Supplementary 3 – Cross Validation Using the Lock Box Feature

Sample Size: 26 Pipelines (5%) – Lock Box Subset of Participants

Output 1: Median Predictive Accuracy and Kolmogorov-Smirnov Statistic

Table 1. The median R^2 of the model across the pipelines included in each sample and the full multiverse, and the K-S statistic of the R^2 between each sample and that of the full multiverse. (Sample size = 26 pipelines, 5% of the multiverse, N = 28).

Sampling Approach	Median R^2	K-S Statistic
Full multiverse	0.138	-
Active learning	0.152	0.104**
Random	0.157	0.106
Stratified	0.111	0.167

* = $p < 0.05$, ** = $p < 0.01$, *** $p < 0.001$

Output 2: Predictive Accuracy of the Pipelines

Table 2. The ten pipelines with the highest R^2 from the full multiverse and each of the sampling approaches. Each result includes the pipeline identifier based on the combination of preprocessing options selected across decision nodes and the corresponding R^2 value for that pipeline, categorised by sample. (Sample size = 26 pipelines, 5% of the full multiverse, *N* = 28).

Best Full Sample	R^2	Best Active Learning	R^2	Best Random Sampling	R^2	Best Stratified Sampling	R^2
b-200, rMas, SAV400, CP1CP2PzP3P4	**0.522**	**b-200, rMas, SAV400, CP1CP2**	**0.452**	b-200, rMas, 700600, P3P4CP1CP2	0.387	**b-100, rAvg, 600200, CP1CP2PzP3P4**	**0.450**
b-100, rAvg, 700200, CP1CP2PzP3P4	**0.512**	**b-100, rAvg, SAV400, CP1CP2PzP3P4**	**0.451**	b-200, rAvg, 700600, CP1CP2PzP3P4	0.370	b-200, rMas, 600300, CP1CP2PzP3P4	0.401
b-200, rAvg, 700200, CP1CP2PzP3P4	**0.502**	**b-100, rAvg, 600200, CP1CP2PzP3P4**	**0.450**	b-100, rAvg, 600300, P3P4CP1CP2	0.321	b-200, rMas, 600300, CP1CP2	0.345
b-100, rMas, SAV400, CP1CP2	**0.469**	b-200, rMas, 600300, CP1CP2PzP3P4	0.401	b-200, rAvg, 500300, CP1CP2PzP3P4	0.318	b-200, rAvg, 600300, P3P4CP1CP2	0.338
b-200, rAvg, SAV400, CP1CP2PzP3P4	**0.458**	b-200, rAvg, 600300, CP1CP2	0.353	b-100, rMas, 500200, CP1CP2	0.287	b-200, rMas, 600200, CP1CP2	0.310
b-200, rMas, SAV400, CP1CP2	**0.452**	b-200, rMas, SAV400, P3P4CP1CP2	0.328	b-100, rAvg, 600300, Cz	0.243	b-100, rMas, 600600, P3P4CP1CP2	0.232
b-100, rAvg, SAV400, CP1CP2PzP3P4	**0.451**	b-100, rAvg, 700300, CP1CP2PzP3P4	0.316	b-200, rMas, 500300, CP1CP2	0.218	b-100, rAvg, 700600, Cz	0.182
b-100, rAvg, 600200, CP1CP2PzP3P4	**0.450**	b-100, rAvg, GAV400, P3P4CP1CP2	0.286	b-100, rAvg, 600300, Pz	0.214	b-100, rAvg, 600200, Cz	0.170
b-200, rAvg, 700200, CP1CP2	**0.443**	b-100, rAvg, 500200, CP1CP2PzP3P4	0.257	b-200, rMas, 600300, around	0.213	b-100, rCSD, 500200, Pz	0.151
b-100, rAvg, 700200, CP1CP2	**0.441**	b-100, rMas, SAV400, P3PzP4	0.250	b-200, rAvg, GAV400, Cz	0.192	b-100, rCSD, 500200, P3P4CP1CP2	0.135

Blue = the Pipeline is included within the best pipelines of the full multiverse. b-100 and b-200 = baseline durations of 100 and 200 ms respectively; rAvg, rMas and rCSD = common average, linked mastoids and current source density reference schemes respectively; around = 4 electrodes around the grand average midline peak.

Output 3: Raincloud Plots of Predictive Accuracies

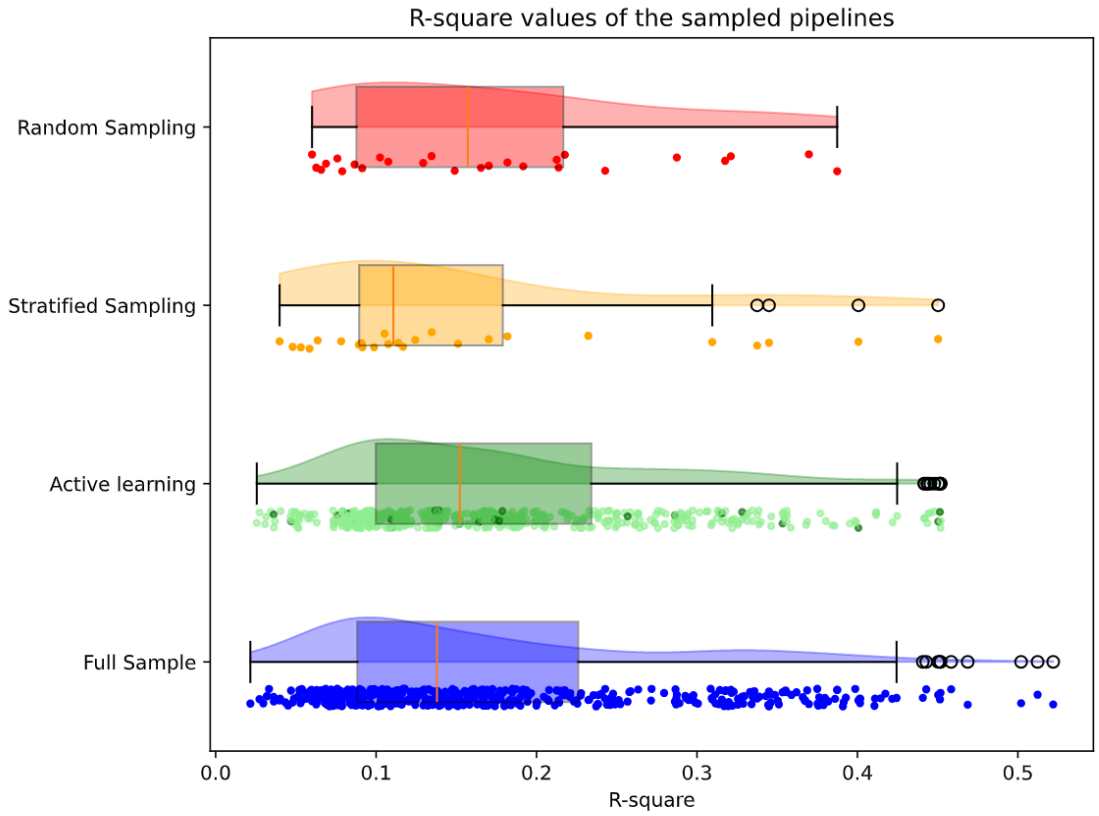

Figure 3. Raincloud plot of R^2-values of the pipelines sampled by each of the three sampling approaches and the full multiverse. Each sampling approach is grouped along the y-axis, and the R^2 of each pipeline included within each sample is plotted along the x-axis. The pipelines directly sampled by the active learning algorithm are in dark green, whereas the pipelines estimated are in light green (sample size = 26 pipelines, 5% of the multiverse, N = 28)

Output 4: Scatter Plots of Spatial Distribution in the Low Dimensional Space

Table 3. The average nearest neighbour distance between each sampling method and the full multiverse (*N* = 28).

Sample Size	Full Multiverse vs. Random Sample	Full Multiverse vs. Stratified Sample	Full Multiverse vs. Active Learning Sample
26 pipelines	0.442	0.517	0.422

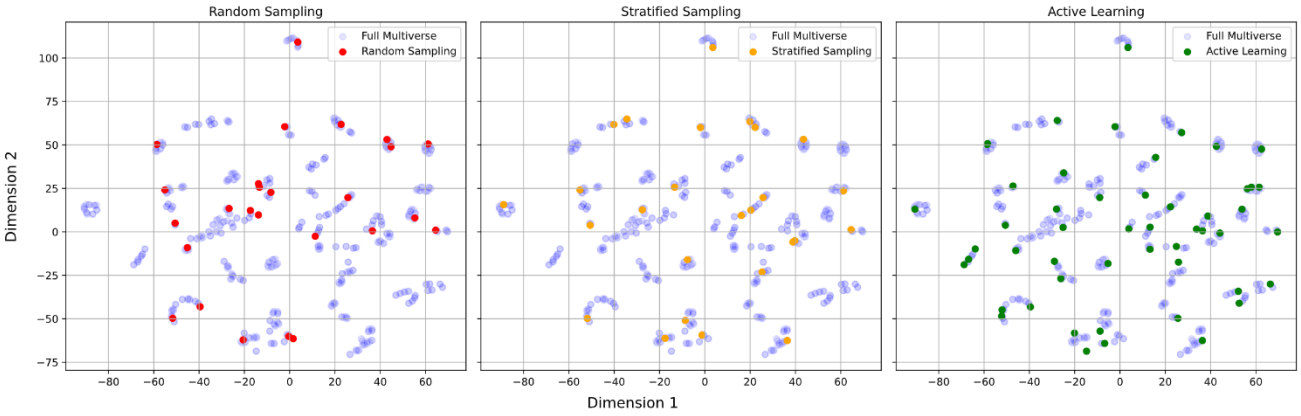

Figure 4. A set of scatter plots displaying the two-dimensional spatial distribution of pipelines, organised by pairwise Euclidean distances of the LPP vector of emotion conditions, for each sampling approach. The x- and y-axes represent dimensions of the low-dimensional embedding space derived from *t*-SNE, where proximity reflects similarity between multiverse pipelines. Green = active learning; red = random; orange = stratified, blue = unsampled pipelines. (Pipeline sample size = 26, 5% of the full multiverse, *N* = 28).

Output 5: Specification Curve

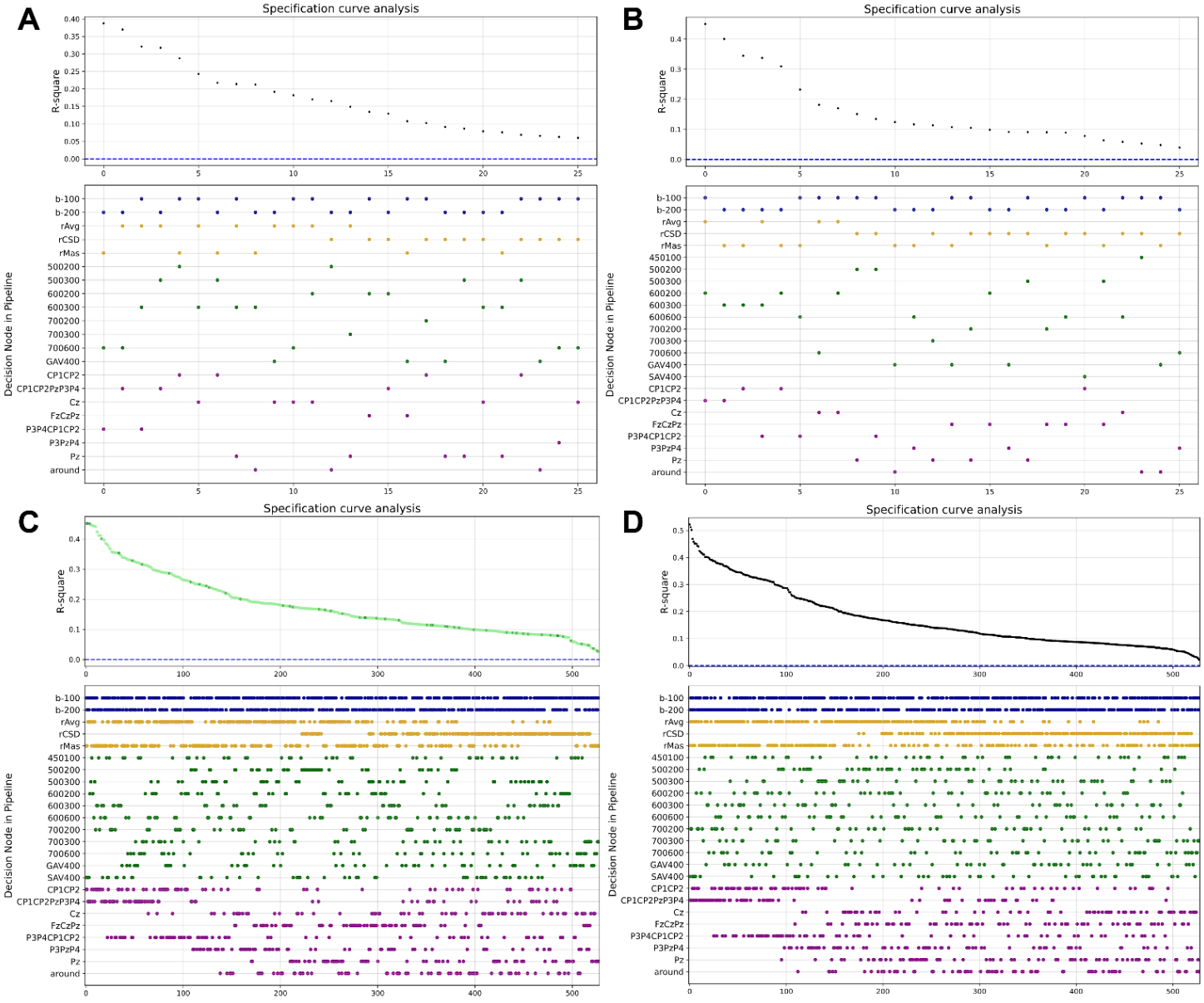

Figure 5. Specification curves displaying the variability in the R2 across each sample, in vertical alignment with the respective pipeline options. Panel A = random sample, Panel B = stratified sample, Panel C = active learning sample, Panel D = full multiverse. Within each specification curve (panels A – D), the y-axis of the upper sub-panel is the R^2 for each pipeline, the y-axis of bottom panel is the options implemented at each decision node for a given pipeline, and the x-axis which runs through the top and bottom panel in alignment is the pipeline within that sample. Each colour in the lower specification panel corresponds to one decision node. Blue = baseline duration, yellow = reference scheme, green = time window, purple = electrode cluster. Each row within a colour represents a different option at that decision node. In the top panel of the active learning plot, dark green points denote pipelines that were directly sampled and light green points denote pipelines that were estimated. (Sample size = 26 pipelines, 5% of the full multiverse, N = 28).