A Bayesian-optimized design for an interpretable convolutional neural network to decode and analyze the P300 response in autism

A Bayesian-optimized design for an interpretable convolutional neural network to decode and analyze the P300 response in autism

Davide Borra[1] (0000−0003−3791−8555)

Elisa Magosso[1,2,3] (0000−0002−4673−2974)

Miguel Castelo-Branco[4] (0000-0003-4364-6373)

Marco Simões[4,5] (0000-0003-3713-2464)

[1]Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, Cesena Campus, Cesena, Italy

[2]Alma Mater Research Institute for Human-Centered Artificial Intelligence, University of Bologna, Bologna, Italy

[3]Interdepartmental Center for Industrial Research on Health Sciences & Technologies, University of Bologna, Bologna, Italy

[4]CIBIT – Coimbra Institute for Biomedical Imaging and Translational Research, ICNAS – Institute of Nuclear Sciences Applied to Health, University of Coimbra, Coimbra, Portugal

[5]CISUC – Center for Informatics and Systems, University of Coimbra, Coimbra, Portugal

Word count: (max:12000)

Correspondence

Davide Borra

email: davide.borra2@unibo.it

ABSTRACT (max: 300 words)

Objective. P300 can be analyzed in autism spectrum disorder (ASD) to derive biomarkers and can be decoded in BCIs to reinforce ASD impaired skills. Convolutional neural networks (CNNs) have been proposed for P300 decoding, outperforming traditional algorithms but they i) do not investigate optimal designs in different training conditions; ii) lack in interpretability. To overcome these limitations, an interpretable CNN (ICNN), that we recently proposed for motor decoding, has been modified and adopted here, with its optimal design searched via Bayesian optimization.

Approach. The ICNN provides a straightforward interpretation of spectral and spatial features learned to decode P300. The Bayesian-optimized (BO) ICNN design was investigated separately for different training strategies (within-subject, within-session, and cross-subject) and BO models were used for the subsequent analyses. Specifically, transfer learning (TL) potentialities were investigated by assessing how pretrained cross-subject BO models performed on a new subject vs. random-initialized models. Furthermore, within-subject BO-derived models were combined with an Explanation Technique (ICNN+ET) to analyze P300 spectral and spatial features.

Main results. The ICNN resulted comparable or even outperformed existing CNNs, at the same time being lighter. Bayesian-optimized ICNN designs differed depending on the training strategy, needing more capacity as the training set variability increased. Furthermore, TL provided higher performance than networks trained from scratch. The ICNN+ET analysis suggested the frequency range [2, 5.8] Hz as the most relevant, and spatial features showed a right-hemispheric parietal asymmetry. The ICNN+ET-derived features, but not ERP-derived features, resulted significantly and highly correlated to ADOS clinical scores.

Significance. This study substantiates the idea that a CNN can be designed both accurate and interpretable for P300 decoding, with an optimized design depending on the training condition.

The novel ICNN-based analysis tool was able to better capture ASD neural signatures than traditional ERP analysis, possibly paving the way for identifying novel biomarkers.

KEYWORDS

Convolutional neural network, Electroencephalography, Autism Spectrum Disorder, P300, Interpretability, Bayesian optimization

1. INTRODUCTION

Autism spectrum disorder (ASD) is a set of neurodevelopmental conditions with persistent deficits in social communication and social interaction across multiple contexts, together with restricted, repetitive patterns of behavior, interests, or activities [1]. ASD people show difficulties in social-emotional reciprocity, in developing, maintaining, and understanding relationships, and in non-verbal communicative behaviors used for social interactions, such as joint attention [2–6]. Joint attention emerges during the first year of life and involves the non-verbal coordination of attention of two individuals towards an object or event [7], playing an important role in the development of social and language capabilities [8,9].

Approaches based on neuroimaging, e.g., diffusion tensor imaging and functional magnetic resonance imaging, are used to characterize and identify potential neural biomarkers of information-processing deficits in children with autism [10,11]. In addition, Event-Related Potentials (ERPs) computed from the electroencephalogram (EEG) provide a less expensive and portable way to study sensory information processing and are applied to study the neural response in autism following an incoming stimulus [12]. Investigations concern alterations in both early ERP components reflecting pre-attentive sensory processing and/or initial orientation and capture of attention, such as P100 [13] and N100 [14], and later components, such as P300 [15,16]. The P300 response is characterized by a positive deflection that occurs while attending a target stimulus; it peaks between 250-500 ms after the stimulus onset and it is mostly distributed on the scalp around midline electrode sites (Fz, Cz, Pz), increasing its amplitude from frontal to parietal sites [17]. This response can be elicited in two-stimuli oddball paradigms, where an infrequent (target) stimulus is immersed into a sequence of more frequent (standard) stimulus. The amplitude of the P300 response was found to be positively correlated with allocation of attention resources, stimulus recognition and updating of working memory [18,19]. Abnormalities in the sensory information processing following an incoming

stimulus were found in ASD, as quantified by a reduced P300 amplitude in auditory and visuo-spatial tasks compared to healthy subjects, reflecting deficiencies in cognitive, attentional, and working memory processes [12,20–27].

Besides being potentially useful as an EEG-based ASD biomarker, the P300 response can be used for ASD intervention. Indeed, Brain-Computer Interfaces (BCIs) proved to be useful personalized therapeutic approaches in ASD [30–34]; these interfaces can be designed to train ASD people via an EEG-based neurofeedback aimed to reinforce social interactions and communication skills (e.g., joint attention [28,29]). In this scenario, the P300 response elicited in visuo-spatial tasks represents an important control signal for the BCI system [28,29]. A crucial stage of a BCI is represented by the decoding algorithm, that detects the P300 response from the EEG and translates it into a command. Challenges to perform this step arise from the noise sensitivity, non-linearity and non-stationarity of the EEG, as these characteristics depend on the subject and on the environment [33]. In particular, the non-stationarity causes shifts in the EEG across trials and recording sessions. In addition, inter-subject variability across subjects, due to anatomical and physiopathological differences, hinders the design of a 'participant-agnostic' BCI. Therefore, due to intra- and inter-subject variabilities, most BCIs require long calibration times on each recording session to achieve satisfactory decoding performance [31]. Furthermore, due to reduced P300 amplitude in ASD, the decoding of the P300 response is even more challenging and, thus, the decoding performance may be negatively affected.

Machine learning algorithms have been widely adopted to learn discriminative patterns from the EEG to perform P300 decoding [32]. Among these algorithms, traditional decoders for P300-based BCIs perform a pre-processing step that includes band-pass filtering within fixed EEG bands (e.g., 0.5-4 Hz and 4-8 Hz [33]; 2-20 Hz and 2-8 Hz [34]; 2-12 Hz [35]), followed by extraction of features in the temporal, frequency and spatial domains; the latter are then

evaluated by a classifier such as linear discriminant analysis, support vector machine or multi-layer perceptron [33,35–40]. In addition to these traditional decoders, significant improvements in performance were found using convolutional neural networks (CNNs) [41–43]. CNNs are feed-forward neural networks including the convolutional operator at least in one layer. Inspired by the hierarchical structure of the ventral stream of the visual system, CNNs are composed by stacked layers of neurons, each neuron characterized by a local receptive field. Neurons in deeper layers have larger receptive field and respond to more complex features [44], enabling the learning of hierarchically structured features from the input signal. At variance with traditional machine learning algorithms, in which a separation between feature extraction, selection, and classification occurs, CNNs solve the decoding task in an end-to-end fashion, by automatically learning the more meaningful features for the addressed problem, i.e., discrimination of P300 events from single EEG trials. Therefore, CNNs do not rely on some characteristics extracted *a priori* from the signals (e.g., spectral contents within fixed bands), but automatically learn the relevant discriminative features to distinguish the P300 response from the input EEG trial. From their first application in P300 decoding with the simple design proposed by Cecotti et al. [45], CNNs were improved by including progressively more convolutional and regularization layers (e.g., dropout [46] and batch normalization [47]) [48]. Among these CNNs, EEGNet [49] and its variants [43,50] were found to be particularly suitable for P300 decoding, outperforming traditional machine learning solutions as well as other CNN-based approaches, also in case of P300-based BCIs aimed at ASD intervention [41,43]. In addition, by adopting specific training strategies (such as within-subject and cross-session strategy or cross-subject strategy), CNNs were found to be capable of learning robust features across sessions and subjects, incapsulating intra-subject and inter-subject variability, thus providing the potentiality of significantly reducing BCI calibration times [43].

However, despite the previous advantages, these algorithms have some limitations. First, they introduce a large number of trainable parameters, i.e., parameters to fit during the training process, and require setting many hyper-parameters, i.e., parameters that define the functional form of the decoder (e.g., convolutional filter size, number of convolutional filters, type of activation function, etc.) that must be set before the training.

Generally, hyper-parameters are selected by testing only a few configurations via empirical evaluations [43,48–50], and, thus, CNNs are built with a sub-optimal design in terms of performance. Furthermore, hyper-parameters are kept the same across different training strategies [43,50]. In this way, the network capacity (i.e., its ability of approximating a wide variety of functions) is kept the same even though the network faces problems with increasing difficulty across the different strategies, e.g., when moving from within-subject and within-session to cross-session and cross-subject decoders, the architecture must learn the relevant features from training distributions with increased variability. Searching for an optimal CNN hyper-parameter configuration, which is also specific for a particular training strategy, is even more necessary when designing a P300 decoder for a BCI-based ASD therapeutic approach, where the P300 response is attenuated and more difficult to distinguish than in case of healthy users.

Moreover, CNNs are scarcely interpretable in their learned features and are often treated as black boxes. As pointed out in a recent survey [51], there is a growing interest to interpret the feature representations provided by deep neural networks and their relation to clinical outcomes quantifying neuropathology. Thus, research is moving from the sole decoding of brain states towards the analysis of EEG features derived from the learning system, for example studying EEG signatures related to movement [52–56], P300 [43,57,58], or depression [59]. Recently, interpretable CNNs (ICNNs), i.e., CNNs that include layers whose learned parameters are directly interpretable, were proposed to decode motor imagery and execution. Zhao et al. [53]

increased feature interpretability in the frequency domain by reparametrizing a convolutional layer to learn Morlet wavelets. Recently, we proposed a lightweight (i.e., with a limited number of trainable parameters) ICNN [52,55] able to increase the interpretability of both spectral and spatial features, by reparametrizing a convolutional layer to learn band-pass filters and by learning spatial filters tied to each band-pass filter. In addition to interpretable layers, explanation techniques (ET) can be used to explain the CNN decision by highlighting which EEG features learned in interpretable domains (e.g., spatial, temporal, spectral) are the most discriminative (i.e., most salient) for the decoding of a specific cognitive state, such as the P300 [43,57,58]. Thus, by leveraging on the increased interpretability embedded into an ICNN, combined with an ET (denoted as ICNN+ET in the following), a data-driven non-linear analysis tool could be realized able to gain insights into the physiopathological neural signatures contained in the EEG associated to P300.

In this study, we aim to contribute to the decoding and, at the same time, to the analysis of the P300 in ASD using an ICNN, to further increase the feasibility of BCI intervention in autism and to potentially characterize novel data-driven biomarkers related to ASD visuo-spatial sensory processing. To this aim, here we adopted an architecture obtained by combining two existing CNN architectures, i.e., Sinc-ShallowNet and EEGNet. Sinc-ShallowNet [52] is an interpretable and lightweight CNN that we recently designed and proposed for decoding purposes other than P300 (specifically, decoding of motor imagery and execution), and that allows a straightforward interpretation of the learned spectral and spatial features. EEGNet represents the state-of-the-art (SOA) CNN for P300 decoding [41,50]. The proposed combination was conceived to obtain a CNN for P300 decoding (named Sinc-ShallowNet-v2 in the following), being both interpretable and with high decoding capabilities, as it inherits the interpretable characteristic from Sinc-ShallowNet and, at the same time, it includes some

design aspects of EEGNet. Then, using this architecture, we address the following two issues as main points of novelty of this study:

i. Investigate the optimal ICNN design (in terms of performance) and the role of the main ICNN hyper-parameters under different training conditions, by performing automatic hyper-parameter search (AHPS) based on Bayesian optimization (BO), separately for each training condition. The investigated training strategies were within-subject and within-session, within-subject and cross-session, and leave-one-subject-out. Furthermore, we also tested the capability to transfer the knowledge from other participants to a new one, adopting a transfer learning strategy to reduce calibration time. Each of these training strategies may represent a different practical scenario of BCI intervention to improve joint attention in ASD.

ii. Design of a novel algorithm based on the combination of the ICNN with saliency representations (ICNN+ET) to highlight the most relevant spectral and spatial features that correspond to the visuo-spatial P300 correlate in autism. These features were then used to define clusters of subjects characterized by shared neural signatures. Then, we investigated whether these features were related to clinical scores measuring the severity of ASD symptoms as derived by developmental and behavioral ASD assessment tools, and whether the ICNN+ET analysis better enhanced useful ASD neural signatures compared to a canonical ERP analysis.

Both the transfer learning and the explanation technique were applied to models derived from Bayesian optimization, as the latter, optimizing the decoding capabilities of the models, contributes to increase the performance and reliability of the obtained results.

2. MATERIALS AND METHODS

2.1. Dataset description

In this study we used the BCIAUT-P300 dataset, an EEG dataset of a feasibility clinical trial [28,29] publicly released for the IFMBE 2019 scientific challenge (https://www.kaggle.com/disbeat/bciaut-p300). EEG signals were recorded from 15 high-functioning ASD participants (22 years old, on average) while testing a P300-based BCI (based on statistical classifiers) aimed to improve their joint attention. During a baseline visit, the following clinical scores, useful also to diagnose ASD, were collected: Autism Diagnostic Observation Schedule (ADOS) [60], Autism Diagnostic Interview-Revised (ADI-R) [61], and Intelligence Quotients (IQs, measured by WAIS-III [62]). These scores are reported in Table 1.

For each participant, the BCI paradigm was carried out during 7 recording sessions (over 4 months) and in each session it was based on an immersive virtual environment presented to the user, consisting of a bedroom with an avatar, common furniture (e.g., shelves, a bed, etc.) and eight objects of interest. These objects were: 1) books on a shelf, 2) a radio on top of a dresser, 3) a printer on a shelf, 4) a laptop on a table, 5) a ball on the ground, 6) a corkboard on the wall, 7) a wooden plane hanging from the ceiling, and 8) a picture on the wall (see Figure 1). Participants were instructed to read non-verbal social agent cues from the avatar, i.e., avatar head turning to a particular object, and to pay attention to that target object. The 8 objects randomly flashed in the virtual scene and, thus, a visuo-spatial P300 response was elicited when the attended object flashed.

Each recording session was composed by a calibration phase, consisting of $N_b = 20$ blocks, and an online phase, consisting of $N_b = 50$ blocks. Each block was related to a particular object selected as target by the avatar that the user tried to identify (see Figure 1 for a scheme about the organization of blocks, runs, and trials). For each block, K runs were repeated ($K = 10$ in

the calibration phase, while K varied in the online phase with $K = 7.095$ on average across online blocks); each run consisted of each of the 8 objects flashing once in a random sequence. This resulted in 1600 trials (20 blocks x 10 runs x 8 EEG trials) per each participant and session during the calibration phase, and in 2838 trials on average during the online phase. During the calibration phase, the BCI statistical classifiers were trained to predict the object the participant was paying attention to; during the online phase, the trained classifiers were applied to identify whether the subject attended the target object correctly and, in that case, to provide positive feedback to the user.

The IFMBE 2019 scientific challenge was organized to stimulate researchers to develop decoders maximizing the object detection accuracy based on the P300 response. In the challenge, for each subject and each session, trials recorded during the calibration phase were released to *tune* decoders (i.e., to set their parameters), while trials recorded during the online phase were used to *test* decoders. In the present study, we adopted this same split as defined by the challenge, as this choice may also allow a fair and direct comparison with the results of decoders that participated to the challenge [41]. More specifically, for each subject and each session, we further split the 1600 calibration trials into two sets: 80% of trials (1280 trials), were randomly sampled and used as *training set* to optimize the trainable parameters of the decoders, while the remaining 20% of trials (320 trials) were used as *validation set* to optimize the hyper-parameters of the decoders. This further splitting was based on our winning solution of the challenge [41,50] to select the number of training epochs (i.e., to perform early stopping). Therefore, each subject-specific and session-specific dataset was separated into 3 different sets, as commonly performed in the literature [41,43,45,49,50,52,55,57,63]: training set (1280 trials), validation set (320 trials), and test set (2838 trials on average). However, since different training strategies were adopted here, by differently aggregating the training set and validation set across sessions and subjects, the overall number of training and validation trials were

different depending on the strategy (see Section 2.5.1). Finally, test trials were used to test the tuned decoders on a held-out set.

EEG signals were recorded at 250 Hz from C3, Cz, C4, CPz, P3, Pz, P4, and POz locations ($C = 8$ electrode sites), with the reference placed at the right ear and the ground at AFz. These signals were acquired notch filtered at 50 Hz and filtered between 2 and 30 Hz [61]. In this study, as in the winning solution [50] that we proposed for the challenge, each trial contained signals from -0.1 s to 1 s relative to the stimulus onset and signals were downsampled to 128 Hz, so that each trial contained $T = 140$ time steps.

2.2 Problem definition

Based on the previous description, the collection of trials associated to the s-th subject acquired during the r-th recording session can be formalized as the collection $D^{(s,r)} = \left\{\left(X_0^{(s,r)}, y_0^{(s,r)}\right), \dots, \left(X_i^{(s,r)}, y_i^{(s,r)}\right), \dots, \left(X_{M^{(s,r)}-1}^{(s,r)}, y_{M^{(s,r)}-1}^{(s,r)}\right)\right\}$. $M^{(s,r)}$ denotes the total number of trials for that subject and that session, $X_i^{(s,r)} \in \mathbb{R}^{C \times T}$ represents the pre-processed EEG signals of the i-th trial ($0 \leq i \leq M^{(s,r)} - 1$), and $y_i^{(s,r)}$ represents the label of the i-th trial, i.e., $y_i^{(s,r)} \in L = \{l_0, l_1\} = \{\text{non-P300, P300}\}$, where the label P300 was assigned to those trials where the flashing object coincided with the object the avatar was looking at.

In this study, each decoder consisted in a parametrized classifier f (representing the ICNN and having a different functional form depending on the hyper-parameters), which solves a binary classification task: $f\left(X_i^{(s,r)}; \theta\right): \mathbb{R}^{C \times T} \rightarrow L$, where θ represents the array of trainable parameters. Thus, the ICNN input was represented by $X_i^{(s,r)}$ that can be viewed as a 2D matrix of shape $(C, T) = (8,140)$ with electrodes along the height and time steps along the width, and the output consisted of two neurons corresponding to either one or the other class. The validation set was used to perform the automatic search of ICNN hyper-parameters via

Bayesian optimization, and the ICNN learned automatically from the training set the relevant features to assign the correct label to unseen input trials belonging to the test set. The knowledge learned by the ICNN during the training process was stored in its trainable parameters θ. These parameters, thanks to their increase interpretability, can be exploited to gain insights into the neural signatures related to a specific class (e.g., P300 class) in a data-driven way, without relying on handcrafted features based on expected EEG responses.

2.3. An update of Sinc-ShallowNet: Sinc-ShallowNet-v2

Sinc-ShallowNet [52] is an ICNN that we developed to decode motor imagery and execution from single EEG trials. This ICNN is composed of two blocks, each consisting of several stacked layers: an interpretable spectral and spatial (ISS) feature extractor followed by a fully-connected (FC) block that performs classification. The ISS block was designed to increase the interpretability of the learned parameters, at the same time keeping limited the model size, i.e., the number of trainable parameters, and included a temporal convolutional layer (with a reparameterization of the kernels), a depthwise spatial convolution and an averaging pooling layer (see also below for a more detailed description of the ISS block). Crucially, in Sinc-ShallowNet the output of the ISS block was provided directly as input to the FC block. Here, we proposed an updated version for P300 decoding, named Sinc-ShallowNet-v2, by integrating in the design also structure elements of EEGNet, a CNN proved to be particularly suitable to decode the P300 response from EEG [41,49,50] but not designed to be interpretable. Specifically, the updated Sinc-ShallowNet-v2 embraces three main blocks by including an additional block, the fixed-scale temporal (FST) feature extractor (inspired by EEGNet), between the ISS block and the FC classification block. This is an important modification, since Sinc-ShallowNet-v2 further processes the output of the ISS block by learning features in the temporal domain, and this may help to better capture intra- and inter-subject variability of P300

in time. A schematization of Sinc-ShallowNet-v2 is reported in Figure 2a; a more detailed description reporting the hyper-parameters, output shape and number of trainable parameters per layer, is reported in Table 2. In the following, the blocks are described in detail.

i. Block 1: Interpretable spectral and spatial (ISS) feature extractor

The first block was inspired by the ISS block of Sinc-ShallowNet [52] and was devoted to separately learn spectral and spatial features from the input EEG trials in an easy interpretable way. The very first layer of the ISS block was a temporal sinc-convolutional layer [52,55,64], learning ${K_0}^{ISS}$ filters with filter size ${F_0}^{ISS} = (1,65)$, unitary stride and zero-padding ${P_0}^{ISS} = (0,32)$ to preserve the number of input temporal samples. This temporal convolutional layer was devoted to filter each electrode signal in time. By using the sinc-convolutional layer to perform such processing step instead of a conventional convolutional layer, each convolutional filter is forced to describe a band-pass filter in the temporal domain. Denoting with k_j the j-th convolutional kernel, in a conventional convolutional layer each value of the filter (i.e., $k_j[0,n], n \in [0,64]$) has to be learned during the optimization process; conversely, in a sinc-convolutional layer, each value of the filter is defined by a parametrized function, forcing the filters to belong to a specific subset of temporal filters (here band-pass filters). Therefore, in a sinc-convolutional layer a re-parametrization of each convolutional kernel occurs:

$$k_j{}'\left[0,n;\{f_{0,j},f_{1,j}\}\right] = 2f_{1,j}sinc\left(2\pi f_{1,j}n\right) - 2f_{0,j}sinc\left(2\pi f_{0,j}n\right). \quad (1)$$

In Equation 1, $\{f_{0,j},f_{1,j}\}$ are the trainable parameters related to the j-th kernel, including only the inferior ($f_{0,j}$) and superior ($f_{1,j}$) cutoff frequencies of the band-pass filter. In this way, the number of trainable parameters reduces from 65 ($= {F_0}^{ISS}[0] \cdot {F_0}^{ISS}[1]$) to 2, for each temporal filter. Lastly, to alleviate the effects of the inevitable truncation of $k_j{}'$ on the characteristics of each filter, $k_j{}'$ is multiplied by a Humming window:

$$\begin{cases} {k_{w,j}}'\left[0,n;\{f_{0,j},f_{1,j}\}\right] = k_j{}'\left[0,n;\{f_{0,j},f_{1,j}\}\right] \cdot w[n] \\ w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{{F_0}^{ISS}[1]-1}\right) \end{cases}. \quad (2)$$

Accordingly, the temporal sinc-convolution computes the convolution between the input and $k_{w,j}{}'[0, n; \{f_{0,j}, f_{1,j}\}]$, learning only the following 2 parameters for each kernel:

$$\theta_{spect,j} = \{f_{0,j}, f_{1,j}\} \in \theta, 0 \leq j \leq {K_0}^{ISS} - 1. \quad (3)$$

Thus, the output of this first layer consists of stacked feature maps containing band-pass filtered versions of the input EEG trial within specific frequency ranges that were explicitly learned during training.

The use of a temporal sinc-convolutional layer, besides substantially reducing the parameters to fit, promotes the learning of more meaningful and well-defined temporal filters, and provides a straightforward interpretation of the learned spectral features [64], being the cutoff frequencies of the learned band-pass filters. Conversely, in case of a conventional temporal convolutional layer, the learned spectral features are not immediately accessible and interpretable.

Downstream the temporal sinc-convolutional layer, a batch normalization layer [47] was included. Then, a spatial depthwise convolutional layer was introduced: for each band-pass filtered map, ${D_1}^{ISS}$ spatial filters were learned having size $(C, 1)$, unitary stride and no zero-padding, i.e., ${D_1}^{ISS}$ spatial combinations of electrodes were learned for each band-pass filtered map. Therefore, a total number of ${K_1}^{ISS} = {K_0}^{ISS} \cdot {D_1}^{ISS}$ spatial filters were learned and constrained to have a norm upper bounded by ${c_1}^{ISS}$ (kernel max-norm constraint). This type of convolution does not exploit dense connections across feature maps as in traditional convolutional layers, reducing the number of trainable parameters. In addition, the combination of temporal sinc-convolution with spatial depthwise convolution leads to an interpretable spectral-spatial feature learning, as each group of ${D_1}^{ISS}$ spatial filters is strictly tied (via depthwise convolution) to a specific band-pass filter, i.e., to a specific frequency range:

$$\theta_{spat,j} = \{\theta_{j0}, \dots \theta_{jk}, \dots, \theta_{j{D_1}^{ISS}-1}\} \in \theta, 0 \leq j \leq {K_0}^{ISS} - 1, \quad (4)$$

indicating with θ_{jk} the k-th spatial filter ($0 \leq k \leq {D_1}^{ISS} - 1$) tied to the j-th band-pass filter.

The whole parameters learned in these two first convolutional layers can be denoted as the set:

$$\theta_{ISS} = \left\{\theta_{ISS,0}, \dots, \theta_{ISS,j}, \dots, \theta_{ISS,{K_0}^{ISS}-1}\right\} =$$

$$\left\{\left(\theta_{spect,0}, \theta_{spat,0}\right), \dots, \left(\theta_{spect,j}, \theta_{spat,j}\right), \dots, \left(\theta_{spect,{K_0}^{ISS}-1}, \theta_{spat,{K_0}^{ISS}-1}\right)\right\}, \quad (5)$$

where each pair $\left(\theta_{spect,j}, \theta_{spat,j}\right)$, $0 \leq j \leq {K_0}^{ISS} - 1$ contains the cutoff frequencies of the j-th band-pass filter and the associated ${D_1}^{ISS}$ spatial filters exploited to decode the input EEG trial.

Output activations of the interpretable spectral-spatial feature extractor were then normalized via batch normalization [47] and activated via an Exponential Linear Unit (ELU) non-linearity [65], i.e., $f(x) = x, x > 0$ and $f(x) = exp(x) - 1, x \leq 0$. Average pooling was introduced to reduce the temporal dimension of the activations by a factor of 4 (${F_p}^{ISS} = {S_p}^{ISS} = (1, 4)$) from $T = 140$ to $T//4 = 35$ (indicating with // the floor division operator). Lastly, a dropout layer [46] was added with dropout rate p^{ISS}.

ii. Block 2: Fixed-scale temporal (FST) feature extractor

This block was inspired by EEGNet, and in Sinc-ShallowNet-v2 was used to learn how to summarize the feature maps provided by the ISS block in the time domain; it contained a separable convolutional layer [66], defined by a temporal depthwise convolution with ${D_0}^{FST} = 1$, filter size ${F_0}^{FST}$, unitary stride and zero padding ${P_0}^{FST} = (0, {F_0}^{FST}[1]//2)$, followed by a pointwise convolution with ${K_0}^{FST}$ filters. In the first layer of this composition, depending on ${F_0}^{FST}$, temporal features were learned on the input ISS feature maps within a temporal window of specific size, i.e., at a fixed temporal scale, and without using dense connections across feature maps. The temporal window in which features are learned corresponds to ${F_0}^{FST}[1]/\ (sf//4)$ s, indicating with sf the sampling frequency of the input EEG signals. In the second layer of the separable convolutional, feature maps at the output of the previous layer

are recombined, and, this is performed learning compressed, equal or overcomplete representations, depending on whether ${K_0}^{FST} < {K_1}^{ISS}$, ${K_0}^{FST} = {K_1}^{ISS}$ or ${K_0}^{FST} > {K_1}^{ISS}$, respectively. Then, activations were normalized via batch normalization [47] and activated via an ELU non-linearity [65]. An average pooling layer was introduced to further reduce the temporal dimension of the activations by a factor of 8 (${F_p}^{FST} = {S_p}^{FST} = (1, 8)$), i.e., from $T//4 = 35$ to $T//32 = 4$. Lastly, a dropout layer [46] was added with dropout rate p^{FST}.

iii. Block 3: Fully-connected (FC) block

This block traduces the activations provided by the FST block into conditional probabilities, finalizing the decoding task. At first, the input feature maps of the FC block were unrolled along one single dimension using a flatten layer, producing a single feature array. Then, this array was provided as input to a single fully-connected layer with $N^{FC} = 2$ neurons associated to the non-P300 and P300 classes. Here, trainable parameters were constrained to have a norm upper bounded by ${c_0}^{FC}$. Lastly, these 2 neurons were activated using a softmax activation function to obtain the output conditional probabilities $p\left(l_k \middle| X_i^{(s,r)}\right)$, $k = 0, 1$.

2.4. Performance metric and comparison of Sinc-ShallowNet-v2 with EEGNet and Sinc-ShallowNet

In this study, we adopted the object-level accuracy as performance metric, i.e., accuracy in decoding the flashing object the participant was paying attention to (among the 8 possible objects, see Section 2.1 for additional details). This performance metric was the same as adopted in the IFBME 2019 competition to evaluate decoders [41]. To this aim, the trial-level EEG decoding provided by the CNN (binary classification, i.e., non-P300 vs. P300) was used to produce an object-level decoding (8-class classification, i.e., discrimination of the attended object among 8 objects). Considering a specific recording block, associated to a specific object the participant was paying attention to, the following processing was performed. Indicating

with l_1 the P300 condition, we considered the probabilities $p\left(l_1 \middle| X_i^{(s,r)}\right)$ predicted for the EEG trials $X_i^{(s,r)}$ within that block when each of the 8 objects flashed (including the attended one). These probabilities where averaged across runs separately for each object, obtaining the average probability $\bar{p}_o, 0 \leq o \leq 7$ that the participant was paying attention to the o-th object Then, the object with the highest probability was predicted as the one attended in that block.

We compared the performance of Sinc-ShallowNet-v2 against our approach derived from EEGNet [50] that won the IFMBE 2019 challenge. In the competition, that approach significantly outperformed traditional machine learning solutions as well as other deep neural networks, including both CNNs [48] and recurrent neural networks [41], thus, it can be considered the current SOA algorithm for the addressed decoding problem. Furthermore, Sinc-ShallowNet-v2 performance was compared against Sinc-ShallowNet [52], to assess whether the inclusion of the FST block in the updated version led to potential benefit in P300 decoding compared to the previous architecture (lacking this block).

2.5. Trainable parameter optimization

2.5.1. Training strategies

The structure of the adopted dataset BCIAUT-P300 (including several sessions per subject and a large number of trials per session), allows to train and evaluate P300 decoders in a large spectrum of training conditions, each reflecting a different scenario in which a P300-based BCI intervention might be applied. In this study, we assessed the optimal design of the proposed ICNN for P300 decoding, separately for each of different training strategies, characterized by increasing variability in the dataset, i.e., simulating BCI paradigms applied progressively to more sessions and subjects.

As introduced in Section 2.1, the trials of each session and subject were divided into training validation sets (80% and 20% of trials of the calibration phase, randomly sampled) and test set

(trials of the online phase). Then, depending on whether training and validation sets were considered separately for each subject or session, or were differently aggregated across sessions and subjects, four different training strategies were designed and investigated. It is worth mentioning that, despite different training strategies were employed, the definition of the test set was the same across training strategies to perform a fair comparison across them, that is, the test set always included trials belonging to the 50 online blocks. These strategies are described in the following. Furthermore, Figure 2b provides a schematization of the adopted strategies, and Table 3 a summary of the number of trials belonging to the training, validation and test sets in each strategy.

i. Within-subject and within-session (WS-WS) strategy.

In this strategy, subject- and session-specific training and validation sets were used to tune subject- and session-specific CNNs. The test set was defined as the subject- and session-specific test set, i.e., trials of the online phase belonging to the subject and session the CNN was trained for (see Table 3A). Furthermore, to simulate a practical scenario where limited trials are available in a single recording session, we trained and validated CNNs using a progressively increasing number of calibration blocks (see Section 2.1), and thus, using variable-sized training and validation sets. This condition was implemented by sampling training and validation examples from the first n calibration blocks, where n was progressively increased from 2 to 20 with a step of 2 blocks (see Table 3B), where 20 is the overall number of calibration blocks in each session.

ii. Within-subject and cross-session (WS-CS) strategy.

In this strategy, subject-specific training and validation sets were used to tune subject-specific CNNs. For the s-th subject, training and validation sets were merged across all recording sessions of that specific subject (see Table 3A). incorporating within-subject variability. The architecture was then tested separately on each subject- and session-specific

test set, i.e., on trials of the online phase of each session belonging to the subject the CNN was trained for.

iii. Leave-one-subject-out (LOSO) strategy.

In this strategy, cross-subject training and validation sets were used to tune cross-subject, cross-session and subject-agnostic CNNs, i.e., the training and validation sets included only examples sampled from other subject distributions. In particular, for each subject s, named "held-back subject", training and validation sets were merged across all sessions from all the other subjects (different from the s-th subject, see Table 3A), incorporating between-subject and within-subject variability. The architecture was then tested separately on each subject- and session-specific test set, i.e., on trials of the online phase of each session belonging to the held-back subject (s-th subject).

iv. Transfer learning on single sessions (TL-WS).

Transfer learning focuses on transferring the knowledge across domains/tasks. It is inspired by the human capability to use the knowledge learned in a source domain/task to improve the performance and/or reduce the training time in a related domain/task [67]. In this strategy, as for the WS-WS strategy (2.5.1-i), training and validation trials were subject- and session-specific. In particular, the definition of training and validation sets was the same as the one adopted in the WS-WS strategy when analyzing variable-sized training and validation sets (depending on the number of included calibration blocks), while keeping unchanged the subject- and session-specific test set, i.e., the trials of the online phase of that subject and session (see Table 3B). However, differently from the WS-WS strategy where the trainable parameters were initialized randomly, in the TL-WS strategy, for the s-th subject and r-th recording session the CNN was initialized using the CNN trained with LOSO strategy when the s-th subject was held-back. That is, the knowledge learned during the LOSO strategy from many subjects except the held-back one was transferred on the held back subject. The use of

TL-WS strategy could be useful when a new user approaches the BCI system in a new recording session and a calibration stage, as short as possible, is needed to tune an accurate decoder. In this view, the knowledge embedded in LOSO models, i.e., incorporating between-subject and within-subject variabilities, could represent a better initialization point in the space of parameter θ than the random one, potentially leading to an improvement in performance and/or to a reduction of training and validation examples needed to achieve high performance. The potentialities of transfer learning were tested by comparing the performance of the TL-WS models with the performance of the WS-WS models, while increasing the size of the training and validation sets, thus, assessing the benefits of the use of models pre-trained on other subjects compared to models trained from scratch.

2.5.2. Training settings

CNN optimization consisted of the minimization of the cross-entropy between the empirical probability distribution defined by the training labels, and the probability distribution defined by the model. This corresponds to minimize the Kullback-Leibler divergence between the two probability distributions at trial-level, and thus also at block-level, i.e., object-level. Adaptive moment estimation (Adam) [68] was used as optimizer with learning rate lr, mini-batch size $bs = 64$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for computing the running averages of the gradient and its square, and $\varepsilon = 10^{-8}$ to improve numerical stability. To address class imbalance, parameter updates were weighted depending on the class occurrence of the training examples: indicating with M_0 and M_1, the number of trials for non-P300 and P300 conditions in the training set and given that $M_0 > M_1$, class weights were defined as 1 and $M_0/M_1 = 7$, respectively for non-P300 and P300 classes. The maximum number of epochs was set to 1000 and early stopping was performed by interrupting the training loop when the validation loss did not decrease for 50 consecutive epochs. Besides early stopping, which acts as regularizer, Sinc-ShallowNet-v2

implicitly included in its structure also many layers devoted to increase the generalization, such as batch normalization [47] (with a momentum term $m = 0.99$ and $\varepsilon =$ 1e-3 for numerical stability), dropout [46] and kernel max norm constraint.

2.6. Hyper-parameter optimization

Hyper-parameter optimization is devoted to find the optimal hyper-parameter configuration of a learning system associated with the best performance measured on a separate validation set. In the following, an overview of hyper-parameter tuning via Bayesian optimization (BO) [69] is provided; then, the procedure adopted to perform BO hyper-parameter search and to report the obtained results is described.

2.6.1. Hyper-parameter search via Bayesian optimization

Denoting with h the array containing the hyper-parameters of interest ($h \in H$, where H is the hyper-parameter search space), the aim of hyper-parameter optimization is to find $h^* = arg\min_{h \in H} k(h)$, that minimizes the objective function $k(h)$ (the cross-entropy loss, in this study) evaluated on the validation set. However, the evaluation of the objective function $k(h)$ requires a new training stage and a new evaluation stage (on the validation set) for each hyper-parameter configuration h. Thus, depending on the number of hyper-parameters to optimize and on the complexity of the model, which are generally both high in deep learning-based approaches, this process can be expensive. Common hyper-parameter search algorithms, such as grid search and random search, do not take advantage of the results of the previous iterations to select the next hyper-parameters to evaluate (the latter are, thus, uninformed by past evaluations), often wasting computational time on hyper-parameters with poor performance. Conversely, BO methods keep track of results related to past evaluations to update a probabilistic model, mapping hyper-parameters to the probability to obtain a specific score from the objective function. That is, BO can be formalized as a sequential model-based optimization, performing

several iterations each one considering a specific hyper-parameter configuration and updating the probability model. This model is a surrogate of the objective function to be minimized and is easier to optimize than the original objective function. BO methods suggest the hyper-parameters in an informed way after each iteration step, based on a "selection function". By investigating hyper-parameters that seem promising based on past results, BO methods can find better configurations than other approaches within fewer iterations compared to uninformed algorithms (e.g., grid or random search) [69].

2.6.2. Hyper-parameter search settings and analysis

Optimal hyper-parameters were searched via BO (using the Python library Optuna [70], version 2.3.0), using tree-structured Parzen estimator as surrogate function and expected improvement as selection function, and sequential model-based optimization was performed for 100 iterations (which is the default value [70]). The hyper-parameters of Sinc-ShallowNet-v2 subjected to BO, defining the array $h \in H$, were ${K_0}^{ISS}$, ${D_1}^{ISS}$, ${c_1}^{ISS}$, $p^{ISS} = p^{FST} = p$, ${K_0}^{FST}$, ${F_0}^{FST}$, ${c_0}^{FC}$ and lr. Within each BO iteration, the hyper-parameters were sampled from the distributions specified in Table 4. Note that, based on the adopted distributions, since ${K_0}^{FST}$ can be equal to or less than (but not greater than) ${K_0}^{ISS} \cdot {D_1}^{ISS}$, only compressed or equal (but not overcomplete) representations in FST layer were examined. BO was applied to models trained with WS-WS, WS-CS and LOSO strategies to investigate the optimal design of Sinc-ShallowNet-v2 depending on the training strategy (see Figure 2b). BO was not applied to TL-WS, as models trained in this strategy strictly depended on LOSO models. Indeed, in TL-WS the ICNN was initialized with LOSO parameters and then trained on the held-back subject. Therefore, the hyper-parameter configuration adopted in TL-WS models needed to be the same as in the LOSO strategy, i.e., hyper-parameters needed to be kept fixed as the ones of Bayesian-optimized LOSO models. Therefore, also the WS-WS models used for comparison with TL-WS models, while training and evaluating decoders with a progressively increasing size of

training and validation sets, were assigned to the Bayesian-optimized hyper-parameters of the LOSO models, for a fair comparison. It is worth remarking that while analyzing the potentialities of transfer learning with TL-WS and WS-WS models, where hyper-parameters were inherited from LOSO models, the validation set was used only to perform early stopping.

For each training condition (WS-WS, WS-CS and LOSO), and for a given q-th hyper-parameter $\in h$ ($0 \leq q \leq 7$) the optimal values obtained across the Bayesian-optimized ICNNs were extracted (we had 15·7 ICNNs in WS-WS and 15 ICNNs in WS-CS and LOSO). Then, the probability (P) to obtain a specific hyper-parameter value via BO across ICNNs was computed. This allowed a visualization and a comparison of the optimal model design across training conditions. Moreover, for each training condition the importance score z_q was derived by using the fANOVA hyper-parameter importance evaluation algorithm proposed in [71] that fits a random forest regression model predicting the objective value given a parameter configuration. Importance scores sum up to 1 over the investigated hyper-parameters (i.e., $\sum_q z_q = 1$) and the higher the score of a specific hyper-parameter, the higher its importance.

To provide a fair performance comparison with the other two CNN-based approaches, the EEGNet adaptation used in [50] and Sinc-ShallowNet [52] (see Section 2.4), the optimal hyper-parameters of EEGNet and Sinc-ShallowNet were searched using BO for each training condition too. In particular, the hyper-parameters of EEGNet and Sinc-ShallowNet were sampled using the same distributions defined as for Sinc-ShallowNet-v2 (see Table 4); of course, Sinc-ShallowNet lacked the hyper-parameters of the FST block (K_0^{FST} and F_0^{FST}) as this block was absent in this architecture.

2.7. Explanation technique: spectral and spatial features analysis

The interpretation of the features learned by Sinc-ShallowNet-v2 incorporating subject-specific knowledge ($\theta = \theta^{(s)}$, in WS-CS) could provide insights about neural signatures

related to each specific ASD subject in a data-driven way, enabling between-subject variability investigation in ASD. For this analysis, we retrained the ICNN in WS-CS using the most frequent Bayesian-optimized configuration obtained in WS-WS across all subjects and sessions. By doing so, only one fixed hyper-parameter configuration was used to retrain WS-CS models across all subjects, so that the same number of temporal and spatial filters in the ISS block – whose learned features are the objects of the following analysis – was used across subjects. In this way, we were able to evidence differences/similarities among ASD subjects in terms of the learned features, excluding that these differences may arose from differences in ICNN configurations. The fixed configuration was based on the most frequent WS-WS optimal configuration, as the Bayesian-optimized WS-WS models resulted the lightest and fastest to train (see Figure 4 and Table 5). It is worth noticing that the WS-CS retrained models achieved comparable performance with the Bayesian-optimized WS-CS models (see Section 1 of Supplementary Materials).

As described in Section 2.3-i, the adopted architecture was designed to provide interpretable parameters in the array $\theta_{ISS}{}^{(s)}$. The ICNN processes input trials by filtering out P300-unrelated spectral and spatial information while preserving only the most significant ones for P300 decoding. However, features may have a different importance on the discrimination, i.e., a band-pass filtering in a peculiar frequency range and a subset of electrodes may be more relevant to distinguish the P300 response. Therefore, the processing of θ_{ISS} should also include an explanation technique, devoted to highlight contributions of the more important features in ASD related to P300, realizing the combination ICNN+ET.

In the following, the proposed algorithm based on ICNN+ET for the investigation of subject-specific spectral and spatial P300 features is described and a schematic representation is reported in Figure 3.

2.7.1. Computation of the relevance of spectral features (spectral relevance)

At first, given a single input EEG trial of the s-th subject containing the P300 response ($X_i^{(s,r)}$), we computed the saliency [72] for the P300 class of each spatio-temporal sample ($C \cdot T$ samples) within each feature map provided by the temporal sinc-convolutional layer. This ET consists in computing the gradient of the output of the neuron associated to the P300 condition (immediately before the softmax function) with respect to the feature maps provided by the first convolutional layer. The output of the ET is one relevance map for each feature map of the first layer (i.e., $\forall j$, with $0 \leq j \leq {K_0}^{ISS} - 1$), and can be interpreted as a transformation $e\left(X_i^{(s,r)}\right): \mathbb{R}^{C \times T} \rightarrow \mathbb{R}^{{K_0}^{ISS} \times C \times T}$ whose output quantifies how much the spatio-temporal samples in each filtered version of the input affect the P300 prediction. Relevance maps were averaged across P300 trials, and the absolute value of these maps was computed. Therefore, an average relevance map $\bar{G}_j^{(s)} \in \mathbb{R}^{C \times T}, 0 \leq j \leq {K_0}^{ISS} - 1$ was obtained, and finally, the relevance score of each feature map was computed as:

$$g_j^{(s)} = \max_{c,t} \bar{G}_j^{(s)} / \max_{c,t,j} \bar{G}_j^{(s)}, 0 \leq j \leq {K_0}^{ISS} - 1, 0 \leq c \leq C - 1, 0 \leq t \leq T - 1, \quad (6)$$

where $g_j^{(s)} \in [0,1]$ is a scalar quantity that summarizes the importance of each band-pass filter for discriminating the P300 response from input EEG trials.

The spectral relevance scores obtained in Equation 6 were used to weight the associated passbands of the temporal filters, defined by ${\theta_{spect,j}}^{(s)}$ (see Equation 3), computing $r_j^{(s)}(f)$ as follows:

$$\begin{cases} p_j^{(s)}(f) = \begin{cases} 1, \ if \ f_{0,j}^{(s)} \leq f \leq f_{1,j}^{(s)} \\ 0, \ elsewhere \end{cases} & 0 \leq j \leq {K_0}^{ISS} - 1 \\ r_j^{(s)}(f) = g_j^{(s)} \cdot p_j^{(s)}(f) & 0 \leq j \leq {K_0}^{ISS} - 1 \end{cases}. \quad (7)$$

In Equation 7, $p_j^{(s)}(f)$ indicates the probability that a specific frequency f was included in the passband of the j-th band-pass filter, i.e., marking with a probability of 0 and 1 frequencies outside and inside the passband, respectively.

Then, the spectral relevance $r^{(s)}(f)$, quantifying the relevance of each frequency bin, was obtained as:

$$r^{(s)}(f) = \max_j r_j^{(s)}(f). \quad (8)$$

2.7.2. Computation of the relevance of spatial features related to the most relevant spectral features (spatial relevance)

By averaging $r^{(s)}(f)$ across subjects, the frequency ranges retaining a relevance greater than or equal to 0.75 across subjects ($\forall s$) were identified. As described in Section 3.3, only one frequency range was identified, and it is indicated in the following with $[f_0^*, f_1^*]$. This interval is characterized by a high relevance with high agreement (≥75%) across subjects, as the spectral relevance was normalized for each subject. For each subject the following analysis was performed. First, we considered the subset of the band-pass filters, denoted as $S^{(s)}$, that contained in their passband the frequency bins belonging to $[f_0^*, f_1^*]$ and we selected the spatial filters (i.e., the learned features $\theta_{jk}^{(s)}$, see Equation 4) associated to this subset of band-pass filters ($j \in S^{(s)}, 0 \leq k \leq D_1^{ISS} - 1$). As we were interested in the investigation of the relevance at the level of single electrode, spatial filters were considered in their absolute value, as done in [45,52]. Subsequently, the absolute spatial features were averaged together electrode per electrode ($\forall c$) and normalized to the maximum across electrodes, obtaining the spatial relevance:

$$\bar{\theta}_{spat}^{(s)} = \frac{1}{N_{spat}^{(s)}} \sum_{j \in S^{(s)},k} abs(\theta_{jk}^{(s)}) \Big/ \max_c \frac{1}{N_{spat}^{(s)}} \sum_{j \in S^{(s)},k} abs(\theta_{jk}^{(s)}),\ 0 \leq c \leq C - 1, \quad (9)$$

where $N_{spat}^{(s)}$ represents the overall number of spatial filters associated to the subset $S^{(s)}$ of the band-pass filters (D_1^{ISS} spatial filters for each band-pass filter in the subset).

Based on Equation 9, each value of $\bar{\theta}_{spat}^{(s)} \in \mathbb{R}^C$ quantifies the relevance (normalized between [0,1]) of a specific electrode signal filtered in the frequency range most important for

decoding the P300 response. Abnormalities in P300 (e.g., ASD abnormalities) may modulate $\bar{\theta}_{spat}^{(s)}$.

2.7.3. Clustering

The spatial relevance $\bar{\theta}_{spat}^{(s)}$ was clustered, automatically finding clusters of subjects characterized by different patterns of $\bar{\theta}_{spat}^{(s)}$. This was made possible by adopting the WS-CS strategy, as it allows the learning of robust subject-specific neural signatures across recording sessions. Here, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [73] was used as clustering algorithm (using the Python library hdbscan [74], version 0.8.27), using the Euclidean distance between observations as distance metric. With HDBSCAN, clustering is performed without specifying the number of clusters as input parameter, including also an outlier detection algorithm [73]. Therefore, the adopted clustering algorithm provides the optimal number of clusters N_{clust} populated by the more reliable observations and marks as outliers a subset O of observations. As observations belonging to the same cluster share a common structure, these were averaged together, obtaining $\bar{\theta}_{c,spat}$ for the c-th cluster. Therefore, cluster-level representations were obtained by averaging across a small subset of subjects sharing a peculiar spatial pattern as learned by the model, reflecting neural signatures belonging to different clusters of ASD subjects.

2.8. Statistical analysis

2.8.1. Performance

As described in Section 2.4, object-level accuracy was used as performance metric, and it was computed separately for each session-specific test set. In addition, as performed in [41], as the performance resulted robust across recording sessions, the performance metric was averaged across all sessions for each subject. Then, the following tests were performed.

i. A Friedmann test was performed to compare the performance obtained by Sinc-ShallowNet-v2 trained with the three training strategies (WS-WS, WS-CS, LOSO

strategies). Then, as significant differences were found (see Section 3.2), post-hoc pairwise comparisons were performed testing all combinations; a total of 3 tests were performed to check for differences between different strategies.

ii. For each training strategy, pairwise comparisons were performed between Sinc-ShallowNet-v2 and the other two CNNs that inspired Sinc-ShallowNet-v2 design: the winning solution of the IFMBE 2019 challenge based on EEGNet [50], representing the SOA for P300 decoding, and Sinc-ShallowNet [52]. A total of 6 tests were performed to check for differences between the proposed ICNN and the other two CNN-based solutions.

iii. Pairwise comparisons were performed between Sinc-ShallowNet-v2 trained with TL-WS strategy and trained with WS-WS strategy for each number of calibration blocks used for tuning the models. This was done to test the benefits of pre-training the models on other subjects compared to training models from scratch. A total of 10 tests were performed to check for differences.

All previous pairwise comparisons were performed using Wilcoxon signed-rank tests and false discovery rate correction at $\alpha = 0.05$ using the Benjamini–Hochberg procedure [75] to correct for multiple tests.

2.8.2. ASD clinical scores

Finally, we investigated whether the relevance of spatial features related to the most relevant spectral features of the P300 response, as learned by the ICNN, was related to ASD clinical scores (ADOS, ADI-R, and IQs, see Section 2.1). As the cluster analysis highlighted a strong contribution of parietal sites across clusters (see Section 3.3), $\bar{\theta}_{spat}^{(s)}$ was averaged across P3, Pz and P4 sites (thus, resulting in a scalar value for each subject s). Then, the Pearson correlation coefficient between this ICNN+ET-derived measure and each ASD clinical score was computed. Lastly, we performed a similar correlation analysis by using a measure derived

from subject-specific evoked potentials, to investigate whether useful ASD neural signatures were enhanced by the proposed ICNN+ET analysis or they emerged already from evoked potentials. To this aim, for each subject, EEG trials of the test set containing the P300 response were band-pass filtered in the range $[f_0^*, f_1^*]$ and were averaged together. Then, the resulting evoked potential was averaged across P3, Pz and P4 sites and across the time samples where the P300 response was stronger, i.e., between 400-600 ms, see Borra et al. [43]. Thus, after this averaging procedure, a scalar value was obtained for each subject s. Then, the Pearson correlation coefficient between this ERP-derived measure and each ASD clinical score was computed.

3. RESULTS

First, this section describes the results of AHPS via BO and the obtained decoding performance; then, it shows the results obtained via the proposed ICNN+ET algorithm to analyze the neural signatures corresponding to attentional P300 in ASD.

3.1. Hyper-parameter search via Bayesian optimization

The hyper-parameter configurations and their importance scores were separately obtained for each tested training strategy (WS-WS, WS-CS, LOSO). Figure 4 shows the probability to obtain a specific hyper-parameter value (P) among the admitted ones (Figure 4a) and the hyper-parameter importance scores z (Figure 4b) for each training condition.

The more frequent design of the ISS block included a lower number of band-pass filters in WS-WS than WS-CS and LOSO (i.e., ${K_0}^{ISS} = 8$ vs. ${K_0}^{ISS} = 16$) and the same number of spatial filters for each band-pass filter across training conditions (i.e., ${D_1}^{ISS} = 4$). Among the regularization techniques employed, spatial kernel max-norm constraint was applied more frequently in the LOSO strategy, while was progressively applied less frequently (i.e., ${c_1}^{ISS} = None$) moving to WS-CS and WS-WS strategies. The FST block was mostly designed learning an equal representation over the input ISS feature maps (i.e., ${K_0}^{FST} = {K_0}^{ISS} \cdot {D_1}^{ISS}$) for WS-WS and LOSO strategy, while equal representation and compressed representation (with ${K_0}^{FST} = 2$) were almost evenly frequent in WS-CS strategy. Interestingly, the most frequent temporal window in which features were learned in the FST block was approx. of 600 ms for all training strategies ($= {F_0}^{FST}[1]/\,(sf//4)$, where ${F_0}^{FST} = (1{,}21)$ and $sf = 128$ Hz), and this could be related to the temporal dynamics of the P300 response. Regarding the FC block, weights were always constrained in case of LOSO strategy mainly with low upper bound (c_0^{FC} = 0.25 or 0.5), while the probability distribution of this hyper-parameter appeared almost uniform for the other strategies.

Dropout (both in the ISS and in the FST blocks) was mostly absent in LOSO strategy while it was included in the other strategies with a more frequent dropout probability of $p = 0.25$. Lastly, higher learning rates were adopted in WS-WS strategy, i.e., $lr \in [5 \cdot 1e-3, 1e-2)$, compared to the other strategies where $lr \in [5 \cdot 1e-4, 1e-3)$.

As shown in Figure 4b, the learning rate resulted the most important hyper-parameter to optimize in case of WS-WS strategy, while the other hyper-parameters resulted almost equally important. Regarding the WS-CS strategy, the number of band-pass filters (${K_0}^{ISS}$) and of spatial filters for each band-pass filter (${D_1}^{ISS}$) were the most important hyper-parameters, while regarding the LOSO strategy, the number of band-pass filters (${K_0}^{ISS}$) and the dropout probability (p) were the most important ones.

Table 5 lists the model size (expressed as the number of trainable parameters) and training time (expressed in s/epoch) of the models. As expected, WS-WS models were the lightest (overall and within each block) and fastest to train, while LOSO models were the heaviest (overall and within each block) and slowest to train. Interestingly, the proportion of trainable parameters across blocks between WS-WS and WS-CS models did not change substantially (ISS: 28%, 28%; FST: 62%, 63%; FC: 10%, 9%, respectively for WS-WS and WS-CS); compared to WS-WS and WS-CS strategies, LOSO models presented a higher concentration of the trainable parameters (i.e., higher capacity) in the FST block (i.e., ISS: 12%; FST: 80%; FC: 8%).

3.2. Decoding performance

3.2.1. Comparison with EEGNet and Sinc-ShallowNet in different training strategies

The performance metrics averaged across subject-specific and session-specific test sets (see Section 2.8.1) are reported here. Figure 5 displays the performance metrics of the Bayesian-optimized EEGNet, the Bayesian-optimized Sinc-ShallowNet, and the Bayesian-optimized

Sinc-ShallowNet-v2. Sinc-ShallowNet-v2 scored significant different performance across the three training strategies ($p < 0.001$, Friedmann test). As expected, post-hoc comparisons highlighted that the network performed significantly worse in LOSO strategy than both WS-WS and WS-CS ($p < 0.001$ for both comparisons) strategies. Lastly, the network in WS-CS performed significantly better than WS-WS strategy ($p = 2.3 \cdot 1e - 3$). Sinc-ShallowNet-v2 scored object-level accuracies of 88.5±2.3%, 91.5±1.8%, 76.0±3.2%, compared to 85.5±2.5%, 91.7±1.6%, 76.3±3.2%, of EEGNet, and 76.1±3.9%, 84.0±2.9%, 67.6±3.6%, of Sinc-ShallowNet, respectively in WS-WS, WS-CS and LOSO strategies. Remarkably, despite the reduction in model capacity by re-parametrizing part of the architecture to design an interpretable spectral and spatial feature extractor, Sinc-ShallowNet-v2 achieved comparable performance compared to EEGNet for P300 decoding in WS-CS and LOSO strategies ($p = 0.84$), while significantly outperformed it in the WS-WS strategy ($p = 0.01$). Lastly, it is worth noticing that the updated version of Sinc-ShallowNet-v2 significantly outperformed Sinc-ShallowNet in all training strategies.

3.2.2. Transfer learning

Transfer learning was adopted to test the capability of models to transfer the knowledge from other subjects to a new one, for a more practical usage of the decoder in a BCI for ASD treatment (i.e., reduction of BCI calibration times). To this aim, Figure 6 shows the accuracy obtained while transferring the knowledge from the other participants to the held-out participant (TL-WS strategy, see Section 2.5.1-iv), using a progressively increasing number of calibration blocks of the held-out participant (from 2 to 20, where 20 corresponds to the entire subject-specific calibration set, see Section 2.5.1-i, iv). The WS-WS performance are reported together with TL-WS performance, to highlight the potential benefit of transfer learning compared to a random initialization (proper of the WS-WS strategy). Remarkably, transfer

learning was found to be beneficial with significant improvements ($p < 0.01$) for all the tested number of calibration blocks; using just 2 calibration blocks per subject, TL-WS reached an object-level performance as high as 80.3±3.0% with an average improvement as high as 37.1% compared to WS-WS.

3.3. Subject-specific ASD neural signatures related to P300

Figure 7 (top panel) shows the spectral relevance (average ± standard error of the mean across subjects). The frequency range retaining most of the relevance (≥ 0.75) resulted $[f_0^*, f_1^*] = [2, 5.8]$ Hz. Then, by clustering the relevance of spatial features related to this frequency range, two clusters were obtained and their $\bar{\theta}_{c,spat}$ are reported in the bottom panel of Figure 7 (left). Each of these cluster-level representations evidenced a peculiar strategy at the level of scalp that the system automatically learned for P300 decoding. Cluster 0 showed a strong and wide contribution of parietal sites (peaking at Pz) symmetrical across hemispheres, while cluster 1 – that resulted the most populated cluster (with 10 observations) – showed a strong right-lateralized contribution at parietal sites (P4). Lastly, in addition to these clusters, two outliers were detected, with a different spatial relevance.

The results of the statistical analysis conducted on the relevance of spatial features ($\bar{\theta}_{spat}^{(s)}$) at parietal sites is reported in Figure 8. Among all clinical scores, ADOS scores (both A-Communication, B-Social Interaction and A+B-Communication-Social Interaction scores) were the ones that mostly correlated to the ICNN+ET-derived measures, with high (r=0.770, r=0.657 and r=0.801, respectively) and significant ($p < 0.01$) positive correlations. Conversely, by using ERP-derived measures no significant correlations with clinical scores were found.

4. DISCUSSION

4.1. Hyper-parameter search

Depending on the training strategy adopted, AHPS selected different hyper-parameter configurations for Sinc-ShallowNet-v2. A peculiar trend of the probability distributions related to ${K_0}^{ISS}$, ${D_1}^{ISS}$, ${K_0}^{FST}$, ${F_0}^{FST}$ (see Figure 4a) was observed while moving from the WS-WS to WS-CS and LOSO strategies. In particular, the probability distributions of these hyper-parameters progressively moved from distributions more focused on small values (WS-WS) to distributions more focused on higher values (LOSO) among the admitted ones (see Table 4 for the admitted values), with the WS-CS strategy exhibiting an intermediate behavior in between the previous strategies. As an example, the probability distributions related to ${K_0}^{ISS}$ moved from a distribution more focused on lower values including also the lowest admitted value (${K_0}^{ISS} = 4$), in the WS-WS strategy, to a distribution entirely focused on the highest admitted value ($P({K_0}^{ISS} = 16) = 1$), in the LOSO strategy. That is, Sinc-ShallowNet-v2, moving from WS-WS to WS-CS and LOSO strategies, required progressively more filtered versions of the input EEG trial (${K_0}^{ISS}$), more electrode combinations tied to each band-pass filtered representation (${D_1}^{ISS}$), more temporal representations to learn (${K_0}^{FST}$) and a wider window size in which learn these representations (${F_0}^{FST}$).

Among regularization techniques, the needing of dropout was stronger in WS-WS strategy, using mostly $p = 0.5$ and $p = 0.25$, with a progressively reduction in WS-CS and LOSO strategies, up to $p = 0$ corresponding to no dropout applied (denoted by $p = None$). That is, the smaller and less variable (in terms of intra-subject and inter-subject variabilities) the dataset, the higher the needing of dropout to provide a better generalization. Similar to other regularization techniques, dropout acts reducing the algorithm capacity. Thus, from the previous considerations about ${K_0}^{ISS}$, ${D_1}^{ISS}$, ${K_0}^{FST}$, ${F_0}^{FST}$ and from the consideration on p, as the training strategy involves a more challenging decoding task, the architecture needs more

capacity to solve the task with high performance. This is particularly relevant in strategies such as WS-CS and LOSO, where single-trial EEG decoding is performed using signals collected across several recording sessions (training set with high intra-subject variability) and across several subjects and sessions (training set with high intra-subject and inter-subject variability), respectively. Interestingly, this result was confirmed by looking to the hyper-parameter importance scores (see red and light blue bars in Figure 4b), where ${K_0}^{ISS}$ and ${D_1}^{ISS}$, and ${K_0}^{ISS}$ and p were the more important hyper-parameters to optimize for the WS-CS and LOSO strategies, respectively. Furthermore, the need of an increased capacity as observed in the structural hyper-parameters of Sinc-ShallowNet-v2 is reflected onto its model size (see Table 5), resulting in 1207, 1655, and 4638 (on average) trainable parameters for WS-WS, WS-CS, and LOSO, respectively. In addition, while progressively increasing the variability in the training examples, the capacity of the model increased differently across the network. While increasing the intra-subject variability (WS-CS) the number of trainable parameters increased by the same proportion across ISS, FST, FC blocks compared to WS-WS, suggesting that increasing the capacity equally across the network could help addressing an increased intra-subject variability in the training examples. Conversely, while also increasing the inter-subject variability (LOSO), the network needed more abstract temporal features to learn in deeper layers compared to the models trained in the other strategies, i.e., 80% (vs. ~60%) of the parameters were in the FST block (see Table 5). This suggests that increasing the capacity at the deeper layers of the architecture could help addressing an increased inter-subject variability.

Lastly, the probability distribution of the learning rate was characterized by being more biased towards higher learning rates in the WS-WS strategy and towards lower learning rates in the LOSO strategy, with the WS-CS representing an intermediate condition. That is, as the decoding task became less challenging, i.e., moving from the LOSO to WS-WS strategies, a

higher learning rate resulted beneficial. As for the previous hyper-parameters, this result was confirmed by looking to the importance scores for the WS-WS strategy (Figure 4b), where the learning rate was the most important hyper-parameter to search.

4.2. Decoding performance

In a preliminary analysis, we investigated the utility of AHPS, by comparing the performance of architectures optimized via BO for each session or each subject instead of using a fixed configuration. Results showed that the use of a fixed configuration provided significantly lower accuracies (see Section 2 of Supplementary Materials). This suggests that when a new subject or a subject in a new session approaches the BCI, BO should be performed again to achieve higher performance.

Object-level accuracy varied using different training strategies. Despite the larger within-subject variability in the input distributions (due to the involvement of many recording sessions), CNNs trained with WS-CS were able to significantly outperform CNNs trained with WS-WS, although the improvement is modest. A possible explanation for this result may be that some subjects exhibited high variability even intra-session, leading to smaller performance in WS-WS than in WS-CS (subjects 1,13,14 in Supplementary Figure 1), on average. Indeed, in these cases WS-CS took great advantage of the larger training set across all sessions, allowing object-level accuracy to be increased up to 8.6% compared to WS-WS. However, WS-WS strategy reached high performance (i.e., >90%) with small variance in most of the subjects (subjects 2, 4, 6, 7, 8, 10, 11, 15 in Supplementary Figure 1), suggesting consistency in within-subject responses both intra- and inter-session. In these cases, WS-CS can still benefit from the 7-time increased training set, slightly improving accuracy compared to WS-WS. Lastly, further increasing the decoding difficulty, by including the cross-subject variability into

the training distributions, the challenge represented by the LOSO strategy was reflected into the lowest performance across all the training strategies.

4.2.1. Comparison with EEGNet and Sinc-ShallowNet in different training strategies

Sinc-ShallowNet-v2 significantly outperformed the solution based on EEGNet [50] in the WS-WS strategy, while it was comparable in the other strategies (WS-CS and LOSO). Due the introduction of the temporal sinc-convolution in Sinc-ShallowNet-v2, the number of trainable parameters to perform band-pass filtering were only 1.5% of the ones used in EEGNet that exploits conventional temporal convolution. This reduction of trainable parameters provided a beneficial effect when using the training strategy involving the most compact training set (WS-WS), in which a limited algorithm capacity resulted optimal, while no effect was observed in the conditions with larger and more variable training sets (WS-CS and LOSO), requiring a higher algorithm capacity). Therefore, the introduction of the interpretable spectral and spatial feature extractor, not only resulted beneficial for feature interpretability, but also did not negatively affect the performance. Indeed, surprisingly the performance of Sinc-ShallowNet-v2 was comparable or even significantly higher, while providing at the same time a straightforward access to the spectral and spatial features. Furthermore Sinc-ShallowNet-v2 significantly outperformed the previous version Sinc-ShallowNet [52] in all training conditions. These results suggest that the inclusion of the FST block in Sinc-ShallowNet-v2, by learning temporal features on the output of ISS block, enables the extraction of more relevant P300-related features than the ISS block only, easing the discrimination of the P300 response. In particular, the FST block may better cope with the large variability in time of P300; the lower performance of Sinc-ShallowNet is then ascribable to the lack of this block and to a reduced ability to catch intra-session, inter-session and inter-subject variability.

4.2.2. Transfer learning

Transfer learning could be used (TL-WS strategy) to reduce the calibration time on a new subject recorded in a new recording session (i.e., requiring WS-WS decoding). Indeed, results reported in Figure 6 highlight a significant benefit in transferring the knowledge from other participants to a new one recorded in a new session, for each calibration set adopted. Remarkably, the performance improvement was particularly relevant (on average 37.1%) in the lowest data regime (i.e., 2 calibration blocks corresponding to 160 trials), suggesting that the proposed interpretable approach was also capable of transferring the knowledge from other participants enabling its usage with extremely compact-sized calibration sets. Lastly, transfer learning not only improved the performance obtained in a WS-WS strategy using a small portion of calibration blocks, which could be useful in practice to reduce calibration time, but also improved the performance when using all 20 the calibration blocks (i.e., 1600 trials).

4.3. Subject-specific ASD neural signatures related to P300

The results obtained by analyzing the neural signatures related to P300 response via the proposed ICNN+ET algorithm suggest that the more relevant (and shared across subjects) features to distinguish P300 were obtained by filtering EEG signals including the frequency range [2, 5.8] Hz in the passband. This is in line with other studies performing P300 classification both on healthy and ASD people, where features from EEG signals were computed mainly in [2,4] Hz, [4,8] Hz, [2,8] Hz and [2,12] Hz frequency ranges [33–35]. In these previous studies, EEG signals recorded from ASD people were pre-processed based on other P300 decoding approaches validated on healthy subjects, adopting fixed cutoff frequencies. Conversely, here the learning system was left free to explore all frequency contents, adapting them to EEG signals of ASD people in an unbiased way. Indeed, some learned filters included also higher cutoff frequencies too, but only the filters containing lower

frequency content resulted more important to provide the correct discrimination of the P300 response. Furthermore, the spectral relevance peaked approximatively in [2,4] Hz, matching the findings of the wavelet analysis conducted by Demiralp et al. [36] where single EEG trials were successfully decoded from features designed in delta ([2,4] Hz) frequency range. These results suggest that future studies on traditional machine learning applications for P300 decoding in ASD, could design filter banks focusing especially on the interval [2, 5.8] Hz, i.e., in the delta and theta ranges.

The spatial relevance showed a strong right-lateralization for most of the observations (see cluster 1 in Figure 7). This indicates that the visuo-spatial sensory processing during the BCI task involved mainly the right hemisphere, differing from the classic P300 scalp distribution. In the literature, there is evidence of hemispheric asymmetries underlying social perception [76], e.g., right-lateralization while processing facial expressions related to emotions [77–79]. Furthermore, a right-lateralization was found also in the P300 response in Amaral et al. [80] and was associated to the high-level characteristics in the realism of the animated paradigms provided via the virtual environment (e.g., reflexive attention generated by social gaze orientation). The BCI intervention investigated in this study was based on a visuo-spatial oddball task which exploits a complex animated paradigm (see Section 2.1) and here this right-lateralization emerged in the electrode discriminatory power related to P300, as learned by the ICNN. Thus, our results further substantiate the idea that neural signatures related to social perception (e.g., perception of gaze, faces and related gestures) are characterized by a peculiar right-hemispheric asymmetry. Two clusters of spatial relevance for the P300 discrimination were obtained (see Section 3.3), potentially highlighting a modulation of the right-hemispheric asymmetry: from an asymmetric involvement of mainly P4 (cluster 1, the most populated – 10 subjects) to the symmetric involvement of mainly Pz, P3 and P4 (cluster 0 – 3 subjects).

From the correlation analysis between the ASD clinical scores and spatial relevance ($\bar{\theta}_{spat}^{(s)}$) of parietal electrodes (P3, Pz, P4), a strong positive and significant correlation with ADOS scores was observed (see Figure 8). This result is particularly interesting, as ADOS together with ADI-R are two important assessment tools used to characterize and diagnose ASD [81]. In addition, recent results [82] suggested that ADOS-related scores are more useful than the combination of ADOS and ADI-R scores to diagnose ASD, which is in line with our finding. Notably, at variance with the results obtained with the ICNN+ET-derived measure, no significant correlations were found between ASD clinical scores and the P300 ERP-derived measure.

Overall, we believe that the proposed method presents some points of novelty and strength, compared to previous CNN-based approaches [41,43,45,48–50,57,58], with perspective not only for practical engineering applications but also for theoretical neuroscience knowledge.

First, from a decoding perspective, we optimized our CNN-based decoders in their structure, by tuning their hyper-parameters to achieve higher decoding performance, and this was done separately for different training strategies. To the best of our knowledge, this is the first time that such analysis is performed on P300 CNN-based decoders, using an automatic strategy (Bayesian optimization) and in a training-specific way. Indeed, previous studies mainly proposed CNN structures with hyper-parameters manually selected, without discussing the specific choice [41,45,48–50,57], or performing sensitivity analysis only on few hyper-parameters and for a single training strategy [43,58]. A significant result of our analysis is that for each investigated training strategy (WS-WS, WS-CS, LOSO), different optimal hyper-parameters were found, depending on the variability incorporated into the training examples, and that a different architecture with a proper capacity should be adopted depending on the specific real-life scenario in which the P300-BCI is used. Thus, the results obtained in this study may help researchers in the design of CNN decoders for P300-based BCI applications,

by driving them in the choice of the more appropriate structure that takes into account the adopted training strategy and the level of variability inside the training examples.

Second, we introduced an interpretable layer in our CNN decoders, using convolutional kernels defined as function of only two and directly interpretable parameters per kernel. The performance evaluation of our decoders compared to EEGNet (which is the state-of-the-art for P300 decoding), indicates that this modification, while decreasing the number of trainable parameters, did not reduce the decoding accuracy. Thus, a CNN can be made intrinsically more interpretable without losing decoding capabilities. ICNNs are receiving growing interest in EEG applications, moving beyond the use of CNNs as black-box decoders only, towards a neurophysiological interpretation of the learned features. In this regard, it is highly important to fist verify the decoding performance of ICNNs, to ensure reliability of the learned features as discriminative of the process under investigation, otherwise the neurophysiological interpretation of these features would be unreliable. So far, ICNNs have been adopted in literature to decode and analyze motor states [52,53] and not the P300 response and not even in neurological disorder conditions such as ASD. Thus, the present study serves also for validating ICNNs for P300 decoding in ASD subjects.

Third and related to previous point, by taking advantage of the interpretability embedded into the proposed ICNN, we have formalized an analysis workflow that combines our ICNN with an explanation technique (ET) to derive novel ASD biomarkers related to P300 by exploiting the knowledge learned by the ICNN. The ICNN+ET combination was able to capture and enhance, better than a canonical ERP analysis, meaningful spectral and spatial characteristics underlying visuo-spatial sensory and attentional processing that are related to autism. The enhancement of P300-related features is in line with the results of our recent study [57] on healthy subjects, where representations derived from a CNN+ET framework better highlighted P300 subcomponents (i.e., P3a and P3b) from single trial level, compared to ERP

canonical analysis. The remarkable aspect is that not only the ICNN-derived ASD markers obtained here have a straightforward neurophysiological interpretation, but they also appeared more sensitive than markers derived from traditional ERP-analysis.

Of course, the present study has also some limitations that can be the subject of future improvements, and can foster further investigations along this line of research.

First, the optimal models obtained via Bayesian optimization may be task-related i.e., may result optimal only for EEG decoding of P300 response elicited through the specific kind of stimuli applied during the adopted oddball paradigm (visual stimuli presented inside a virtual environment). It would be of great value to test these same models on other BCI-based P300 paradigms, i.e., using a different type of visual stimuli or a different sensory modality (acoustic stimuli), or even on BCI paradigms based on event-related potentials other than P300. This would test the ability of the proposed models to generalize across different BCI paradigms. This kind of analysis is very important to promote a more efficient use of decoders, limiting the proliferation of novel decoders for every new paradigm and task, but it is only rarely performed (see [52] as a useful example of this analysis).

Another limitation concerns the limited number of subjects (15 subjects) included in the examined dataset. Thus, the biomarkers derived from the ICNN+ET analysis that were found to be correlated with ASD social impairment scores, although promising, require a more extensive validation on a larger set of participants. Furthermore, in future research, the workflow ICNN+ET, applied here to study P300 in autism, could be focused on analyzing other ERP components in autism, and/or in other neurological and neurodevelopmental disorders that involve ERP abnormalities (e.g., schizophrenia, depression, etc.) [83].

5. CONCLUSION

In conclusion, in this study we investigated the Bayesian-optimized design of an ICNN in different training strategies when decoding the P300 response for BCI intervention in ASD. While performing AHPS, models were found needing more capacity and lower learning rates to decode EEG signals when more variability was included in the training distribution (i.e., including progressively intra-subject and inter-subject variabilities), i.e., depending on the practical scenario in which the decoder is used in the BCI system. Despite its interpretable and lightweight nature, the proposed ICNN performed as well as EEGNet, even significantly outperforming it in the strategy with the lowest variability in the training examples (within-subject and within-session). Furthermore, Sinc-ShallowNet-v2 significantly outperformed the previous Sinc-ShallowNet design. Lastly, transferring the knowledge from other users to a new one proved to lead to a substantial reduction of calibration times. All these results contribute to the development of optimal decoders for P300-based BCI for ASD interventions, by specifically improving CNN designs, performance, and calibration times. Furthermore, in this study we leveraged the interpretable nature embedded into the ICNN to design an ICNN+ET algorithm for the analysis of the visuo-spatial P300 in ASD in the frequency and spatial domains. The analysis on spectral features matched known P300-related correlates, and while analyzing spatial features, a right-hemispheric asymmetry was found, in line with the literature of social perception. The modulation of this asymmetry, as provided by the ICNN+ET analysis, was found to be correlated to ADOS scores, while no significant correlations with any clinical score were found by using a simpler ERP analysis. This suggests that the ICNN+ET algorithm was capable of better characterize and enhance useful ASD-related features than a canonical analysis. In the future, the analysis on optimal ICNN designs for P300 decoding may be extended to other oddball recording paradigms involving different stimuli properties. Furthermore, the proposed ICNN+ET combination could be applied on more subjects for

1 further validation, generalized to other ERP components and neurological disorders to study
2 alterations in ERP components in a data-driven way, and also possibly extended on other
3 recording modalities of neural activity, e.g., magnetoencephalography.

ACKNOWLEDGMENTS

The authors declare no conflict of interest.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN V used for this research. The provider was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

The dataset adopted was collected in the scope of the European project BRAINTRAIN-FP7-HEALTH-2013-INNOVATION-1-602186.

REFERENCES

[1] American Psychiatric Association 2013 *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association)

[2] Baron-Cohen S 1989 Perceptual role taking and protodeclarative pointing in autism *British Journal of Developmental Psychology* **7** 113–27

[3] Baron-Cohen S, Baldwin D A and Crowson M 1997 Do Children with Autism Use the Speaker's Direction of Gaze Strategy to Crack the Code of Language? *Child Development* **68** 48

[4] Swettenham J, Baron-Cohen S, Charman T, Cox A, Baird G, Drew A, Rees L and Wheelwright S 1998 The Frequency and Distribution of Spontaneous Attention Shifts between Social and Nonsocial Stimuli in Autistic, Typically Developing, and Nonautistic Developmentally Delayed Infants *Journal of Child Psychology and Psychiatry* **39** 747–53

[5] Klin A, Jones W, Schultz R, Volkmar F and Cohen D 2002 Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism *Arch Gen Psychiatry* **59** 809

[6] Dawson G, Toth K, Abbott R, Osterling J, Munson J, Estes A and Liaw J 2004 Early Social Attention Impairments in Autism: Social Orienting, Joint Attention, and Attention to Distress. *Developmental Psychology* **40** 271–83

[7] Bakeman R and Adamson L B 1984 Coordinating Attention to People and Objects in Mother-Infant and Peer-Infant Interaction *Child Development* **55** 1278

[8] Charman T 1998 Specifying the Nature and Course of the Joint Attention Impairment in Autism in the Preschool Years: Implications for Diagnosis and Intervention *Autism* **2** 61–79

[9] Charman T 2003 Why is joint attention a pivotal skill in autism? ed U Frith and E L Hill *Phil. Trans. R. Soc. Lond. B* **358** 315–24

[10] Travers B G, Adluru N, Ennis C, Tromp D P M, Destiche D, Doran S, Bigler E D, Lange N, Lainhart J E and Alexander A L 2012 Diffusion Tensor Imaging in Autism Spectrum Disorder: A Review: Diffusion tensor imaging *Autism Res* **5** 289–313

[11] Minshew N J and Keller T A 2010 The nature of brain dysfunction in autism: functional brain imaging studies *Current Opinion in Neurology* **23** 124–30

[12] Townsend J, Westerfield M, Leaver E, Makeig S, Jung T-P, Pierce K and Courchesne E 2001 Event-related brain response abnormalities in autism: evidence for impaired cerebello-frontal spatial attention networks *Cognitive Brain Research* **11** 127–45

[13] Luck S J, Hillyard S A, Mouloua M, Woldorff M G, Clark V P and Hawkins H L 1994 Effects of spatial cuing on luminance detectability: Psychophysical and electrophysiological evidence for early selection. *Journal of Experimental Psychology: Human Perception and Performance* **20** 887–904

[14] Heinze H J, Luck S J, Mangun G R and Hillyard S A 1990 Visual event-related potentials index focused attention within bilateral stimulus arrays. I. Evidence for early selection *Electroencephalography and Clinical Neurophysiology* **75** 511–27

[15] Herrmann C S and Knight R T 2001 Mechanisms of human attention: event-related potentials and oscillations *Neuroscience & Biobehavioral Reviews* **25** 465–76

[16] Sokhadze E, Baruth J, Tasman A, Sears L, Mathai G, El-Baz A and Casanova M F 2009 Event-related Potential Study of Novelty Processing Abnormalities in Autism *Appl Psychophysiol Biofeedback* **34** 37–51

[17] Polich J 2007 Updating P300: An integrative theory of P3a and P3b *Clinical Neurophysiology* **118** 2128–48

[18] Azizian A and Polich J 2007 Evidence for Attentional Gradient in the Serial Position Memory Curve from Event-related Potentials *Journal of Cognitive Neuroscience* **19** 2071–81

[19] Mecklinger A and Pfeifer E 1996 Event-related potentials reveal topographical and temporal distinct neuronal activation patterns for spatial and object working memory *Cognitive Brain Research* **4** 211–24

[20] Cui T, Wang P P, Liu S and Zhang X 2017 P300 amplitude and latency in autism spectrum disorder: a meta-analysis *Eur Child Adolesc Psychiatry* **26** 177–90

[21] Ciesielski K T, Courchesne E and Elmasian R 1990 Effects of focused selective attention tasks on event-related potentials in autistic and normal individuals *Electroencephalography and Clinical Neurophysiology* **75** 207–20

[22] Courchesne E, Kilman B A, Galambos R and Lincoln A J 1984 Autism: Processing of novel auditory information assessed by event-related brain potentials *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* **59** 238–48

[23] Courchesne E, Lincoln A J, Kilman B A and Galambos R 1985 Event-related brain potential correlates of the processing of novel visual and auditory information in autism *J Autism Dev Disord* **15** 55–76

[24] Courchesne E, Lincoln A J, Yeung-Courchesne R, Elmasian R and Grillon C 1989 Pathophysiologic findings in nonretarded autism and receptive developmental language disorder *J Autism Dev Disord* **19** 1–17

[25] Dawson G, Finley C, Phillips S, Galpert L and Lewy A 1988 Reduced P3 amplitude of the event-related brain potential: Its relationship to language ability in austism *J Autism Dev Disord* **18** 493–504

[26] Verbaten M N, Roelofs J W, van Engeland H, Kenemans J K and Slangen J L 1991 Abnormal visual event-related potentials of autistic children *J Autism Dev Disord* **21** 449–70

[27] Kemner C, van der Gaag R J, Verbaten M and van Engeland H 1999 ERP differences among subtypes of pervasive developmental disorders *Biological Psychiatry* **46** 781–9

[28] Amaral C, Mouga S, Simões M, Pereira H C, Bernardino I, Quental H, Playle R, McNamara R, Oliveira G and Castelo-Branco M 2018 A Feasibility Clinical Trial to Improve Social Attention in Autistic Spectrum Disorder (ASD) Using a Brain Computer Interface *Frontiers in Neuroscience* **12** 477

[29] Amaral C P, Simões M A, Mouga S, Andrade J and Castelo-Branco M 2017 A novel Brain Computer Interface for classification of social joint attention in autism and comparison of 3 experimental setups: A feasibility study *Journal of Neuroscience Methods* **290** 105–15

[30] Yger F, Berar M and Lotte F 2017 Riemannian Approaches in Brain-Computer Interfaces: A Review *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 1753–62

[31] Saha S and Baumert M 2020 Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review *Front. Comput. Neurosci.* **13** 87

[32] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A and Yger F 2018 A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update *Journal of Neural Engineering* **15** 031005

[33] de Arancibia L, Sánchez-González P, Gómez E J, Hernando M E and Oropesa I 2020 Linear vs Nonlinear Classification of Social Joint Attention in Autism Using VR P300-Based Brain Computer Interfaces *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1869–74

[34] Krzemiński D, Michelmann S, Treder M and Santamaria L 2020 Classification of P300 Component Using a Riemannian Ensemble Approach *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1885–9

[35] Chatterjee B, Palaniappan R and Gupta C N 2020 Performance Evaluation of Manifold Algorithms on a P300 Paradigm Based Online BCI Dataset *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* IFMBE Proceedings vol 76, ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1894–8

[36] Demiralp T, Ademoglu A, Istefanopulos Y, Başar-Eroglu C and Başar E 2001 Wavelet analysis of oddball P300 *International Journal of Psychophysiology* **39** 221–7

[37] Bostanov V and Kotchoubey B 2006 The t-CWT: A new ERP detection and quantification method based on the continuous wavelet transform and Student's t-statistics *Clinical Neurophysiology* **117** 2627–44

[38] Bittencourt-Villalpando M and Maurits N M 2018 Stimuli and Feature Extraction Algorithms for Brain-Computer Interfaces: A Systematic Comparison *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 1669–79

[39] Bittencourt-Villalpando M and Maurits N M 2020 Linear SVM Algorithm Optimization for an EEG-Based Brain-Computer Interface Used by High Functioning Autism Spectrum Disorder Participants *XV Mediterranean Conference on Medical and*

Biological Engineering and Computing – MEDICON 2019 ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1875–84

[40] Miladinović A, Ajčević M, Battaglini P P, Silveri G, Ciacchi G, Morra G, Jarmolowska J and Accardo A 2020 Slow Cortical Potential BCI Classification Using Sparse Variational Bayesian Logistic Regression with Automatic Relevance Determination *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* IFMBE Proceedings vol 76, ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1853–60

[41] Simões M, Borra D, Santamaría-Vázquez E, GBT-UPM, Bittencourt-Villalpando M, Krzemiński D, Miladinović A, Neural_Engineering_Group, Schmid T, Zhao H, Amaral C, Direito B, Henriques J, Carvalho P and Castelo-Branco M 2020 BCIAUT-P300: A Multi-Session and Multi-Subject Benchmark Dataset on Autism for P300-Based Brain-Computer-Interfaces *Front. Neurosci.* **14** 568104

[42] Craik A, He Y and Contreras-Vidal J L 2019 Deep learning for electroencephalogram (EEG) classification tasks: a review *J. Neural Eng.* **16** 031001

[43] Borra D, Fantozzi S and Magosso E 2021 A Lightweight Multi-Scale Convolutional Neural Network for P300 Decoding: Analysis of Training Strategies and Uncovering of Network Decision *Frontiers in Human Neuroscience*

[44] Lindsay G 2020 Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future *Journal of Cognitive Neuroscience* 1–15

[45] Cecotti H and Graser A 2011 Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33** 433–45

[46] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *The Journal of Machine Learning Research* **15** 1929–58

[47] Ioffe S and Szegedy C 2015 Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift *Proceedings of the 32nd International Conference on Machine Learning* Proceedings of Machine Learning Research vol 37, ed F Bach and D Blei (Lille, France: PMLR) pp 448–56

[48] Manor R and Geva A B 2015 Convolutional Neural Network for Multi-Category Rapid Serial Visual Presentation BCI *Frontiers in Computational Neuroscience* **9** 146

[49] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *Journal of Neural Engineering* **15** 056013

[50] Borra D, Fantozzi S and Magosso E 2020 Convolutional Neural Network for a P300 Brain-Computer Interface to Improve Social Attention in Autistic Spectrum Disorder *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1837–43

[51] Zhang X, Yao L, Wang X, Monaghan J, McAlpine D and Zhang Y 2021 A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers *J. Neural Eng.* **18** 031002

[52] Borra D, Fantozzi S and Magosso E 2020 Interpretable and lightweight convolutional neural network for EEG decoding: Application to movement execution and imagination *Neural Networks* S0893608020302021

[53] Zhao D, Tang F, Si B and Feng X 2019 Learning joint space–time–frequency features for EEG decoding on small labeled data *Neural Networks* **114** 67–77

[54] Liang S, Hang W, Yin M, Shen H, Wang Q, Qin J, Choi K-S and Zhang Y 2022 Deep EEG feature learning via stacking common spatial pattern and support matrix machine *Biomedical Signal Processing and Control* **74** 103531

[55] Borra D, Fantozzi S and Magosso E 2020 EEG Motor Execution Decoding via Interpretable Sinc-Convolutional Neural Networks *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* ed J Henriques, N Neves and P de Carvalho (Cham: Springer International Publishing) pp 1113–22

[56] Vahid A, Mückschel M, Stober S, Stock A-K and Beste C 2020 Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control *Commun Biol* **3** 112

[57] Borra D and Magosso E 2021 Deep learning-based EEG analysis: investigating P3 ERP components *Journal of Integrative Neuroscience* **In press**

[58] Farahat A, Reichert C, Sweeney-Reed C and Hinrichs H 2019 Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization *Journal of Neural Engineering*

[59] Wu W, Zhang Y, Jiang J, Lucas M V, Fonzo G A, Rolle C E, Cooper C, Chin-Fatt C, Krepel N, Cornelssen C A, Wright R, Toll R T, Trivedi H M, Monuszko K, Caudle T L, Sarhadi K, Jha M K, Trombello J M, Deckersbach T, Adams P, McGrath P J, Weissman M M, Fava M, Pizzagalli D A, Arns M, Trivedi M H and Etkin A 2020 An electroencephalographic signature predicts antidepressant response in major depression *Nat Biotechnol* **38** 439–47

[60] Lord C, Risi S, Lambrecht L, Cook, Jr. E H, Leventhal B L, DiLavore P C, Pickles A and Rutter M 2000 The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism *Journal of Autism and Developmental Disorders* **30** 205–23

[61] Lord C, Rutter M and Le Couteur A 1994 Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders *J Autism Dev Disord* **24** 659–85

[62] Spek A A, Scholte E M and van Berckelaer-Onnes I A 2008 Brief Report: The Use of WAIS-III in Adults with HFA and Asperger Syndrome *J Autism Dev Disord* **38** 782–7

[63] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Human brain mapping* **38** 5391–420

[64] Ravanelli M and Bengio Y 2018 Speaker Recognition from Raw Waveform with SincNet *2018 IEEE Spoken Language Technology Workshop (SLT)* pp 1021–8

[65] Clevert D-A, Unterthiner T and Hochreiter S 2015 Fast and accurate deep network learning by exponential linear units (elus) *arXiv preprint*

[66] Chollet F 2016 Xception: Deep Learning with Depthwise Separable Convolutions *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1800–7

[67] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H and He Q 2020 A Comprehensive Survey on Transfer Learning *arXiv:1911.02685 [cs, stat]*

[68] Kingma D P and Ba J 2017 Adam: A Method for Stochastic Optimization *arXiv:1412.6980 [cs]*

[69] Bergstra J, Bardenet R, Bengio Y and Kégl B 2011 Algorithms for hyper-parameter optimization *Advances in neural information processing systems* **24**

[70] Akiba T, Sano S, Yanase T, Ohta T and Koyama M 2019 Optuna: A Next-generation Hyperparameter Optimization Framework *arXiv:1907.10902 [cs, stat]*

[71] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown 2014 An Efficient Approach for Assessing Hyperparameter Importance *Proceedings of the 31st International Conference on Machine Learning* ed Eric P. Xing and Tony Jebara (PMLR) pp 754–62

[72] Simonyan K, Vedaldi A and Zisserman A 2014 Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps *arXiv:1312.6034 [cs]*

[73] Campello R J G B, Moulavi D, Zimek A and Sander J 2015 Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection *ACM Trans. Knowl. Discov. Data* **10** 1–51

[74] McInnes L, Healy J and Astels S 2017 hdbscan: Hierarchical density based clustering *JOSS* **2** 205

[75] Benjamini Y and Hochberg Y 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289–300

[76] Brancucci A, Lucci G, Mazzatenta A and Tommasi L 2009 Asymmetries of the human social brain in the visual, auditory and chemical modalities *Phil. Trans. R. Soc. B* **364** 895–914

[77] Hartikainen K M 2021 Emotion-Attention Interaction in the Right Hemisphere *Brain Sciences* **11** 1006

[78] Demaree H A, Everhart D E, Youngstrom E A and Harrison D W 2005 Brain Lateralization of Emotional Processing: Historical Roots and a Future Incorporating "Dominance" *Behavioral and Cognitive Neuroscience Reviews* **4** 3–20

[79] Simões M, Monteiro R, Andrade J, Mouga S, França F, Oliveira G, Carvalho P and Castelo-Branco M 2018 A Novel Biomarker of Compensatory Recruitment of Face Emotional Imagery Networks in Autism Spectrum Disorder *Front. Neurosci.* **12** 791

[80] Amaral C P, Simões M A and Castelo-Branco M S 2015 Neural Signals Evoked by Stimuli of Increasing Social Scene Complexity Are Detectable at the Single-Trial Level and Right Lateralized ed S Ben Hamed *PLoS ONE* **10** e0121970

[81] Hiremath C S, Sagar K J V, Yamini B K, Girimaji A S, Kumar R, Sravanti S L, Padmanabha H, Vykunta Raju K N, Kishore M T, Jacob P, Saini J, Bharath R D, Seshadri S P and Kumar M 2021 Emerging behavioral and neuroimaging biomarkers for early and accurate characterization of autism spectrum disorders: a systematic review *Transl Psychiatry* **11** 42

[82] Kamp-Becker I, Tauscher J, Wolff N, Küpper C, Poustka L, Roepke S, Roessner V, Heider D and Stroth S 2021 Is the Combination of ADOS and ADI-R Necessary to Classify ASD? Rethinking the "Gold Standard" in Diagnosing ASD *Front. Psychiatry* **12** 727308

[83] Picton T W 1992 The P300 Wave of the Human Event-Related Potential *Journal of Clinical Neurophysiology* **9** 456–79

FIGURES

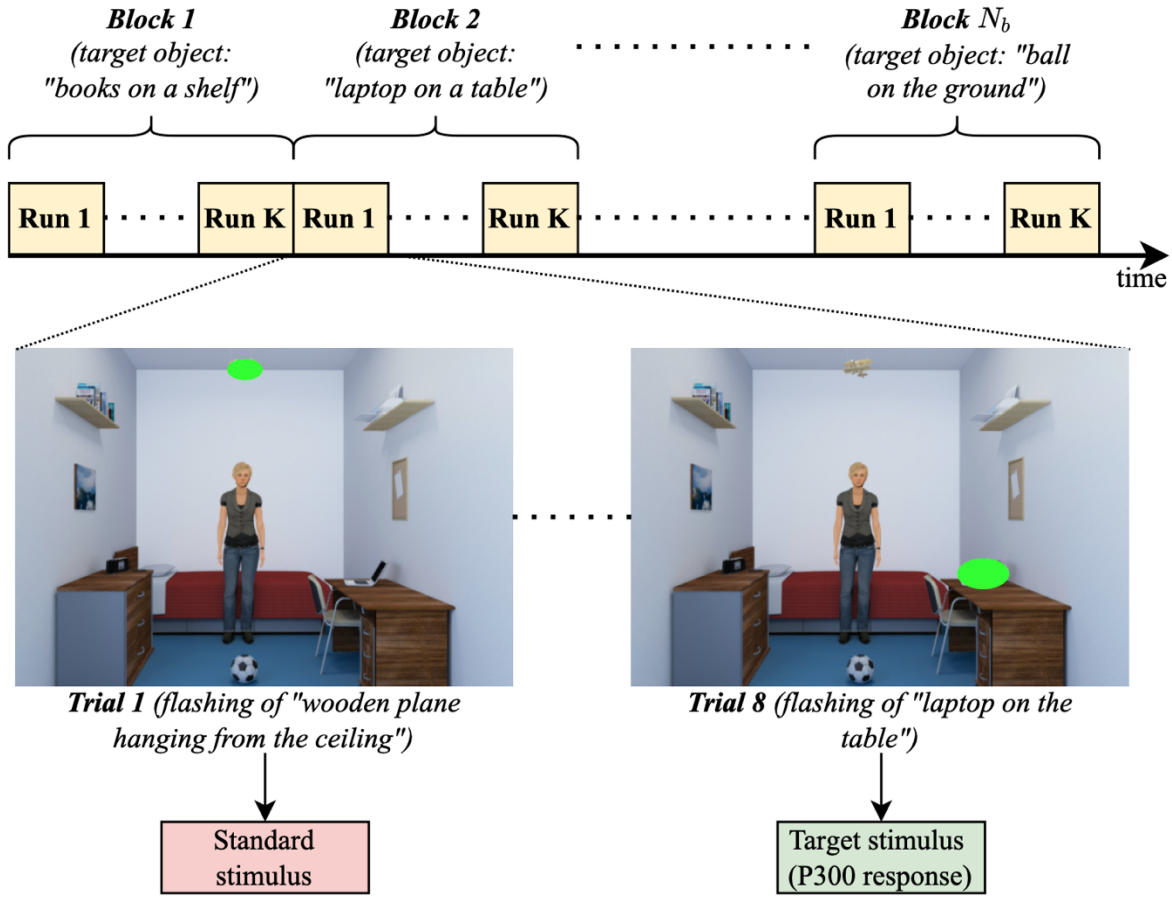

Figure 1 – Structure of the BCI paradigm in its division into blocks, runs, and trials. Each recording block was used to identify one of the eight target objects visualized in the virtual environment, with a total of $N_b = 20$ (calibration phase) and of $N_b = 50$ (online phase) blocks. Within each block, K runs were recorded, where $K = 10$ (calibration phase) and $K = 7.095$ (online phase) on average across subjects and sessions. In each run, the 8 objects randomly flashed (schematized by the green ellipses in the figure) and 8 EEG trials per run where recorded. Therefore, overall, within each recording session $N_b \cdot K \cdot 8$ EEG trials were recorded, corresponding 1600 trials in the calibration phase and 2838 trials on average in the online phase, respectively.

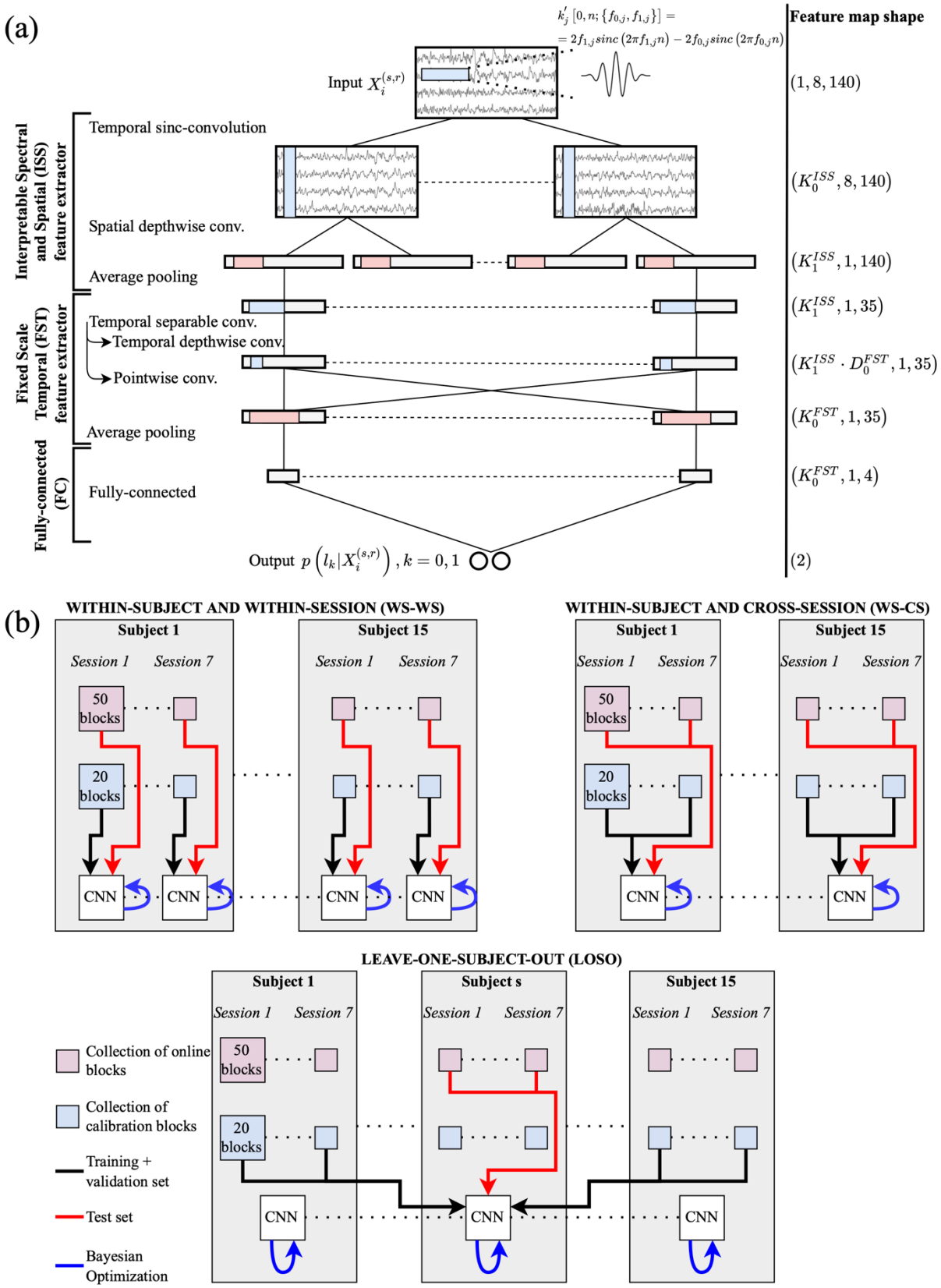

Figure 2 – Sinc-ShallowNet-v2 (a) and training strategies investigated while performing Bayesian optimization (b). In Figure 2a, blocks and main layers are listed on the left side. To keep the figure as clear and simple as possible, only the main trainable layers (i.e., convolutional and fully-connected layers) in addition to pooling layers (to highlight the temporal dimension reduction) were displayed. Boxes represent the output feature maps of

each layer, and colored rectangles represent convolutional (blue) and pooling (red) kernels. Tuples reported on the right side represent the shape of the feature maps. For all outputs except the last, tuples are composed by three numbers representing the number of the feature maps, the number of spatial samples and the number of temporal samples within each map, respectively. The input layer provides an output of shape $(1, C, T) = (1,8,140)$, as it just replicates the original input EEG trial with shape $(8,140)$, providing a single feature map as output. The temporal dimension changed from $T = 140$ to $T//4 = 35$ and to $T//32 = 4$ along the entire CNN due to the average pooling operations (where // indicates the floor division operator). See Section 2.2, 2.3 and Table 2 for further details. Figure 2b shows a schematic representation of how training, validation, and test examples (by means of black and red arrows) were sampled from calibration (blue boxes) and online (purple boxes) blocks recorded in the BCI paradigm when training decoders in Bayesian optimization in within-subject and within-session (WS-WS), within-subject and cross-session (WS-CS), and leave-one-subject-out (LOSO) training conditions. For brevity, in the LOSO strategy the aggregation across blocks is reported only for the s-th subject. See Section 2.1 and Section 2.5.1 for further details about the definition of recording blocks and the training strategies, respectively.

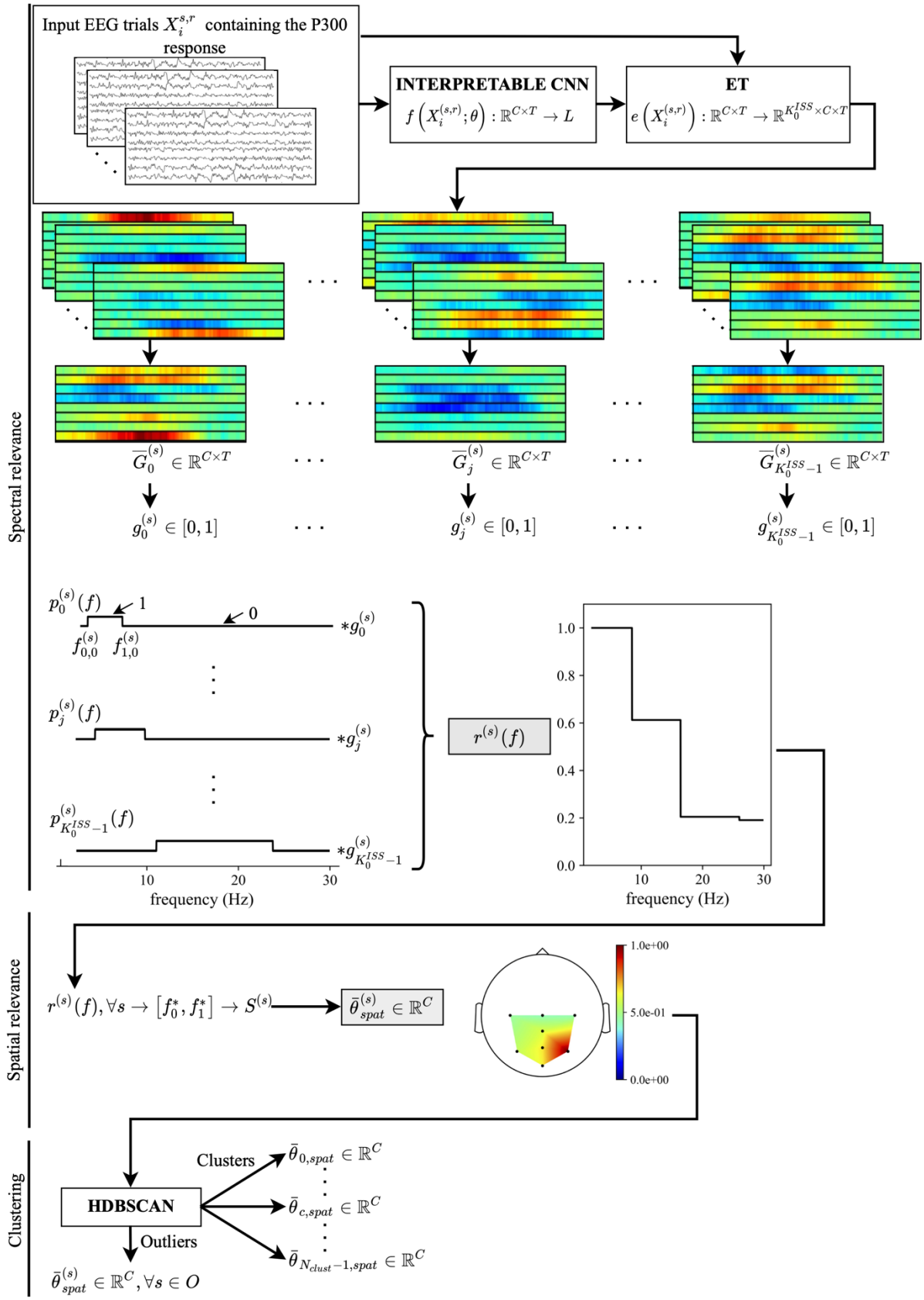

Figure 3 – Schematic representation of the algorithm based on ICNN+ET adopted to gain insights about the neural signatures of the visuo-spatial P300 correlate in autism in its three main steps: spectral relevance computation, spatial relevance computation, and clustering.

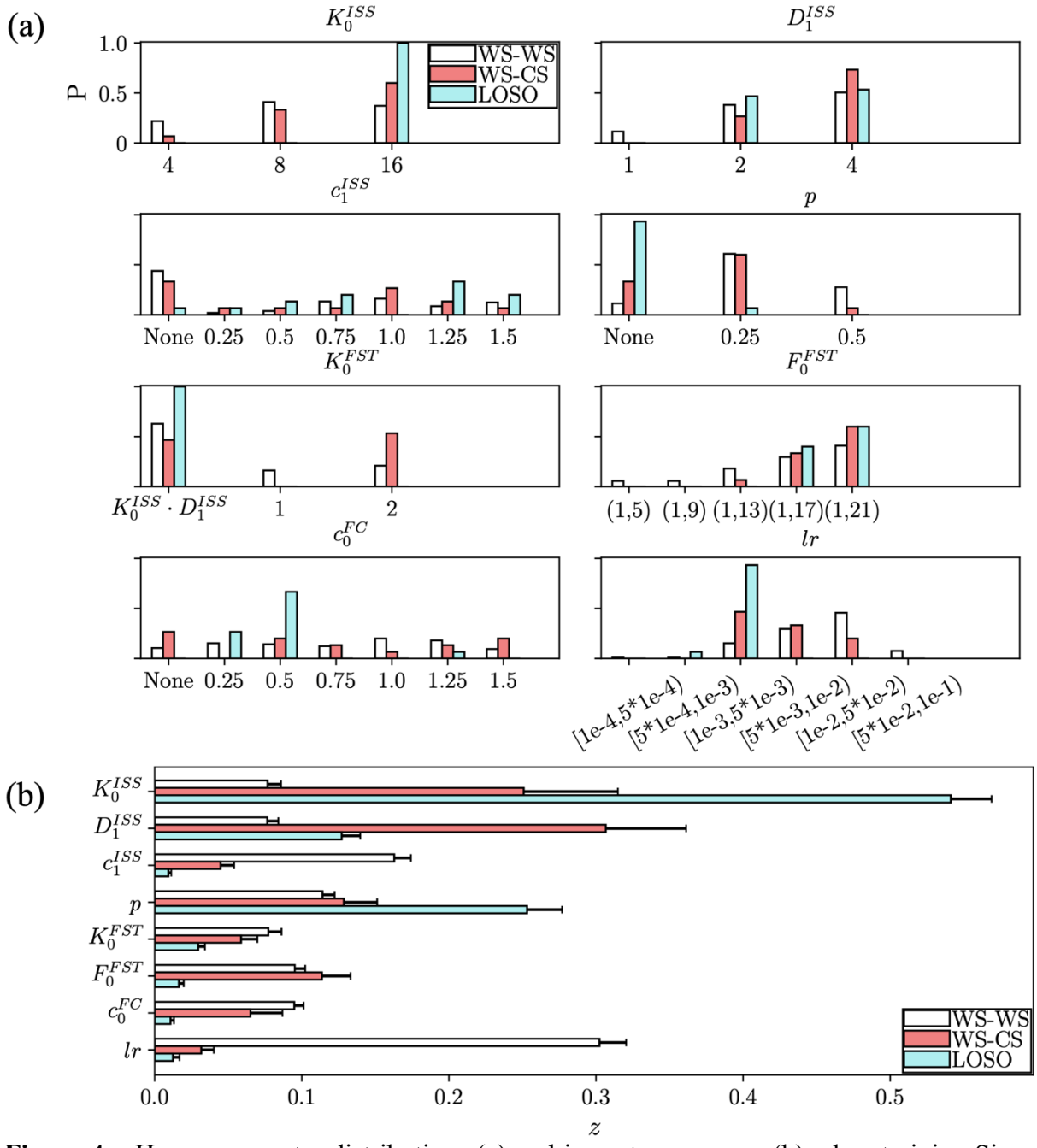

Figure 4 – Hyper-parameter distributions (a) and importance scores (b) when training Sinc-ShallowNet-v2 with within-subject and within-session (WS-WS), within-subject and cross-session (WS-CS) and leave-one-subject-out (LOSO) strategies. Hyper-parameter distributions were reported representing the probability (P) that a specific hyper-parameter value resulted as optimal via BO, and it is reported separately for the different training strategies. Hyper-parameter importance scores were reported in their mean values across decoders trained with each training strategy, separately. Bar heights represent the mean value, while the error bars represent the standard error of the mean.

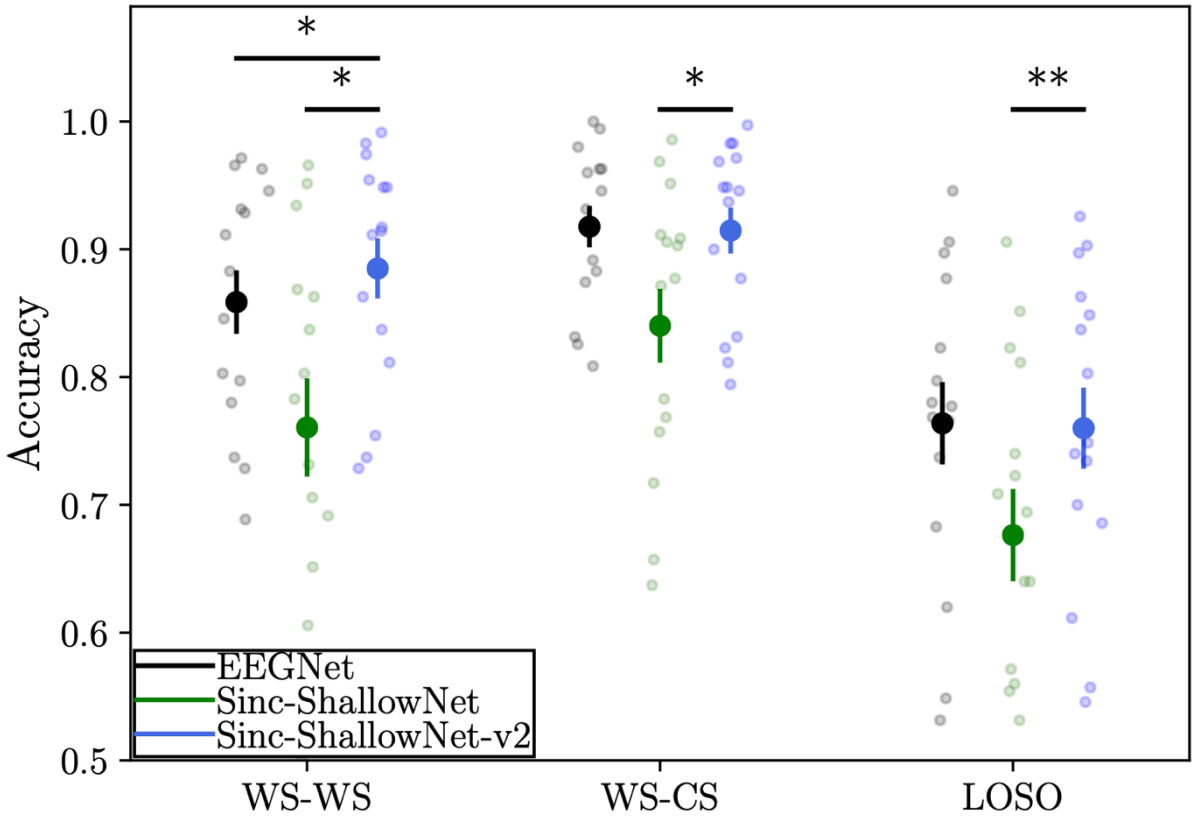

Figure 5 – Accuracies of the Bayesian-optimized EEGNet (black), Bayesian-optimized Sinc-ShallowNet (green) and Bayesian-optimized Sinc-ShallowNet-v2 (blue) scored in the WS-WS, WS-CS and LOSO strategies. Bigger dots represent the mean value across subjects, while the error bars represent the standard error of the mean. Smaller dots represent the accuracy scored for each subject (i.e., 15 data points). See part A of Table 3 for details about the number of examples defining the training, validation and test sets. Wilcoxon signed-rank tests were performed to compare the performance of the Bayesian-optimized Sinc-ShallowNet-v2 with the Bayesian-optimized EEGNet and with the Bayesian-optimized Sinc-ShallowNet, within each training strategy. P-values were corrected for multiple comparisons (6 in total) via the Benjamini–Hochberg procedure and significant comparisons are marked once applied the correction (*$p<0.05$, **$p<0.01$, ***$p<0.001$). See Section 2.8.1 for further details about the statistical tests.

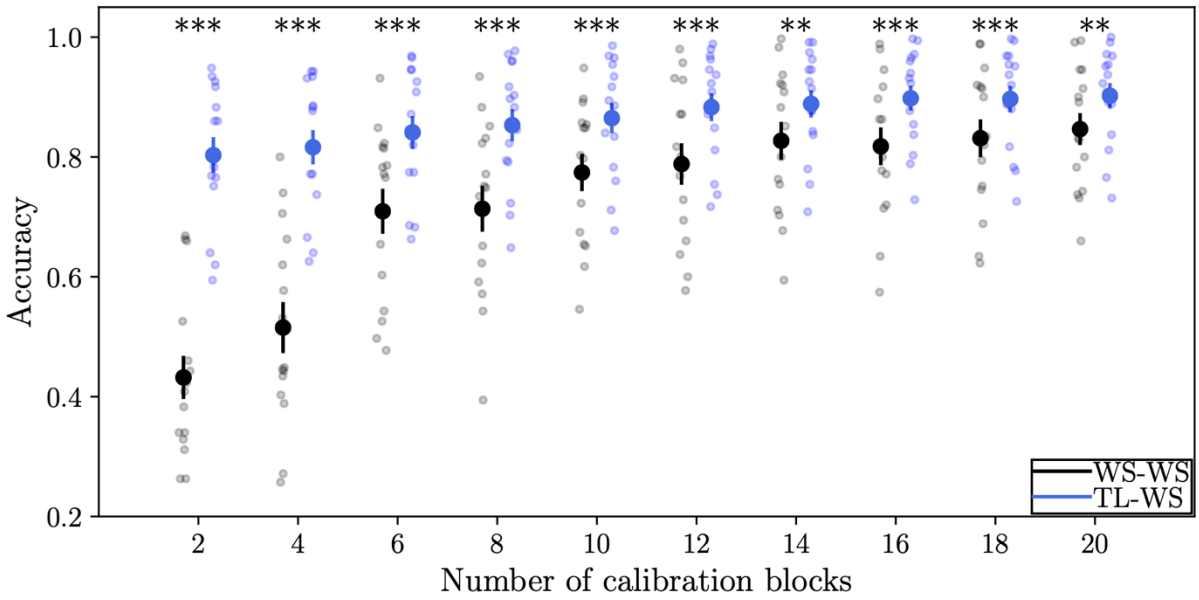

Figure 6 – Results of transfer learning as a function of the number of calibration blocks of the subject. Blue distributions show the accuracies obtained with the TL-WS models while transferring the knowledge from multiple subjects on a new subject (i.e., using the knowledge embedded in a LOSO model to initialize the WS-WS model on the held-out subject). Black distributions show the accuracies obtained by randomly initializing the CNN (i.e., WS-WS models trained from scratch, without exploiting knowledge from other subjects). The performance metric is reported for different numbers of calibration blocks used for training the TL-WS and WS models. Bigger dots represent the mean value across subjects, while the error bars represent the standard error of the mean. Smaller dots represent the accuracy scored for each subject (i.e., 15 data points). See part B of Table 3 for details about the number of examples defining the training, validation, and test sets. Wilcoxon signed-rank tests were performed to compare the performance of TL-WS models and WS-WS models. P-values were corrected for multiple comparisons via the Benjamini–Hochberg procedure and significant comparisons are marked once applied the correction (*$p<0.05$, **$p<0.01$, ***$p<0.001$). See Section 2.8.1 for further details about the statistical tests.

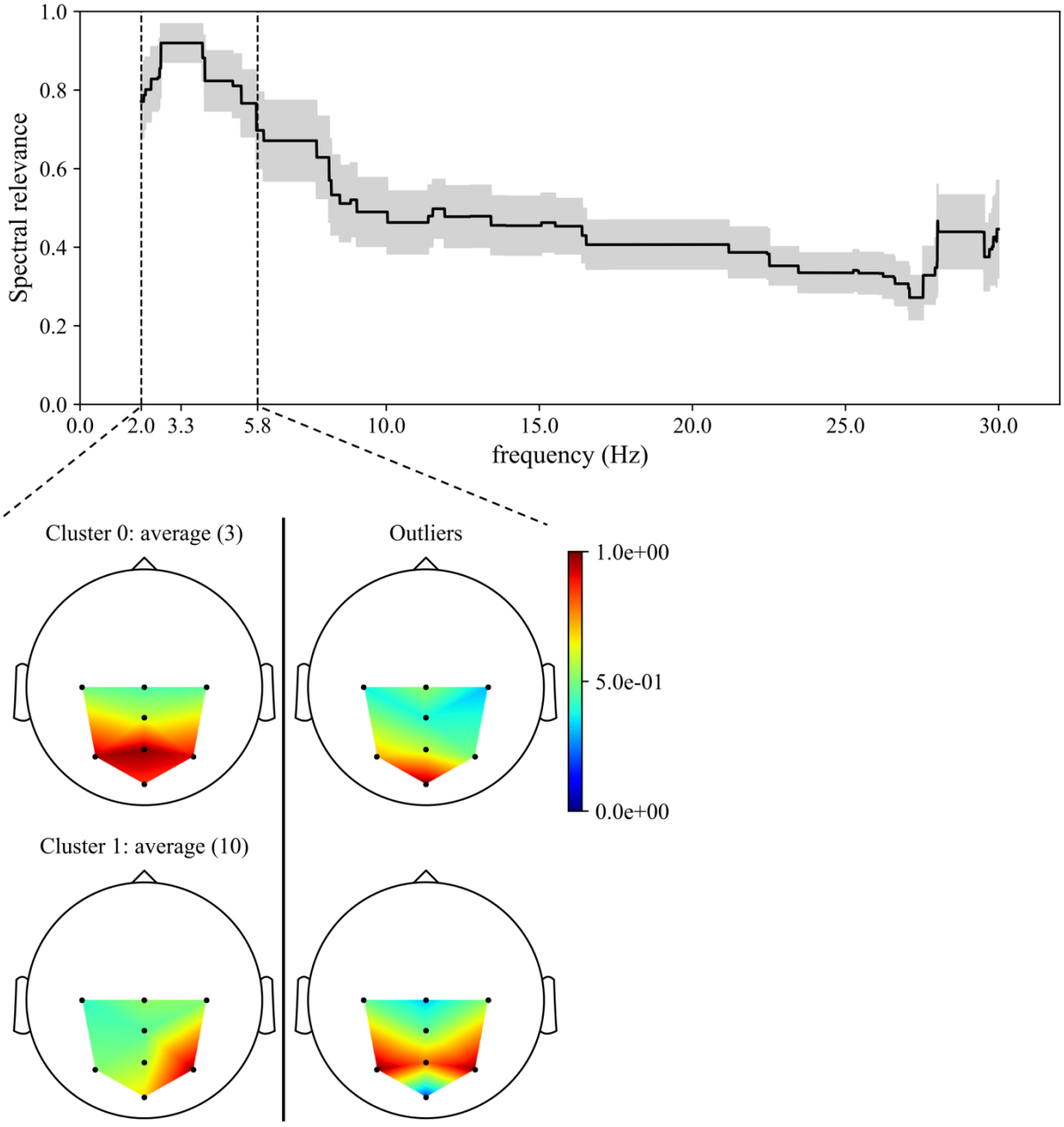

Figure 7 – Spectral relevance (top) and spatial relevance (bottom). The spectral relevance is reported in its mean value (black line) and standard error of the mean (grey shaded area) across subjects. The more relevant frequency range (associated to a relevance ≥0.75) is delimited by the two dashed vertical lines. The spatial relevance is displayed for the two clusters (averaged within each cluster) and to the outliers detected. The number of subjects in each cluster is reported within brackets.

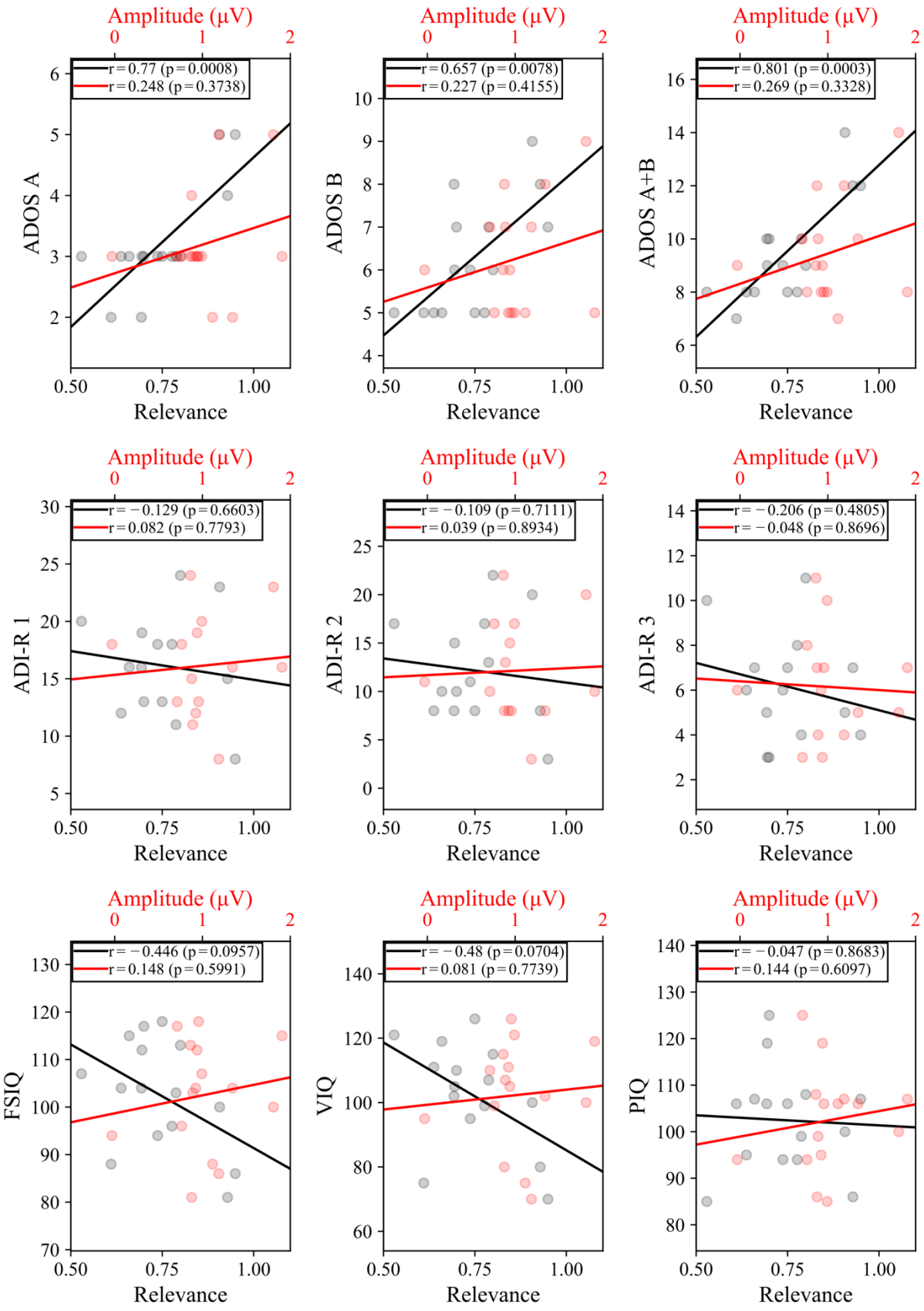

Figure 8 – Correlation analysis between each ASD clinical score and the ICNN+ET-derived (black) or ERP-derived measures (red). Subject-specific observations are reported with dots together with regression lines. Pearson's correlation coefficients together with p-values are reported inside each plot.

TABLES

Table 1 – Measured ASD clinical scores, i.e., ADOS, ADI-R, and IQs (mean ± standard deviation across subjects).

Measure	Value
ADOS A: Communication	3.20±0.86
ADOS B: Social Interaction	6.27±1.29
ADOS A+B: Communication-Social Interaction	9.47±1.86
ADI-R 1: Social Interaction	16.14±4.39
ADI-R 2: Communication and Language	12.14±5.19
ADI-R 3: Restricted and Repetitive Behavior	6.14±2.33
FSIQ: Full Scale IQ	102.53±11.24
VIQ: Verbal IQ	102.33±16.06
PIQ: Performance IQ	102.46±10.59

Table 2 – Sinc-ShallowNet-v2. Each layer is provided with its name, main hyper-parameters and number of trainable parameters. See Sections 2.2 and 2.3 for the meaning of the symbols. In all layers, where not specified, stride (S) and padding (P) were set to $(1,1)$ and $(0,0)$, respectively. The structural hyper-parameters (i.e., hyper-parameters of the architecture structure) reported in bold were searched via Bayesian optimization.

Block	Layer name	Hyper-parameters	Number of trainable parameters
	Input	$K_0 = 1$	0
ISS	Sinc-Conv2D	$\boldsymbol{K_0^{ISS}}$, $F_0^{\mathrm{ISS}} = (1,65)$, $P_0^{\mathrm{ISS}} = (0, F_0^{ISS}[1]//2)$	$2 \cdot K_0^{ISS} \cdot K_0$
	BatchNorm2D		$2 \cdot K_0^{ISS}$
	Depthwise-Conv2D	$\boldsymbol{D_1^{ISS}}$, $K_1^{ISS} = K_0^{ISS} \cdot D_1^{ISS}$, $F_1^{ISS} = (C, 1)$, $\boldsymbol{c_1^{ISS}}$	$F_1^{ISS}[0] \cdot F_1^{ISS}[1] \cdot K_1^{ISS}$
	BatchNorm2D		$2 \cdot K_1^{ISS}$
	ELU		0
	AvgPool2D	$F_p^{ISS} = S_p^{ISS} = (1,4)$	0
	Dropout	$\boldsymbol{p^{ISS}} = \boldsymbol{p^{FST}} = \boldsymbol{p}$	0
FST	Separable-Conv2D	$\boldsymbol{K_0^{FST}}$, $\boldsymbol{F_0^{FST}}$, $D_0^{FST} = 1, P_0^{FST} = (0, F_0^{FST}[1]//2)$	$F_0^{FST}[0] \cdot F_0^{FST}[1] \cdot K_1^{ISS} + K_1^{ISS} \cdot K_0^{FST}$
	BatchNorm2D		$2 \cdot K_0^{FST}$
	ELU		0
	AvgPool2D	$F_p^{FST} = S_p^{FST} = (1,8)$	0
	Dropout	$\boldsymbol{p^{FST}} = \boldsymbol{p^{ISS}} = \boldsymbol{p}$	0
FC	Flatten		0
	Fully-Connected	$N^{FC} = 2$, $\boldsymbol{c_0^{FC}}$	$N^{FC} \cdot (T//32 \cdot K_0^{FST} + 1)$
	Softmax		0

Table 3 –Part A: The number of trials in the training set and validation set used to tune the models (under Bayesian optimization) and number of trials in the test set used to test the models' performance, for the different training strategies. Part B: The number of trials in the training set, validation set and test sed used to tune and to test the WS-WS models and the TL-WS models, to evaluate the beneficial effect of transfer learning. Note that in the computational experiments of part B, the validation set was used only for early stopping, as the other hyper-parameters were inherited from the LOSO models tuned in experiments of part A. The trials in the test set were 2838 (on average) in each training strategy, as each model was tested separately on each session-specific test set, and then the performance was averaged across all sessions for each specific subject.

	Training strategy	No. of trials in the training set	No. of trials in the validation set	No. of trials in the test set
A	*Within-subject and within-session (WS-WS)*	1280	320	2838
	Within-subject and cross-session (WS-CS)	8960	2240	2838
	Leave-one-subject-out (LOSO)	125440	31360	2838
B	*Within-subject and within-session (WS-WS)*	From 128 to 1280 (step of 128)	From 32 to 320 (step of 32)	2838
	Transfer learning on single session (TL-WS)	From 128 to 1280 (step of 128)	From 32 to 320 (step of 32)	2838

Table 4 – Searched hyper-parameters of Sinc-ShallowNet-v2: distributions and admitted values sampled during Bayesian optimization. Curly brackets denote discrete admitted values, while square brackets denote interval of admitted values.

Hyper-parameter	Distribution	Values
${K_0}^{ISS}$	uniform	{4, 8, 16}
${D_1}^{ISS}$	uniform	{1, 2, 4}
${c_1}^{ISS}$	uniform	{None, 0.25, 0.5, 0.75, 1, 1.25, 1.5}
$p^{ISS} = p^{FST} = p$	uniform	{None, 0.25, 0.5}
${K_0}^{FST}$	uniform	$\{{K_0}^{ISS} \cdot {D_1}^{ISS}, 1, 2\}$
${F_0}^{FST}$	uniform	{(1,5), (1,9), (1,13), (1,17), (1,21)}
${c_0}^{FC}$	uniform	{None, 0.25, 0.5, 0.75, 1, 1.25, 1.5}
lr	log-uniform	[1e-4, 1e-1]

Table 5 – Model size (expressed as the number of trainable parameters) and training time (expressed in s/epoch) of Sinc-ShallowNet-v2 for each training condition. The total number of trainable parameters (mean ± standard error of the mean) is reported together with the number of parameters specific for each block, indicating within brackets the percentage of parameters exploited in each block.

	Block	**Within-subject and within-session (WS-WS)**	**Within-subject and cross-session (WS-CS)**	**Leave-one-subject-out (LOSO)**
Model size (# tr. parameters)	-	1207±141	1655±412	4638±559
	ISS	336±21 (28%)	466±47 (28%)	555±43 (12%)
	FST	749±121 (62%)	1042±368 (63%)	3689±483 (80%)
	FC	122±13 (10%)	147±40 (9%)	394±34 (8%)
Training time (s/epoch)	-	0.980±0.014	6.380±0.319	23.1±1.1