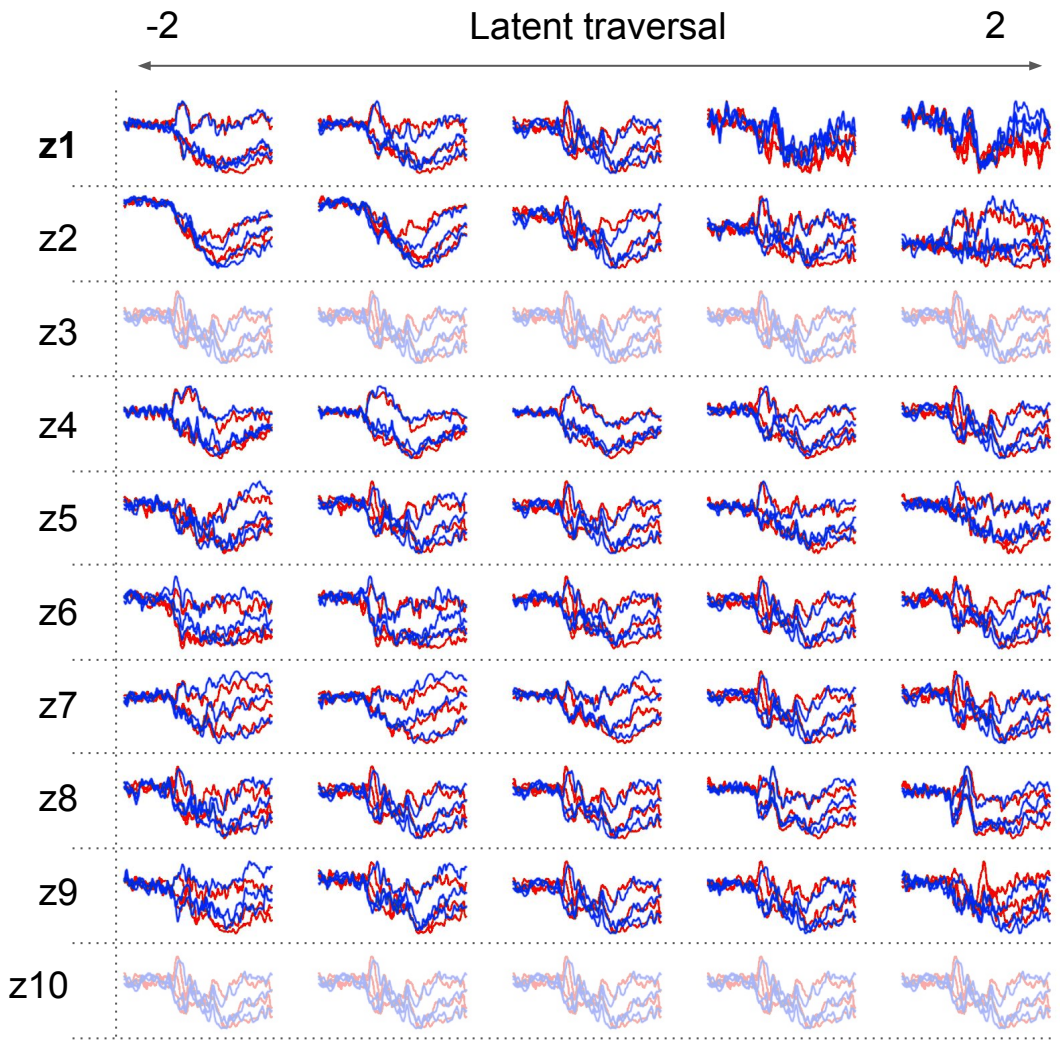

Figure A8: Latent traversals of a well disentangled pre-trained β-VAE with 8 "informative" dimensions (UDR=0.36) indicated by an arrow in Fig. A1. Each row reconstructs an ERP trajectory as the value of each latent dimension is traversed between [-2, 2] while keeping the values of all other latents fixed. Greyed out latents are "uninformative" – the model effectively switched them off by setting their inferred posterior to the unit Gaussian prior. Bold latent was identified by SCAN to be associated with depression.

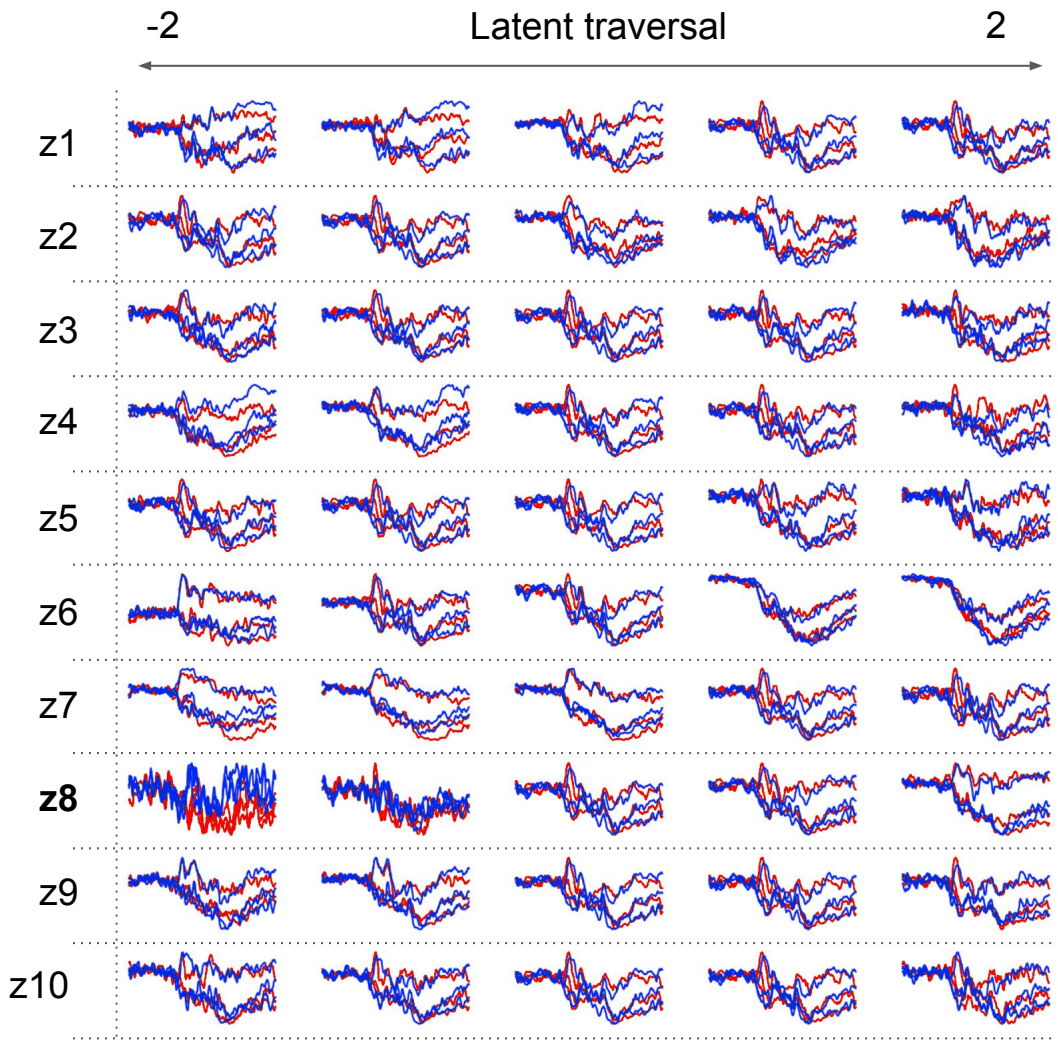

Figure A9: Latent traversals of a well disentangled pre-trained β-VAE with 9 "informative" dimensions (UDR=0.36) indicated by an arrow in Fig. A1. Each row reconstructs an ERP trajectory as the value of each latent dimension is traversed between [-2, 2] while keeping the values of all other latents fixed. Greyed out latents are "uninformative" – the model effectively switched them off by setting their inferred posterior to the unit Gaussian prior. Bold latent was identified by SCAN to be associated with depression.

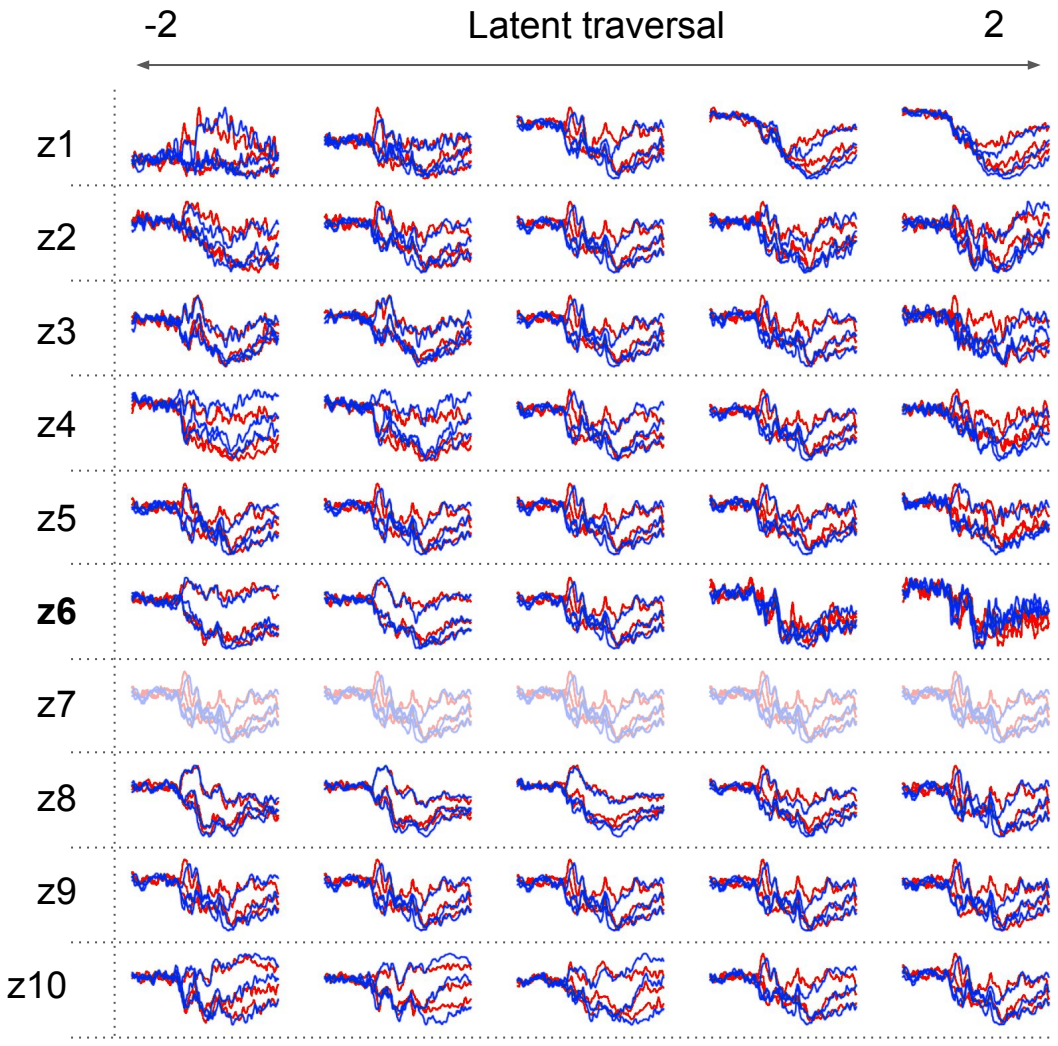

Figure A10: Latent traversals of a well disentangled pre-trained β-VAE with 10 "informative" dimensions (UDR=0.31) indicated by an arrow in Fig. A1. Each row reconstructs an ERP trajectory as the value of each latent dimension is traversed between [-2, 2] while keeping the values of all other latents fixed. Bold latent was identified by SCAN to be associated with depression.