



**S1 Fig:** Bootstraps ( $n = 1000$ ) of submissions for (A) SC1: on/off, (B) SC2: dyskinesia, and (C) SC3: tremor. Team models (black) and their ensembles (blue) are ordered by rank. Boxes correspond to the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles, and individual points are displayed beyond  $1.5 \times \text{IQR}$  (interquartile range) from the edge of the box. For each sub-challenge, a null model (shown in red) estimated as the subject-specific mean of the training labels was used as a benchmark.