scientific reports

Check for updates

OPEN

A deep learning-enriched framework for analyzing brain functional connectivity

Davide Borra[1]✉ & Elisa Magosso[1,2]

Cognitive and motor functions require a coordinated communication among brain regions, with the directionality of interactions playing a key role, as the brain relies on functional asymmetries of reciprocal connections. Predictive models based on deep learning approaches could represent valuable tools for processing functional connectivity. However, these approaches are mainly adopted for decoding different brain states, but not for characterizing the information flow of functional networks. Here, we design a deep learning-enriched framework for analyzing spectral directed functional connectivity. The knowledge learned by a novel interpretable convolutional neural network ('Functional-Connectivity-Net', FCNet) – trained to discriminate brain states from functional connectivity – is used to define novel inflow and outflow measures, characterized for being non-linear, and for combining the information across brain regions and frequencies in an optimally discriminative way. Moreover, network decision is explained via DeepLIFT, revealing the most relevant frequency contents and connectivity inflow/outflow. We apply our approach to EEG functional connectivity estimated at both the scalp and cortex level, during motor imagery tasks. The network explanations match the known markers of spectral connectivity changes underlying motor imagery, and the network-based measures capture connectivity changes with high strength and significance like graph theory measures (in degree, out degree, authority, hubness). Our framework is helpful to elucidate the predictability of brain functional networks, and the most informative frequencies and connectivity inflow/outflows for the analyzed brain states.

Keywords EEG, Functional connectivity, Spectral granger causality, Connectivity inflow and outflow, Interpretable neural networks, Explainable artificial intelligence

Cognitive functions require a coordinated flow of information among functionally specialized brain regions. The dynamics of brain coordination is thought to be mediated by neural oscillations[1–3]. These reflect the synchronized rhythmic fluctuations of local neuronal ensembles, and are considered key mechanisms for facilitating the inter-regional flow of neural information[4]. Inter-regional communication exhibits distinct patterns in resting state and during task execution. To achieve a more complete overview on cognitive functions it is necessary to characterize these patterns, not only to describe normal mechanisms of perception, attention, and learning, but also their pathological alterations such as in neurodegenerative diseases[5]. The hypothesized key role of neural oscillations in establishing and maintaining inter-regional communication for supporting normal brain functions has led to a growth in the design and application of computational methods for estimating and analyzing oscillatory interactions in electrophysiological signals, e.g., electroencephalogram (EEG)[6–10].

As concerning EEG studies, different approaches were proposed in the literature for quantifying the oscillatory interactions between brain regions (see Bastos et al.[8] and Cao et al.[9] for a review), estimating the functional connectivity between them, that is, the statistical interdependence among spatially distant neurophysiological regions[11], as computed from the measured neural time series. These estimates can be computed at the level of EEG sensors (scalp-level description) or at the level of cortical regions (cortex-level description) in case the cortical source activity is reconstructed back from the EEG[12]. Moreover, the estimates can be provided as a function of the frequency, resulting in spectral connectivity estimations. The estimates can be grouped into non-directed and directed, depending on whether the computation considers not only the strength (and the frequency) but also the direction of the interaction. There is a strong interest into directed estimates[13], such as the one based on Granger causality[14], as the directionality of interactions plays a key role in in brain functioning.

[1]Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, Cesena Campus, Cesena 47522, Italy. [2]Alma Mater Research Institute for Human-Centered Artificial Intelligence, University of Bologna, Bologna 40126, Italy. ✉email: davide.borra2@unibo.it

Indeed, the brain organization relies on functional asymmetries of reciprocal connections, e.g., differences between ascending and descending connections in cortical hierarchies[15] or asymmetries in interhemispheric interactions[16,17]. For example, during the imagination of right-hand and left-hand movements, an increased connectivity has been found from left to right and from right to left regions, respectively[18–20].

Analyzing brain interactions, however, is accompanied by the challenge of analyzing hundreds of interactions in the brain network, scaling up exponentially as the number of brain regions analyzed in the network increases (e.g., 462 possible interactions, excluding auto-connections, when using 24 regions of interest), often leading researchers to focus only on few brain regions or on a selection of interactions, based on a priori knowledge. To address this limitation and to analyze the brain network in its entire structure, measures derived from graph theory are widely exploited[21,22]. These measures summarize complex brain networks (consisting of hundreds of interactions) using a few parameters with a clear geometrical meaning, and represent powerful tools for capturing the topological changes in the whole brain networks. Graph theory measures have been successfully applied to directed functional connectivity estimates, advancing the understanding of cognitive processing[23], by studying their modulations between the different brain states. Among the several graph theory measures, centrality indices, such as the in degree, out degree, authority and hubness, measure the strength of the transmission (i.e., outflow) and reception (i.e., inflow) of information in each brain region of an analyzed network, and are particularly useful to provide a quantitative overview of the overall directionality of information flow (e.g., from left to right hemisphere, from anterior to posterior regions, etc.) in a network of interconnected brain regions (see for example Ursino et al.[24] and Borra et al.[7]). Centrality measures based on graph theory have provided neuroscientists valid tools for conveniently analyzing the regional inflow and outflow of information in a directed connectivity network. Simpler inflow/outflow measures of a brain region, within a network of interacting regions, can be computed by equally weighting all entering/exiting connectivity values, i.e., by equally weighting all other regions in the network. This is the case of the in degree and out degree. More refined measures, like authority and hubness, weight the other brain regions depending on the interactions established in the brain network, i.e., specifically for a brain condition. This weighted computation proved to realize more sensitive indices than in degree and out degree[24]. When computing these measures, the frequency dimension is generally analyzed by aggregating connectivity estimates across frequencies belonging to a few frequency bands[13], defined based on a priori knowledge on the addressed cognitive task, for example alpha-band and beta-band for motor imagery[25,26]. Moreover, the measures exploit linear formulations on the connectivity matrix to aggregate the different brain interactions, e.g., the weighted sum of the connectivity values entering in a brain region when computing the in degree. Centrality measures quantifying inflow/outflow could be further engineered to optimally weight both brain interactions and frequency components such that the brain conditions in the cognitive task are maximally separated in the computed measure, and to exploit non-linear formulations in their definition.

Recent advancements in deep learning – originally proposed for decoding and analyzing neural time series – could be transposed and exploited for decoding and analyzing brain functional connectivity estimates. Deep learning techniques have rapidly emerged in the past years for decoding neural time series including EEG signals, primarily relying on convolutional neural networks (CNNs)[27,28]. A CNN processes the input neural activity recorded during a cognitive task (e.g., a motor imagery task) by automatically learning the neural features for mapping the input into one of the output brain states that characterize the task (e.g., a specific motor imagery condition). By doing so, the CNN automatically learns from the data the most salient aspects of the neural activity that are specific for predicting a brain state, distinguishing it from others. Therefore, the knowledge embedded in the neural network is task-specific by construction. Moreover, as the processing operated by the CNN is generally non-linear, it can describe complex and non-linear relationships from the input neural activity. Finally, convolutional operations can be performed in the CNN separately into a given domain (e.g., frequency or spatial domain) for efficiently summarizing the neural activity in that domain with no/minimal pre-processing required (e.g., with no a priori spectral aggregation), thus, avoiding discarding any relevant information. Due to these unique properties, initially exploited only for decoding EEG signals, deep learning-enriched frameworks have been proposed and validated also as tools for analyzing EEG signals[29–35]. To increase the interpretability of the knowledge learned by the CNN – otherwise difficult to be interpreted and exploited in the analysis framework – neural network designs have been proposed including interpretable layers[29–31], that is, layers whose features are easily interpretable in a given domain (e.g., frequency domain) once the CNN is trained. In addition to interpretable CNNs, explanation techniques, e.g., Deep Learning Important FeaTures (DeepLIFT)[35–38], were coupled to the CNN to derive useful visualizations of the most relevant samples in a given domain (space, time, frequency) that mostly drive the neural network decision towards a specific brain state (thus, designing explainable artificial intelligence frameworks)[32–38].

While deep learning approaches have been applied to EEG signals for both decoding and analysis, their application to EEG functional connectivity is still limited and restricted to decoding. In particular, predictive models that process EEG-derived functional connectivity mainly relies on traditional machine learning approaches[39–42] – i.e., realized by manually extracting features from functional connectivity matrices (e.g., centrality measures) and then classifying them. Only a few studies exploited CNNs and their automatic feature learning[43,44]; however, they mainly focused on neural decoding without aiming at realizing novel analysis tools. Moreover, they relied on complex deep and non-interpretable CNN architectures, hindering the possible use of these CNNs for revealing neurophysiological features linked to functional connectivity. Finally, these previous predictive models mainly processed functional connectivity matrices computed in pre-defined EEG bands; thus, potentially useful information contained at the level of single frequency bins may have been masked. Only with fMRI-derived connectivity representations, Ellis et al.[45] applied an explainable artificial intelligence approach based on a 1D-CNN as an analysis tool to uncover the effects of schizophrenia in brain functional networks. However, the used CNN was non-interpretable, and thus the learned hidden features were not easily interpretable and relatable to brain interactions. Following these considerations, the literature still lacks a deep

learning-empowered framework, realized both interpretable and explainable, for analyzing EEG-based spectral directed functional connectivity.

Predictive models based on deep learning approaches could represent valuable tools for processing functional connectivity for two main reasons[46]. First, they could represent important benchmark tools for quantifying the level of predictability (i.e., discriminatory power) of spectral directed functional networks in an end-to-end fashion, that is, linking them to the output cognitive state with a minimal processing of the input connectivity matrices. This would indicate the level of predictability of functional brain networks for specific cognitive states of interest. Second, the adoption of interpretable and explainable components in the design of the framework would offer a unique tool devoted at automatically unveiling the most informative frequency components and connectivity inflow/outflow features for the considered cognitive states, directly from the spectral connectivity matrix. Importantly, inspecting these patterns would provide useful insights to extend traditional graph theory measures, expanding the quantification of directed functional networks.

The present study is positioned within this line of research, by proposing a novel deep learning-enriched framework devoted at processing directed spectral functional connectivity derived from EEG signals. This would help identifying the performance ceiling of functional networks, and characterizing novel measures of information flow that could support traditional measures. To the best knowledge of the authors, this represents the first attempt of applying a deep learning approach with a primary focus on the analysis of brain interactions. The key novel properties of our approach are:

i. The design of a novel interpretable CNN that processes directed spectral connectivity, termed FCNet ('Functional-Connectivity-Net'). The analysis framework is based on this network. FCNet processes connectivity estimates separately in the frequency and spatial domains. It employs a spectral feature extractor – that optimally summarizes the information along the frequency axis – and an interpretable spatial feature extractor – designed for automatically learning the inflow and outflow measures that better separate the output brain states. Remarkably, these measures (a) aggregate inter-regional interactions by weighting brain regions such that the brain states are maximally separated, (b) reflect the contribution of the most salient frequency components, as automatically learned by the neural network, and (c) are non-linear by construction.
ii. The automatic characterization of the most salient frequencies and connectivity inflow/outflows from spectral directed connectivity estimates in an end-to-end manner. It exploits the inflow and outflow measures extracted from FCNet for analyzing the brain functional connectivity, and it couples the neural network with an explanation technique (DeepLIFT) for providing useful visualizations of the frequencies and of inflow/outflow measures most relevant for driving the network decision towards the output brain states.

We show the potentialities of our deep learning-enriched framework by applying it to a motor imagery task involving the imagination of right-hand and left-hand movements. To provide a wider validation of the framework, experiments are conducted on the directed functional connectivity estimated via Granger causality at both scalp level and cortex level using two different EEG datasets (63 total participants).

Finally, it is crucial to highlight that our aim is to present the design and exploitation of an interpretable and explainable artificial intelligence approach for analyzing functional connectivity during motor imagery (i.e., an approach for contributing to neural data analysis), and not to propose a new neural network for improving motor imagery neural decoding. Indeed, we hypothesize that a deep learning-enriched approach, designed to be both interpretable and explainable, is capable of capturing changes of brain functional connectivity occurring during a cognitive task. Such approach could represent a useful tool for supporting the analysis of brain interactions, by revealing the predictability of functional networks, and the informativeness of frequency components and connectivity inflows/outflows related to cognitive states directly from the causal spectra.

Methods

Decoding and analysis of functional connectivity: main concepts and notations

In this study, we developed a framework for decoding and analyzing the brain functional connectivity estimated from EEG recordings in a data-driven way, by exploiting the knowledge learned by a learning system (a CNN in this study). In the following, we present the main concepts and useful notations.

Let us denote the participant-specific dataset composed of pre-processed EEG signals by:

$$D_{EEG}{}^{(p)} = \{(X_0,\, y_0)\,, \ldots\,, (X_i,\, y_i)\,, \ldots\,, (X_{M-1},\, y_{M-1})\}, \tag{1}$$

where M indicates the number of EEG examples (e.g., EEG epochs in a trial-based EEG recording), assumed here to be the same for each participant. $X_i \in \mathbb{R}^{R\times T}$ represents the EEG multi-variate activity recorded in the i-th example from R brain regions at T time samples, and y_i is the associated label among a set L of brain states of interest (e.g., right-hand motor imagery in a motor imagery task). The EEG multi-variate time series can be represented in the spatial domain at the scalp-level – corresponding to signals at EEG sensors – or at the cortex-level – corresponding to signals obtained by back-projecting the EEG activity on the cortical surface. The latter is enabled via cortical source reconstruction[12], a processing step devoted to transform sensor-space signals (scalp signals) into source-space signals (cortical signals), generally aggregated within cortical regions of interest (ROIs). Therefore, depending on the level of the EEG spatial representation, the brain regions characterized in the multi-variate neural activity either are located on the scalp or on the cortical surface.

Starting from pre-processed EEG data, the brain functional connectivity can be estimated, for example by computing directional causal influences between brain regions (at the scalp-level or cortex-level) via Granger causality[13]. Granger causality, as other measures of connectivity, can be formulated in the spectral domain (spectral functional connectivity), quantifying the interactions between regions separately for F different

frequency bins, i.e., the connectivity information is provided as a function of frequency. Therefore, for each example $X_i \in \mathbb{R}^{R\times T}$, containing the multi-variate neural activity, the spectral functional connectivity estimate $A_i \in \mathbb{R}^{R\times R\times F}$ is derived. This matrix quantifies the interaction from the j-th to the k-th brain regions (at the scalp-level or cortex-level) at the f-th frequency bin, having denoted with j, k, f the indices of the matrix A_i (i.e., $A_i[j,k,f]$). By linking each connectivity estimate to the associated brain state (as in Eq. 1), the participant-specific dataset composed of functional connectivity estimates can be represented by:

$$D_{FC}{}^{(p)} = \{(A_0, y_0), \ldots, (A_i, y_i), \ldots, (A_{M-1}, y_{M-1})\}, \tag{2}$$

A CNN-based learning system can be trained to realize a classifier g aimed at discriminating between the brain states of interest, starting from the functional connectivity estimates as input. Thus, the CNN describes the non-linear function g, mapping the brain interactions – contained in A_i–to a brain state:

$$g(A_i;\theta) : \mathbb{R}^{R\times R\times F} \rightarrow L, \tag{3}$$

where g is parametrized in the parameter array θ (whose values are learned during training). By doing so, the system automatically learns the most relevant features about the brain interactions for correctly discriminating the output brain states. Crucially, the knowledge learned by the learning system is encoded in the parameter values θ, and can be leveraged for designing a novel analysis framework for studying the brain functional connectivity in a data-driven way. At a high level, this can be achieved by performing the following analyses.

i. Analysis of the output of intermediate interpretable layers. In case the network architecture includes intermediate interpretable layers, their output can be easily interpreted and analyzed, when processing the input A_i. As an example, in case the network is forced at extracting the connectivity inflow or outflow that maximize the between-class discrimination in an intermediate layer, this layer output directly provides connectivity measures that can be analyzed.
ii. Explanation of the network decision. The network can be combined with an explanation technique to identify, while processing the input A_i, the most relevant samples in the feature maps of a target layer that, based on the learned knowledge θ, drive the decision of the network towards a specific output brain state. In this case, it is crucial that the chosen target layer works on representations living in an interpretable domain. The target layer could be the input layer, which works on the spatial and frequency domains in case of spectral functional connectivity estimates, or a hidden interpretable layer, which forces interpretability in a given domain. As an example, with this procedure it is possible to analyze the frequency components that are maximally relevant for discriminating a brain state under analysis (e.g., the right-hand motor imagery in case of a motor imagery task).

The proposed deep learning-enriched analysis framework encapsulates all these aspects in a unique and comprehensive tool, that could be conveniently used by neuroscientists for analyzing functional connectivity estimates in a data-driven way. In the following, we present the core methodology underlying our framework, and we illustrate the experiments conducted on real data using the proposed framework.

Deep learning-enriched analysis of functional connectivity

Here we describe the proposed deep learning-enriched framework for analyzing brain interactions. At first, we describe the developed neural network for decoding brain functional connectivity. Then, we delineate how to exploit the network knowledge for extracting and analyzing connectivity-related measures, by explaining in detail the data-driven analyses sketched in Sect. "Decoding and analysis of functional connectivity: main concepts and notations" (see points i. and ii).

Network architecture and training

We designed a compact convolutional neural network – termed FCNet ('Functional-Connectivity-Net') – for decoding functional connectivity estimates derived from EEG signals. FCNet processes the functional connectivity matrix $A_i \in \mathbb{R}^{R\times R\times F}$ by first extracting features in the frequency domain (*spectral block*), and then in the spatial domain (*spatial block*). The processing operated in the spatial block is interpretable by design, and it is characterized by the extraction of measures of connectivity inflow and outflow. Finally, the last *classification block* finalizes the decoding problem. Specifically, in our experiments FCNet was applied to a binary classification task consisting of the discrimination between right-hand vs. left-hand motor imagery (see Sect. "Experiments"). In Fig. 1 we provide a high-level scheme of the processing operated by the fundamental blocks of FCNet. The main FCNet blocks are described in the following, and a complete overview of the network structure can be found in Table 1.

i. Spectral block. The spectral block aims at searching features that summarize the causal spectra in the frequency domain. First, a 3D convolutional layer is applied, learning $K = 32$ kernels with size of W=(1,1,9), using a unitary stride. Feature maps are pooled (average pooling) in the frequency domain, to reduce the computational cost, utilizing a pooling size and stride of $W = S = (1, 1, 4)$, and neurons of the pooled feature maps are dropped out during training (dropout probability $p = 0.1$). Then, a second 3D convolutional layer is applied, continuing the learning of spectral features. This layer learned $K = 32$ kernels too, with size of $W = (1, 1, 9)$. To keep limited the number of trainable parameters, separable convolution (i.e., depthwise convolution followed by pointwise convolution) is implemented, utilizing a depth multiplier $D = 1$. Feature maps are average pooled in the frequency domain, utilizing a pooling size and stride

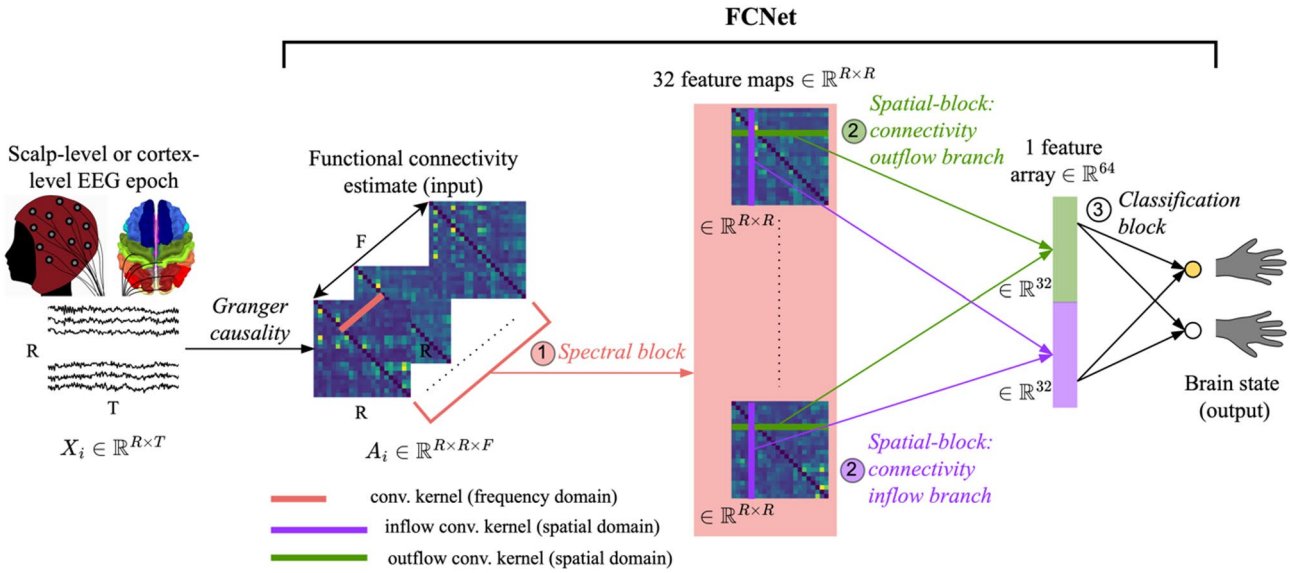

Fig. 1. High-level graphical representation of the processing operated by FCNet blocks. FCNet comprises three main blocks: a *spectral block*, a *spatial block* (composed by two connectivity outflow and inflow parallel branches), and a *classification block*. Here we schematize the processing by reporting the output of each FCNet block while processing the input spectral functional connectivity $A_i \in \mathbb{R}^{R\times R\times F}$, computed from the EEG activity $X_i \in \mathbb{R}^{R\times T}$ at the scalp-level or cortex-level. The output is represented by the right-hand and left-hand motor imagery conditions.

of $W = S = (1, 1, 4)$, and neurons of the pooled feature maps are dropped out during training (dropout probability $p = 0.1$). This block, with the chosen hyperparameters, provides as output 32 feature maps with shape (R, R), each summarizing the spectral functional connectivity presented as input to the network, across all frequency bins.

ii. Spatial block. The spatial block aims at learning connectivity inflow- and outflow-related features from the high-level connectivity feature maps provided by the spectral block. To this aim, this block exploits a parallel dual-branched structure (*connectivity inflow branch* and *connectivity outflow branch*). Here, each branch operates in parallel on the input connectivity feature maps by separately learning inflow and outflow features. Specifically, the connectivity inflow branch includes a depthwise 2D convolutional layer (depthwise multiplier of $D = 1$) – termed *inflow convolutional layer* – learning 32 kernels with size $W = (R, 1)$. This way, the network learned 32 features (one per feature map of the spectral block, as $D = 1$) aimed at computing the connectivity inflow associated to each brain region from the connectivity feature maps provided by the spectral block. Therefore, the output of the inflow convolutional layer provides a quantification (in 32 variants) of the brain connectivity inflow at each brain region; remarkably, the mathematical formulation of these inflow measures – parametrized in the parameters of the inflow convolutional kernels – is automatically learned by the network from the input data. Then, the network learns how to optimally recombine together the 32 brain connectivity inflows by performing a pointwise 2D convolutional layer, learning 32 kernels with size $W = (1, 1)$. The combination of these depthwise and pointwise convolutions realizes a separable convolutional layer. Finally, by exploiting an average pooling layer (pooling size and stride of $W = S = (1, R)$), the recombined inflow measures are pooled together across brain regions, obtaining a single scalar value for each feature map. As concerning the connectivity outflow branch, the same processing as in the inflow branch is performed (same sequence of layers) but operating in space along the other dimension, that is, using convolutional kernels with size $W = (1, R)$ in the depthwise 2D convolutional layer, and pooling feature maps using a pooling size and stride of $W = S = (R, 1)$. Similar to the connectivity inflow branch, in the connectivity outflow branch the output of the depthwise 2D convolutional layer – termed *outflow convolutional layer* – provides a quantification (in 32 variants) of the brain connectivity outflow at each brain region. Overall, each parallel branch (connectivity inflow branch and connectivity outflow branch) outputs a feature array with 32 elements, summarizing the inflow and outflow measures learned by the network across all the brain regions.

iii. Classification block. This block finalizes the decoding problem starting from the two feature arrays provided by the dual-branched spatial block. The feature arrays are concatenated together into a single feature array with 64 elements, and then provided as input to a fully-connected layer with N units, one per brain state to classify. As the experiments were conducted on a right-hand vs. left-hand motor imagery classification task (2 motor brain states, see Sect. "Experiments"), the fully-connected layer includes $N = 2$ units, activated via softmax function. Therefore, the network output produces the conditional probability $p\left(l_c|A_i\right)$, $\forall l_c \in L = \{\text{right-hand}, \text{left-hand}\}$, where c is the index of a specific output class (i.e., output brain state).

Block	Layer name	Main hyper-parameters	Output shape
	Input	–	$(1, R, R, F = 81)$
Spectral	*Conv3D*	$K = 32, W = (1, 1, 9)$	$(32, R, R, 73)$
	BatchNorm3D	–	$(32, R, R, 73)$
	ReLU	–	$(32, R, R, 73)$
	AvgPool3D	$W = S = (1, 1, 4)$	$(32, R, R, 18)$
	Dropout	$p = 0.1$	$(32, R, R, 18)$
	SepConv3D	$K = 32, D = 1, W = (1, 1, 9)$	$(32, R, R, 10)$
	BatchNorm3D	–	$(32, R, R, 10)$
	ReLU	–	$(32, R, R, 10)$
	AvgPool3D	$W = S = (1, 1, 4)$	$(32, R, R, 1)$
	Dropout	$p = 0.1$	$(32, R, R, 1)$
	Squeeze	–	$(32, R, R)$
Spatial – connectivity inflow branch	*InflowConv2D*	$K = 32, D = 1, W = (R, 1)$	$(32, 1, R)$
	PointConv2D	$K = 32, W = (1,1)$	$(32, 1, R)$
	BatchNorm2D	–	$(32, 1, R)$
	ReLU	–	$(32, 1, R)$
	AvgPool2D	$W = S = (1, R)$	$(32, 1, 1)$
	Flatten	–	(32)
Spatial – connectivity outflow branch	*OutflowConv2D*	$K = 32, D = 1, W = (1, R)$	$(32, R, 1)$
	PointConv2D	$K = 32, W = (1,1)$	$(32, R, 1)$
	BatchNorm2D	–	$(32, R, 1)$
	ReLU	–	$(32, R, 1)$
	AvgPool2D	$W = S = (R, 1)$	$(32, 1, 1)$
	Flatten	–	(32)
Classification	*Concatenate*	–	(64)
	Fully-connected	$N = 2$	(2)
	Softmax	–	(2)

Table 1. FCNet architecture. Each layer is provided with its name, main hyper-parameters, number of parameters to fit and output shape. Where not specified, a unitary stride (S) and no padding (P) were applied. The output shape is reported as a function of the number of brain regions R, that varied in our experiments: $R = 22$ in scalp-level experiments and $R = 24$ in cortex-level experiments. The total number of parameters to fit was 5.4k and 5.6k, respectively for the conducted scalp-level and cortex-level experiments.

For spectral and spatial blocks, after each convolutional layer a batch normalization layer is included, and the resulting feature maps are then activated via rectified linear activations (ReLU functions), introducing non-linearity in the network.

The network trainable parameters – 5.4k and 5.6k, respectively in the scalp-level and cortex-level experiments (see Sect. "Experiments") – were trained for 500 epochs using the cross-entropy as loss function and Adam as optimizer (learning rate of 5·1e-4, and mini-batch size of 32). The network hyperparameters (e.g., the number and size of the convolutional kernels in the spectral block) were determined via preliminary empirical hyperparameter evaluations. A within-subject training strategy was employed, training the network separately for each participant. We adopted this choice as we were interested in designing an algorithm able to highlight participant-specific connectivity signatures, by leveraging on the participant-specific neural decoders' knowledge. We designed the training strategy in this way to open important application scenarios of our analysis framework, such as the prospective application to individual patients, to better support the analysis of the neuropathology and guide the personalization of treatments in patients. To assess the decoding performance, the dataset of the p-th participant ${D_{FC}}^{(p)}$ was partitioned employing a 10-fold cross-validation scheme. Moreover, from the training set we sampled 20% of examples form the validation set, for performing early stopping, arresting the learning at the training epoch with the highest validation accuracy. Therefore, the remaining 80% of examples were effectively used to train the network. The confusion matrix and accuracy served as performance metrics. For each participant and each cross-validation fold, these metrics were computed on the test set, and then averaged across folds.

Analysis of the output of intermediate interpretable layers

The design of FCNet includes two interpretable layers: the inflow convolutional layer and the outflow convolutional layer. By construction these layers operate by automatically learning, via convolutional operators, 32 connectivity inflow measures and 32 connectivity outflow measures, respectively. Overall, the performed

computation is non-linear with respect to the input spectral connectivity, encompassing the spectral block and the inflow/outflow convolutional layer. Moreover, the computation exploits the knowledge learned by the interpretable CNN during the training process, such that it optimally combines the spectral information (frequency domain: spectral block) and it optimally weights the brain interactions (spatial domain: inflow/outflow convolutional layer) in order to maximally separate the output brain states. To process these network-based measures, we performed the following steps.

For each participant and each cross-validation fold (that is, for each trained network), we performed a forward pass through the network using the test examples A_i as input (i.e., $\forall\, i | A_i \in \text{test set}$). We extracted the 32 inflow measures and 32 outflow measures – each $\in \mathbb{R}^R$ (i.e., quantifying the inflow/outflow at each brain region) – at the output of the inflow convolutional layer and outflow convolutional layer, respectively. That is, we exploited the neural network up to the interpretable layers as a tool for extracting representations useful to the analysis of brain connectivity. For each trained network, each of the 32 inflow measures was averaged across the test examples associated to the same brain state classified by the network (i.e., one of the two motor imagery conditions, see Sect. "EEG pre-processing and processing"). The same was done for the outflow measures. After this procedure, for each trained network (i.e., for each participant and each cross-validation fold), 32 average inflow measures and 32 average outflow measures were obtained for each brain state (i.e., for each motor imagery condition). It is worth noticing that these resulting measures came out from independently trained models (one model per participant and cross-validation fold), that is, from models with different trained convolutional kernels. Therefore, we performed a clustering procedure for appropriately ordering the measures and aggregating them across models. The clustering procedure was applied separately to inflow and outflow measures. Specifically, constrained K-means[47] was adopted, and applied to the inflow (/outflow) measures averaged across the brain states. The clustering algorithm searched for 32 clusters (one per measure), while imposing no links between the measures within each participant and each cross-validation fold. That is, the clustering of the 32 inflow (/outflow) measures was achieved only across – and not within – the participant and fold dimensions. This ensured that each cluster was characterized by a peculiar connectivity inflow (/outflow) measure, that was consistent across participants and cross-validation folds. Once found the 32 clusters of inflow/outflow computations, for each brain state, we ordered the measures of each participant and each cross-validation fold according to the cluster they belong to, and we averaged each measure across the 10 cross-validation folds. Overall, this processing resulted into 32 average inflow measures and 32 average outflow measures for each participant and each brain state, termed as *FCNet-based inflow measures* and *FCNet-based outflow measures*. These 64 neural network-based connectivity measures were under investigation in this study.

Explanation of the network decision

FCNet decision was explained to understand the samples in the frequency and spatial domains that mostly drove the network decision towards the appropriate brain state. The network explanation analysis has been included in the proposed framework to not only extract useful measures for analyzing brain connectivity but also to visualize the most relevant contributions for the learning system in interpretable domains (frequency and spatial domains).

For each trained neural network (i.e., for each participant and each cross-validation fold), we explained the network decision by using the Deep Learning Important FeaTures (DeepLIFT) algorithm[48] while the network processed the test set examples A_i as input (i.e., $\forall i | A_i \in \text{test set}$). Once performed the forward propagation of the information using an input example A_i, DeepLIFT backpropagates the output prediction back to a target layer (e.g., the input layer), providing a relevance representation map with the same shape of the layer output, quantifying the positive or negative contribution to the output prediction. This explanation technique provides a measure of the change in the output from a reference value with respect to the change in the input from a reference input value. In this study we exploited a reference input connectivity matrix of zeros, which is also the default option for DeepLIFT. DeepLIFT is widely adopted in the literature for explaining the network decision when applying deep neural networks to multi-variate neural activity[35–38], and represents a good candidate for explaining the decision of the proposed neural network, that processes a connectivity matrix extracted from a multi-variate neural activity. Moreover, a recent benchmark study on explanation techniques applied on simulated EEG time series showed that DeepLIFT is consistently more accurate and robust to detect neural changes in the temporal, spatial, and frequency domains compared to other techniques, e.g., saliency maps, gradient-weighted class activation mapping (Grad-CAM), and guided GradCAM[35].

We derived DeepLIFT relevance representations associated to the output neuron of the correct class (i.e., the brain state associated to the input: right-hand or left-hand motor imagery in this study), separately with respect to the input layer and to the inflow/outflow convolutional layers. Crucially, the output of all these layers is interpretable, thus, the derived relevance representations can be easily extracted and interpreted in a domain of interest. For each trained network (i.e., each participant and cross-validation fold), and each input test example $A_i \in \mathbb{R}^{R \times R \times F}$ we derived three types of representations, describing how the network processed the input in distinct domains for providing the correct brain state as output. The first representation quantified the relevance of each input connection, between each pair of brain regions, as a function of the frequency, resulting into an input relevance map $\in \mathbb{R}^{R \times R \times F}$ (relevance computed with respect to the input layer). The second and the third representations quantified the relevance of each connectivity inflow and outflow measure (64 in total) for each brain region, resulting into 32 inflow and 32 outflow relevance maps $\in \mathbb{R}^R$ (relevance computed with respect to the inflow/outflow convolutional layers). These three types of representations were separately averaged across the test examples and underwent the following processing, for easing our understanding of the most relevant connectivity-related features in the frequency and spatial domains.

i. Frequency domain. As concerning the relevance representations computed with respect to the input layer, we averaged the DeepLIFT maps ($\in \mathbb{R}^{R\times R\times F}$) across cross-validation folds, and all brain regions. Therefore, for each participant we obtained an average frequency pattern (*spectral relevance*, $\in \mathbb{R}^{F}$), quantifying the relevance of each frequency bin of the input connectivity matrix, for discriminating the correct brain state. This relevance pattern could be useful for visualizing the frequency range contained in the input connectivity estimate most important for decoding the different brain state conditions (here different motor imagery conditions).
ii. Spatial domain. As concerning the relevance representations computed with respect to the connectivity inflow and outflow layers, we averaged each of the DeepLIFT maps (i.e., each of the 32 inflow relevance maps and each of the 32 outflow relevance maps $\in \mathbb{R}^{R}$) across participants, cross-validation folds, and brain regions. The averaging across participants and folds was accomplished by ordering the 32 relevance maps within each participant and fold according to the results of the clustering procedure described in Sect. "Analysis of the output of intermediate interpretable layers". After this procedure, we obtained a scalar relevance score specific of each of the 32 inflow measures and of the 32 outflow measures (*inflow/outflow measure relevance score*), that, in its absolute value, could be useful for ranking the 64 inflow/outflow measures based on their contribution towards the correct output brain state. This score could be exploited for visualizing the most relevant connectivity inflow and outflow, as extracted from the neural network (out of the total 64 measures), when analyzing the output brain states.

Experiments

EEG pre-processing and processing

The proposed framework was applied on functional connectivity estimates derived at both the scalp level and cortex level, to provide a more complete evaluation. We addressed the decoding and analysis of right-hand vs. left-hand motor imagery (2 brain states under analysis), which is a common and simple task performed in the literature, mainly exploited for brain-computer interface applications[28,29]. In the following, the datasets used in our experiments are briefly presented, together with the performed pre-processing and processing steps.

i. Scalp-level experiments. We exploited the BNCI2014-001 dataset[51], also known as 'dataset IIa' from BCI competition IV. Widely recognized in the literature as a primary benchmark for assessing the efficacy of new neural networks in EEG applications, this dataset has been extensively employed in previous studies[27]. It consists of 22-channel EEG sampled at 250 Hz, recorded from 9 healthy participants across 2 recording sessions. Electrodes were placed according to 10–10 international system as displayed in Fig. 2a. The imagination of the right-hand and left-hand movements was performed for 4 s. For each participant and each session, 144 trials were recorded, balanced across classes (288 trials in total, for each participant). Signals were band-pass filtered between 1 and 40 Hz using a 4th order zero-phase Butterworth filter. Then, the EEG was epoched by extracting 4-s length epochs from 0 s to 4 s with respect to the motor imagery onset. Finally, to mitigate potential spurious connections between sensors arising from volume conduction when estimating brain connectivity (see Sect. "Functional connectivity estimation"), a spherical spline surface Laplacian transformation was applied[8]. After this pre-processing procedure, each EEG epoch was $X_i \in \mathbb{R}^{22\times 1000}$ ($R = 22$, $T = 1000$), and the corresponding label was $y_i \in L = \{\text{right-hand}, \text{left-hand}\}$. These epochs defined the ${D_{EEG}}^{(p)}$ for the p-th participant (see Eq. 1).
ii. Cortex-level experiments. We exploited the Lee2019-MI dataset[52]. This dataset consists of 62-channel EEG sampled at 1000 Hz, recorded from 54 participants across 2 recording sessions. The imagination of the right-hand and left-hand movements was performed for 4 s. For each participant and each session, 100 trials were recorded, balanced across classes (200 trials in total, for each participant). Signals were band-pass filtered between 1 and 40 Hz using a 4th order zero-phase Butterworth filter, and downsampled to 250 Hz (same sampling frequency used for the dataset described in point i.). Then, the EEG was epoched by extracting 4-s length epochs from 0 s to 4 s with respect to the motor imagery onset. Finally, the cortical activity was derived as follows from each EEG epoch. Scalp signals were transformed into cortical signals using MNE Python library (version 1.2.2)[53]. To this aim, a template head anatomy was adopted using the FSaverage template, with the source space restricted to the cortex and discretized into 20,484 vertices. The forward problem[54] was solved via the boundary element method with MNE default parameters. The inverse problem[12] was solved using eLORETA (exact Low Resolution Electromagnetic Tomography)[55] with MNE default parameters, with identity noise covariance matrix, and with the dipole source orientation constrained to be perpendicular to the cortex. By doing so, each cortical vertex was associated to one source signal. The cortical surface was parcellated into 24 ROIs (12 per hemisphere) extracted from the Desikan-Killiany atlas[56], covering frontal, parietal, and occipital areas; these ROIs are known to be involved in the fronto-parietal network active during motor execution and imagination[57,58], and were also considered in prior studies[7,59] when analyzing the cortical activity over a selection of relevant ROIs during movements. The list of the 24 ROIs and of their abbreviations is reported in Table 2, and their location over the cortex is displayed in Fig. 2b. It is worth noting that we adopted this coarser cortical resolution with large cortical ROIs in order to mitigate the effects of spatial blurring and localization inaccuracy due to the relatively low EEG spatial resolution (62 channels). A waveform representative of the neural activity of each ROI was derived by averaging all signals of the vertices belonging to that ROI. To avoid cancelling out the neural activity in case of many vertices within the ROI having dipole orientations in opposite directions, the signs of source signals that were not oriented as the 'dominant direction' were flipped before averaging[60]. The dominant direction corresponded to the first principal direction of orientations of dipoles belonging to the ROI. After these pre-processing and processing procedures, each EEG epoch was $X_i \in \mathbb{R}^{24\times 1000}$ ($R = 24$, $T = 1000$), and the corresponding

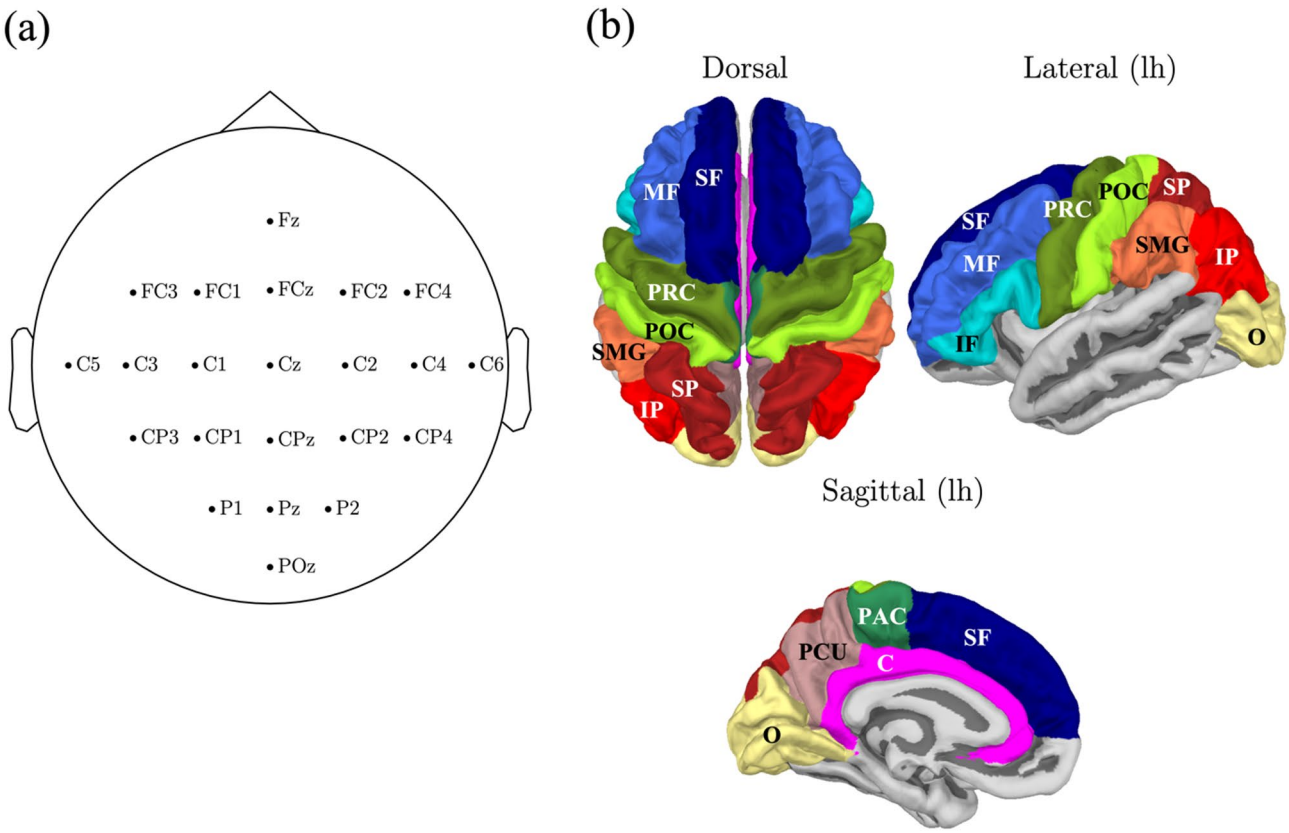

Fig. 2. Position of the EEG electrodes and cortical regions of interest (ROIs) of the data used in the experiments. (**a**) Location of the electrodes in the scalp-level experiments (data from the BNCI2014-001 dataset[51]). (**b**) Location of the cortical ROIs in the cortex-level experiments (data from the Lee2019-MI dataset[52]). Here, the dorsal view, and the lateral and sagittal views relative to the left hemisphere (lh) are displayed. See Table 2 for the complete name of ROIs (here abbreviated).

label was $y_i \in L = \{\text{right-hand}, \text{left-hand}\}$. These epochs defined the ${D_{EEG}}^{(p)}$ for the p-th participant (see Eq. 1).

Functional connectivity estimation

The directional influences between brain regions (either at the scalp-level or at the cortex-level) were estimated by computing pairwise spectral Granger causality (GC)[14], exploiting an order $p = 30$ (as done in prior studies[7,61]) in the bivariate autoregressive model. Since in both experiments the signals were band-limited in the range 1–40 Hz, spectral GC was computed in this range of frequency for a set of $F = 81$ frequencies, by using a frequency resolution of about 0.5 Hz. For the generic f-th frequency bin, the spectral GC is represented by a non-symmetric matrix $\in \mathbb{R}^{R \times R}$, with the off-diagonal jk-th value quantifying the influence exerted by the j-th signal (representing the j-th brain region) onto the k-th signal (representing the k-th brain region) at that frequency, i.e., $GC_{j \to k,f}$. For each participant, the directional influences in the frequency domain were estimated separately for each EEG epoch. Thus, we obtained a functional connectivity matrix $A_i \in \mathbb{R}^{R \times R \times F}$ for each EEG epoch X_i. To emphasize how much each connectivity value contributed to the overall connectivity across the brain regions and to mitigate inter-trial variability, the connectivity matrix at the f-th frequency bin ($f = 0, \ldots, 80$) was normalized such that the sum of all off-diagonal connectivity values was 1, i.e., $A_i[:,:,f] = A_i[:,:,f] \Big/ \sum_{j,k,j \neq k} A_i[j,k,f]$, as done in a previous study[7].

All functional connectivity estimates and the associated motor imagery conditions were collected to form the dataset ${D_{FC}}^{(p)}$ for the p-th participant (see Eq. 2). Each 3D matrix A_i represented the network input, and the corresponding motor imagery y_i represented the network output. Therefore, according to the adopted network training (see Sect. "Network architecture and training"), 207/144, 52/36, 29/20 examples were used in the training, validation, and test sets, respectively (scalp-level/cortex-level experiments).

Inflow and outflow measures: statistical analysis between motor states

To assess the effect size and statistical significance of the inflow and outflow measures, a statistical analysis on these measures was performed, by comparing each of them between the considered brain states (i.e., right-hand vs. left-hand motor imagery). These measures included the FCNet-based measures (see Sect. "Analysis of the

ROI name	Lobe	ROI abbreviation
Superior frontal gyrus	Frontal	SF
Middle frontal gyrus	Frontal	MF
Rostral division		
Caudal division		
Inferior frontal gyrus	Frontal	IF
Pars opercularis		
Pars triangularis		
Pars orbitalis		
Precentral gyrus	Frontal	PRC
Postcentral gyrus	Parietal	POC
Paracentral lobule	Frontal	PAC
Cingulate cortex	Frontal, Parietal	C
Rostral anterior division		
Caudal anterior division		
Posterior division		
Isthmus division		
Superior parietal cortex	Parietal	SP
Inferior parietal cortex	Parietal	IP
Supramarginal gyrus	Parietal	SMG
Precuneus cortex	Parietal	PCU
Occipital cortex	Occipital	O
Lingual gyrus		
Pericalcarine cortex		
Cuneus cortex		
Lateral occipital cortex		

Table 2. List of the cortical regions of interest (ROIs) adopted in the cortex-level experiments. The cortex was parcellated into 12 ROIs per hemisphere (24 ROIs in total), based on the Desikan Killiany atlas[56]. The complete name of each ROI is reported in the left column, the lobe to which the ROI belongs to is indicated in the middle column, while the ROI abbreviation is reported on the right. Additionally, the table reports the divisions/cortices composing each selected ROI, according to the chosen atlas.

output of intermediate interpretable layers") and also measures extracted with a classic approach (graph theory) used for reference. Specifically, the classic analysis approach consisted of the following processing. For each experiment (scalp-level and cortex-level), participant and motor imagery condition:

i. Computation of the total, alpha-band, and beta-band connectivity matrices, by integrating the connectivity matrix A_i ($\forall i$), across the entire frequency axis (1–40 Hz), the alpha-band frequency range (8–12 Hz), and the beta-band frequency range (12–30 Hz), respectively. Thus, we obtained $A_{i,tot}$, $A_{i,\alpha}$, $A_{i,\beta}$ connectivity matrices ($\in \mathbb{R}^{R\times R}$).
ii. Computation of centrality indices derived from the graph theory, to synthetize some changes in the topology of the brain connectivity network between the motor imagery conditions. The brain connectivity can be described by a weighted graph, where the magnitude of the connectivity between two brain regions (EEG sensors or cortical ROIs) is represented as the weight of an edge, while the two brain regions connected by an edge represent two nodes of the graph. A centrality index provides a measure of importance of a particular node in the graph. By considering all nodes in the brain graph, a centrality measure can be represented by an array $\in \mathbb{R}^R$. Here we focus on centrality indices that take into account the direction of connections, specifically the *in degree*, *out degree*, *authority* and *hubness*, and were computed on each connectivity matrix derived from the previous point i. (i.e., on $A_{i,tot}$, $A_{i,\alpha}$, $A_{i,\beta}$). These indices were considered as prior studies that analyzed changes in the brain network connectivity during motor imagery observed a peculiar directionality in the connectivity pattern[18–20]. The mathematical formulation of these indices is provided in the following, by considering a generic adjacency matrix A (one among $A_{i,tot}$, $A_{i,\alpha}$, $A_{i,\beta}$), i.e., a matrix defined by all edges' weights, with the jk-th element of the matrix ($A[j,k]$) representing the weight of the edge connecting the node j to node k, where a node corresponds to a brain region, either on the scalp or on the cortex. The in degree and out degree (in_k and out_k) for a node (node k) are the sum of the weights of the edges entering and exiting from that node, respectively:

$$in_k = \sum_j A[j,k], \tag{4}$$

$$out_k = \sum_j A\left[k, j\right]. \tag{5}$$

Thus, the in degree and out degree provide a direct interpretation of the nodes most involved in the reception (in degree) and transmission (out degree) of information, by equally weighting the nodes in the graph in the same way across conditions. Authority and hubness are recursive measures that provide a more refined concept of reception and transmission of information than the previous ones. Specifically, the authority ($auth_k$) of a node (node k) is the sum of the weights of edges entering a node, multiplied by the hubness of the node the edge originates from. On the other hand, the hubness (hub_k) of a node (node k) is the sum of the weights of edges exiting from a node, multiplied by the authority of the node the edge points to.

$$auth_k = \sum_j A\left[j, k\right] hub_j, \tag{6}$$

$$hub_k = \sum_j A\left[k, j\right] auth_j. \tag{7}$$

Like the in degree and out degree, these measures provide insights about the nodes that are mostly involved in the reception (authority) and transmission (hubness) of the information, but they take mutually into account the centrality of receiving and sending nodes. This results in a weighted computation, where the weights are specific of the adjacency matrix A on which the centrality index is computed, related to a specific condition. The recursive formulation implies that strong connections exist between nodes with high authority and nodes with high hubness. This could be useful to further emphasize any existing directionality in the connectivity pattern. As the centrality indices were computed separately for each brain region (i.e., graph node), each centrality index resulted into an array $\in \mathbb{R}^R$.

iii. Averaging of the centrality indices computed at point ii. across the examples (i.e., EEG epochs) belonging to the same motor imagery condition, i.e., $\forall i | y_i = \text{right-hand}$, and $\forall i | y_i = \text{left-hand}$.

Therefore, after this procedure, for each experiment (scalp-level and cortex-level), each participant and each motor imagery condition, we derived a set of 4 centrality measures (in degree, out degree, authority, hubness), separately for the brain connectivity matrix representing the total connectivity, alpha-band connectivity, and beta-band connectivity. This results in a total of 12 *graph theory-based centrality measures* (6 inflow and 6 outflow measures).

For each experiment (scalp-level and cortex-level), and each measure derived from the graph theory (12 in total) and from FCNet (64 in total, see Sect. "Analysis of the output of intermediate interpretable layers"), we compared each inflow/outflow measure between the right-hand and left-hand motor imagery condition, by performing a paired permutation test based on t-statistic (two-tail test, 10000 permutations), employing the tmax method for adjusting p-values for multiple comparisons (R tests, one per brain region, for each measure)[62]. The Cohen's d was used to quantify the effect size[63], i.e., $d = t/\sqrt{s}$, where t is the t-value and s is the sample size ($s = 9$ and $s = 54$, corresponding to the number of participants respectively in the scalp-level and cortex-level experiments).

Comparative analysis of FCNet vs. other decoding approaches

In this study we are not focusing on proposing a new approach for improving the decoding performance in motor imagery neural decoding; rather, the key contribution point is the design and exploitation of an interpretable and explainable artificial intelligence approach to analyze functional connectivity during motor imagery (i.e., neural data analysis). However, even though it is not our main aim, as a secondary and additional contribution, we also performed a comparative analysis of FCNet vs. other neural decoders. To this aim, we compared the decoding performance scored by FCNet – on top of which the proposed data-driven framework is built – with two variants of FCNet and with nine other decoding approaches widely adopted in the literature for decoding motor imagery (4 machine learning algorithms and 5 deep neural networks). Specifically, we compared FCNet with:

i. FCNet variants. Two non-interpretable variants of FCNet were considered, by changing the interpretable inflow/outflow convolutions of the network (spatial block), each performing a 2D convolution along one of the two dimensions of the high-level connectivity feature maps $(32, R, R)$ provided by the spectral block (see the interpretable connectivity inflow/outflow branches of the spatial block in Sect. "Network architecture and training"). Specifically, a first non-interpretable variant was obtained by redesigning the spatial block with a single 2D convolutional layer learning 32 kernels with size $W = (R, R)$, activated via ReLU. The convolutional layer was realized using separable convolutions (i.e., depthwise convolution combined with pointwise convolution), as in the original version. A feature array of 32 elements was returned, which was provided as input to the classification block. This way, the network learned patterns on the high-level connectivity feature maps $(32, R, R)$ in a mixed way across brain interactions, without exploiting an interpretable feature learning that enables the distinction between inflow/outflow connectivity features. Finally, a second non-interpretable variant was obtained by redesigning the spatial block with a single 2D average pooling layer (pooling size and stride of $W = S = (R, R)$). The high-level connectivity feature maps $(32, R, R)$

were summarized with a feature array of 32 elements, each containing the mean connectivity value across all brain interactions, and this array was provided as input to the classification block. Like the previous variant, the network summarized the feature maps in a mixed way across brain interactions, without exploiting an interpretable feature learning of inflow/outflow connectivity. The main difference between the variants is that, in the second variant the network does not learn the optimal recombination of the high-level connectivity feature maps across brain interactions, but simply computes the mean value.

ii. State-of-the-art decoders. We selected 9 state-of-the-art decoders based on prior studies. Four of these algorithms were machine learning solutions proposed for decoding graph theory indices derived from brain connectivity networks. The remaining algorithms were deep neural networks proposed for decoding EEG multi-variate time series. As concerning the machine learning algorithms, graph theory measures were extracted, quantifying the inflow and outflow (in degree and out degree, or authority and hubness), optionally used in combination with the clustering coefficient – an index that indicates the clustering degree (i.e., the segregation degree) of the nodes of a brain network – and classified by using linear discriminant analysis (LDA), inspired from previous decoding studies[39,42]. Among the deep neural networks, we included in the benchmark ShallowFBCSPNet and Deep4Net[64] – the first successful CNNs proposed for motor imagery decoding – together with EEGNet[36], a multi-purpose CNN reaching state-of-the-art decoding performance on a variety of tasks (e.g., decoding of motor imagery, P300, steady-state visual evoked potential)[65], also winning international EEG decoding competitions[66,67]. Finally, we also included recent CNN designs that exploit inception modules for learning temporal features, such as EEGInception[68] and EEGITNet[69].

All decoders were trained and tested as FCNet, to provide a fair comparison, see Sect. "Network architecture and training". Then, the decoding accuracy scored by FCNet was compared with the one scored by each of the other algorithms (2 from point i. and 9 from point ii.) by performing a Wilcoxon signed-rank test (two-tail test). To correct p-values for multiple tests (11 in total) we adopted the Benjamini-Hochberg procedure.

Results

Motor imagery-induced changes in brain connectivity: results from graph theory

The graph theory-based measures extracted from the functional connectivity estimates are displayed in Figs. 3 and 4, respectively for inflow and outflow quantifications; specifically, the difference between right-hand and left-hand motor imagery conditions is shown, both as to the scalp-level experiment (Figs. 3a and 4a) and cortex-level experiment (Figs. 3b and 4b).

In both scalp-level and cortex-level experiments, inflow measures were modulated depending on the motor imagery condition, with an increase of functional connectivity entering the sensorimotor regions (e.g., scalp-level: C3/C4 and C5/C6; cortex-level: precentral gyrus, postcentral gyrus, paracentral lobule) in the right hemisphere and a concurrent decrease in the left hemisphere, in the right-hand condition compared to the left-hand condition. However, while at the cortex level the modulation was significant in all the considered frequency ranges, i.e., the entire frequency axis, alpha-band, and beta-band, this was not observed at the scalp level, where statistical significance was observed mainly in the entire frequency range.

A similar lateralization was in general observed in outflow measures: the exiting connectivity tended to increase in the sensorimotor regions of the right hemisphere and to decrease in the ones of the left hemisphere for right-hand condition vs. left-hand condition. This was observed at the cortex level consistently in all the considered frequency ranges (entire frequency range, alpha-band and beta-band). On the contrary, scalp level outflow measures exhibited less consistent results and with lower statistical significance (only one significant brain region in the entire frequency range), resembling those at cortex level only as to the alpha band, but with no statistical significance. Thus, overall, inflow and outflow connectivity tended to increase in the hemisphere ipsilateral to the movement relative to the contralateral one.

Motor imagery-induced changes in brain connectivity: results from the deep learning approach

First, we show the performance of FCNet obtained when decoding motor imagery from connectivity estimates. Figure 5 displays the decoding accuracy scored by the model trained on each participant, separately for the scalp-level and cortex-level experiments. Additionally, we also report the average confusion matrices, across participant-specific models. From Fig. 5, all the trained models performed well-above the chance level (0.5), achieving across participants an accuracy of 0.780 (0.031) and of 0.754 (0.016), respectively for the scalp-level and cortex-level experiments (mean value and standard error of the mean within brackets). Therefore, FCNet was able to detect the changes in the brain connectivity due to the different motor imagery conditions in both experiments (see Supplementary Fig. S1 for the modulations that were observed in the connectivity matrices).

Once checked that FCNet is able to detect motor imagery modulations from connectivity estimates, its knowledge can be exploited in the analysis framework. Figure 6 shows a visualization of the spectral relevance attributed by FCNet while decoding motor imagery. The highest relevance was attributed to the upper alpha-band (peaking at approx. 10–12 Hz) and to the beta-band (peaking at approx. 20–22 Hz), consistently across scalp-level and cortex-level experiments. Additionally, in Fig. 7 we visualize the connectivity inflow measure and outflow measure that resulted the most relevant among the 32 measures extracted by the inflow convolutional layer and by the outflow convolutional layer, respectively. The most relevant measures extracted by FCNet exhibited a hemispheric lateralization, with large and significant differences between right and left motor imagery mainly at sensorimotor regions (central and posterior regions), in the experiments at both the scalp and cortex level.

Figures 8 and 9 report the results of the statistical analysis obtained when contrasting connectivity inflow/outflow measures between the two motor imagery conditions. This analysis was applied separately to each

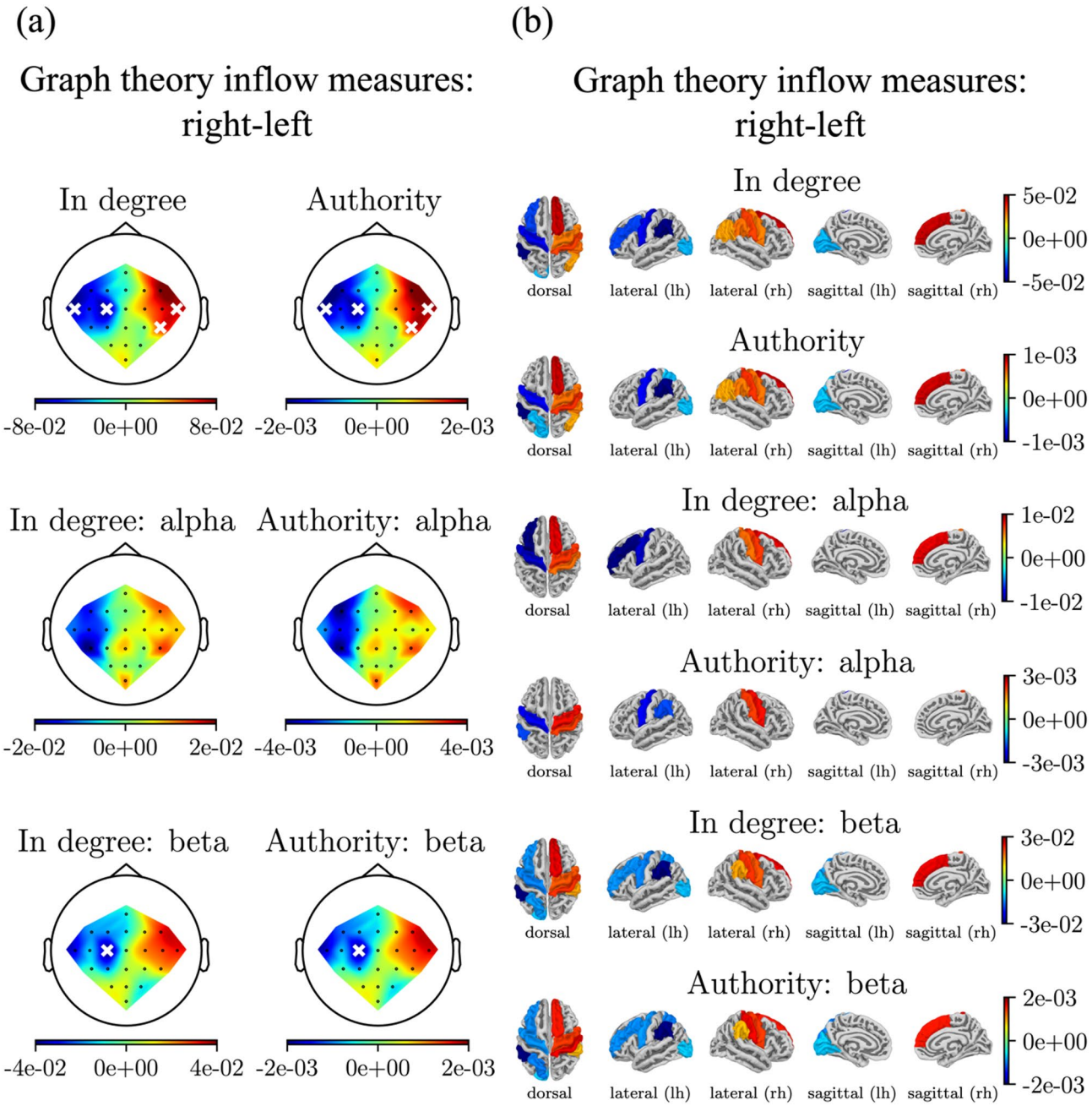

Fig. 3. Graph theory-based connectivity inflow measures: in degree and authority (entire frequency range, alpha-band, beta-band). Here, for each measure, the difference between the right-hand and left-hand motor imagery conditions is displayed, on average across participants, as to the scalp-level experiment (**a**) and cortex-level experiment (**b**). In cortex-level representations, the dorsal, lateral, and sagittal views are shown; the last two views are displayed separately for the left hemisphere (lh) and right hemisphere (rh). The results of the conducted statistical analyses are reported too, by marking with a white cross (panel a) or by color-highlighting (panel b) the regions that were significantly different ($p < 0.05$) between the two motor imagery conditions.

inflow/outflow measure, using both graph theory-based measures (classic approach) and FCNet-based measures (deep-learning enriched approach). Specifically, Figs. 8 and 9 display the absolute value of Cohen's d-value and the p-value averaged across the brain regions that exhibited a significant difference between the two motor imagery conditions. This visualization enables to understand the strength of the effect size (d-value) and of the statistical significance (p-value) of inflow/outflow measures when detecting changes between motor conditions. Notably, the proposed deep learning-enriched approach provided connectivity inflow and outflow measures with both high effect size and statistical significance, similarly to the classic approach, in both scalp-level and cortex-level experiments. The same was observed also when considering the maximum Cohen's d-value (in its absolute value) – instead of the average – across the significant brain regions, see Supplementary Figs. S2 and S3.

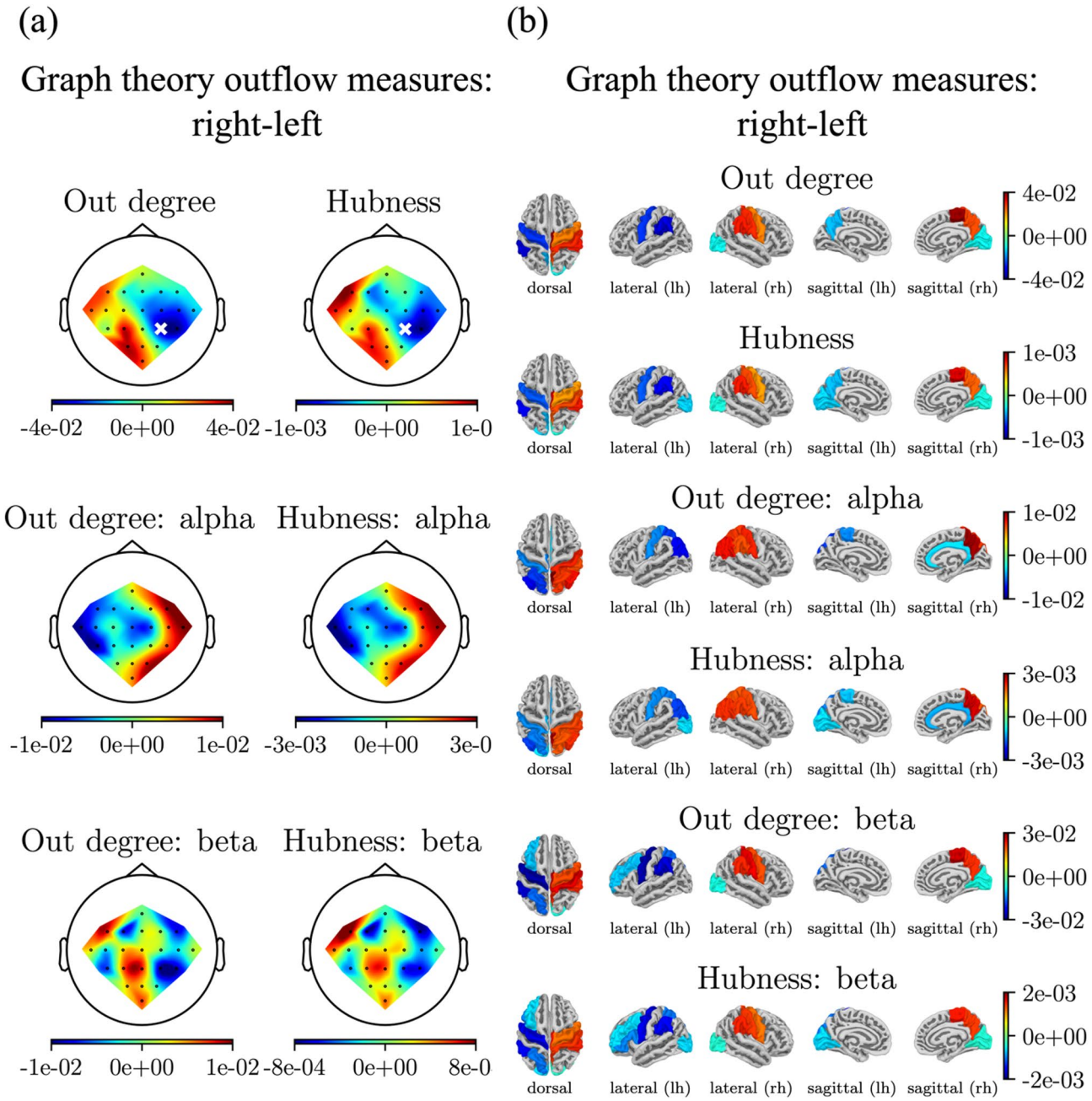

Fig. 4. Graph theory-based connectivity outflow measures: out degree and hubness (entire frequency range, alpha-band, beta-band). Here, for each measure, the difference between the right-hand and left-hand motor imagery conditions is displayed, on average across participants, as to the scalp-level experiment (**a**) and cortex-level experiment (**b**). In cortex-level representations, the dorsal, lateral, and sagittal views are shown; the last two views are displayed separately for the left hemisphere (lh) and right hemisphere (rh). The results of the conducted statistical analyses are reported too, by marking with a white cross (panel a) or by color-highlighting (panel b) the regions that were significantly different ($p < 0.05$) between the two motor imagery conditions.

Finally, as the scalp-level experiments were conducted on a small dataset (BNCI2014-001 dataset[51], consisting of 9 participants), in Figs. 10 and 11 we report the results obtained while replicating our experiments at the scalp-level on a larger population. In this case, we adopted the same dataset used previously in the cortex-level experiments (Lee2019-MI dataset[52], consisting of 54 participants), limiting the analysis only to the scalp-level EEG signals, i.e., by preparing the signals as specified in Sect. "EEG pre-processing and processing" point ii, but without performing the source reconstruction. Figure 10 reports the decoding performance (accuracy and confusion matrix) scored by FCNet in this additional experiment at the scalp-level. In this case too, the scalp-level neural decoders performed significantly above the chance level (0.5), achieving an accuracy of 0.710 (0.014) across participants (mean value and standard error of the mean within brackets). Figure 11 reports the results of

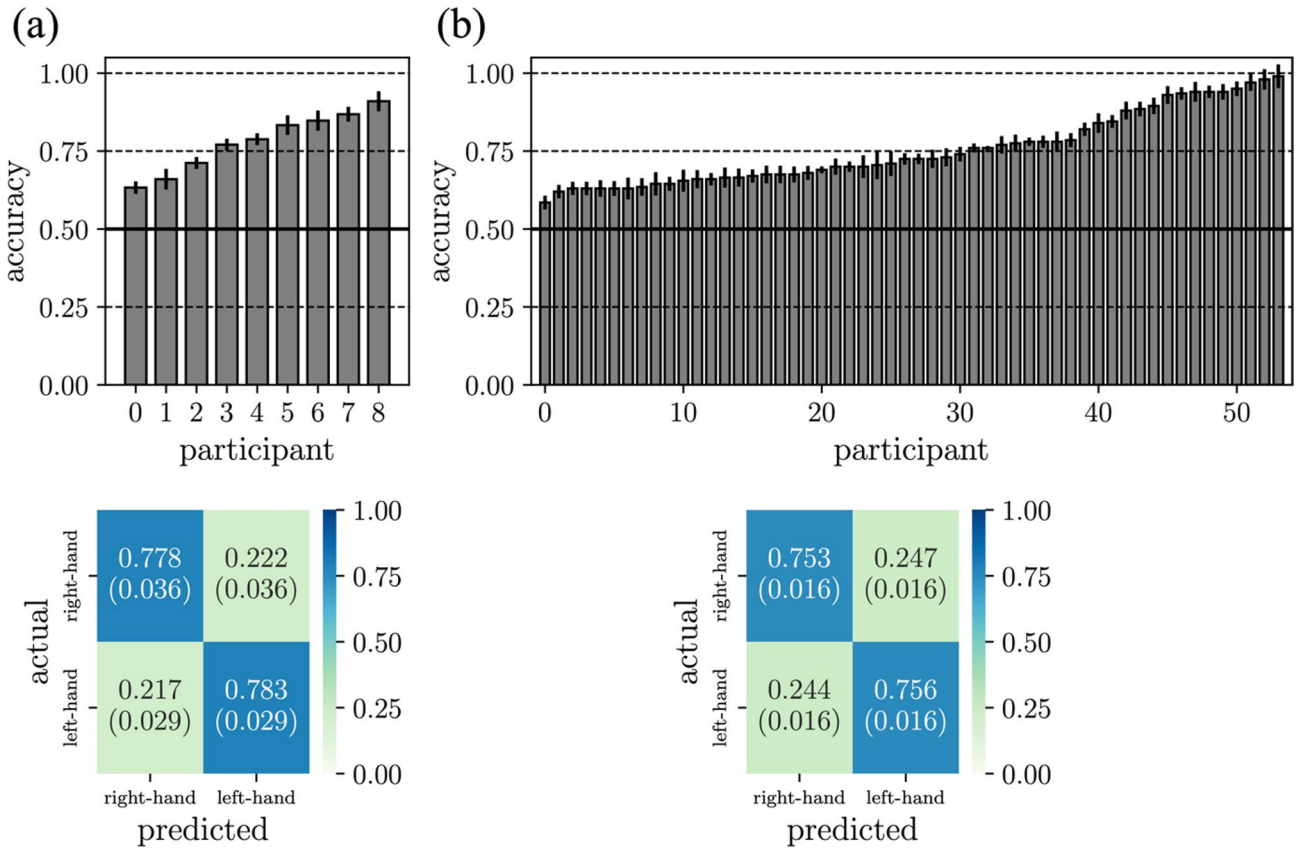

Fig. 5. FCNet decoding performance. The decoding accuracy and the confusion matrix are reported, separately for the scalp-level experiment (**a**) and the cortex-level experiment (**b**). The accuracy is reported separately for each participant. The bar height denotes the mean value, and the error bar the standard error of the mean across cross-validation folds. The black horizontal line represents the chance level (0.5). Participant distributions are displayed sorted from the least accurate to the most accurate. Each cell of the confusion matrix displays the mean value and standard error of the mean (within brackets) across participants.

the statistical analysis obtained when contrasting connectivity inflow/outflow measures between the two motor imagery conditions, in this additional scalp-level experiment. Even when applying our deep learning-enriched approach to this new and larger scalp-level EEG dataset, the FCNet-derived connectivity inflow (Fig. 11a) and outflow (Fig. 11b) measures resulted to have both high effect size and statistical significance, like the classic approach. Collectively, these performance results and statistical analyses results further strengthen the statistical inference of our scalp-level experiments.

Comparative analysis of FCNet vs. other decoding approaches

The results from the performance comparative analysis are reported in Table 3, for both scalp-level and cortex-level experiments.

From the comparative analysis, FCNet significantly outperformed ($p < 0.05$) the two considered FCNet variants, which were designed by changing the interpretable connectivity inflow/outflow branches, replacing interpretable components with non-interpretable components. Therefore, the interpretable components included in FCNet architecture promote not only interpretability of connectivity inflows/outflows, but also an improved decoding accuracy. To further inspect the beneficial effect of FCNet interpretable components, emerging in Table 3 in terms of decoding performance, we compared the spectral relevance resulting from FCNet with the one resulting from the two variants of FCNet. These are reported in Fig. 12 for both scalp-level and cortex-level experiments. From Fig. 12, the frequency-domain processing operated by FCNet (black line), thanks to the inclusion of the interpretable connectivity inflow/outflow elements, attributed a higher importance to alpha- and beta-band frequency components compared to the variants (red and blue lines), consistently across scalp-level and cortex-level experiments.

Finally, FCNet also significantly outperformed ($p < 0.05$) the considered existing learning systems (including both machine learning and deep neural networks) in the cortex-level experiments, and it significantly outperformed ($p < 0.05$) all learning systems except for ShallowFBCSPNet ($p = 0.2$) in the scalp-level experiment. However, it is worth highlighting that ShallowFBCSPNet is characterized by a network architecture specific for extracting features related to sensorimotor rhythms, and thus, for solving motor imagery decoding. On the other hand, FCNet is designed without any constraint in the learned features and can be potentially

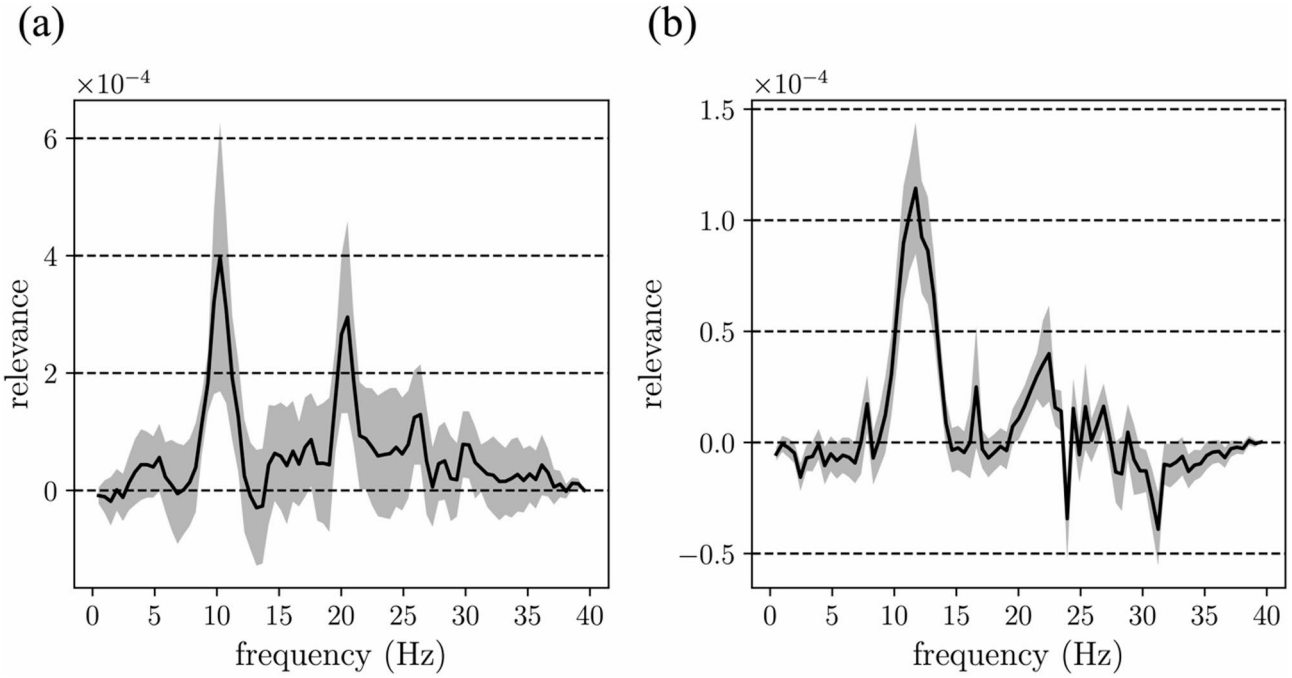

Fig. 6. FCNet-based spectral relevance. The relevance attributed by FCNet in the frequency domain is reported separately for the scalp-level experiment (**a**) and cortex-level experiment (**b**). The tick black line denotes the mean value and the overlayed gray area the standard error of the mean across participants.

used in the future also for solving other neural decoding problems. From this analysis, FCNet emerges as an architecture that significantly outperforms state-of-the-art solutions consistently across the two motor imagery datasets exploited in our experiments, improving the accuracy of about 18% and 4% with respect to the most accurate state-of-the-art machine learning and deep learning approach, respectively. Thus, FCNet is not only characterized by interpretability properties, valuable for its use as an analysis tool, but also by superior performance compared to other state-of-the-art motor imagery decoders commonly used in the literature.

Discussion

In this study, we designed a deep learning-based framework for analyzing directed functional connectivity estimates – quantified by spectral Granger causality – based on a novel interpretable convolutional neural network, named FCNet. This neural network automatically learns, from the connectivity estimates, non-linear relationships for maximizing the between-class discriminability (i.e., to optimally separate the output brain states) at first in the frequency domain, and then in the spatial domain. In the latter, FCNet employs interpretable components, summarizing the entering and the exiting connections across brain regions from the processed connectivity matrices; thus, the neural network is able to learn the connectivity inflow and outflow measures that are most class-discriminative. In our analysis framework, FCNet-based inflow and outflow measures are extracted and used; moreover, the network decision is explained to reveal the most relevant frequency components and the most relevant inflow/outflow measures for predicting the correct output brain state. Our framework was applied on connectivity matrices estimated from real EEG data recorded while participants performed a motor imagery task, by illustrating its potentialities at both the scalp-level and cortex-level. It is crucial to notice that our study does not focus on proposing a new neural network architecture aimed at improving motor imagery decoding; rather, it is intended to be a proof-of-concept study in which we propose a novel connectivity analysis framework based on the knowledge learned by an interpretable convolutional neural network, and we illustrate how neuroscientists can use it on real data collected during a cognitive task (specifically, a motor imagery task) for characterizing brain functional interactions in a data-driven way.

Motor imagery-induced changes in brain connectivity: short literature overview

Prior studies on the modulation of EEG oscillations highlighted a key role of alpha-band and beta-band oscillations[25,26] in both motor execution and motor imagery, reporting attenuated amplitude oscillations in these frequency ranges at sensorimotor areas (e.g., central electrode sites at scalp level, and pre-central and post-central gyri at cortex level)[7,58,70,71]. Studies on scalp-level directed functional connectivity[18–20] evidenced a strong directional connectivity from central to right and from central to left EEG electrodes during both right-hand and left-hand motor imagery. Additionally, a strong connectivity from left to right electrodes was also observed for the right-hand motor imagery condition, vice versa for the left-hand imagery. Therefore, these EEG studies point to a peculiar pattern in the directionality of functional connectivity occurring in right-hand vs.

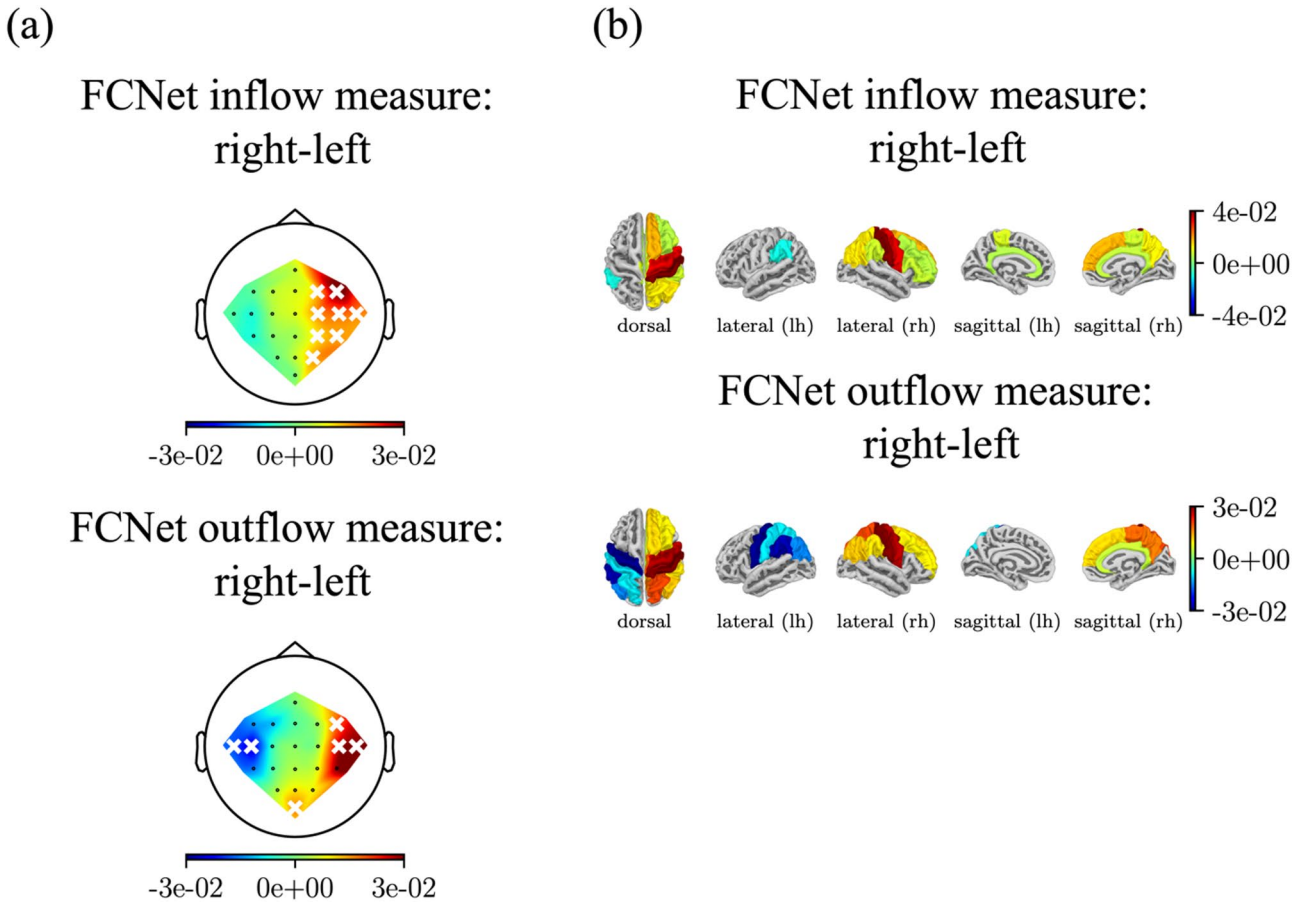

Fig. 7. Most relevant FCNet-based connectivity inflow and outflow measures. Here, for each of the most relevant inflow and outflow measure extracted by FCNet, the difference between the right-hand and left-hand motor imagery conditions is displayed, on average across participants, as to the scalp-level experiment (**a**) and cortex-level experiment (**b**). In cortex-level representations, the dorsal, lateral, and sagittal views are shown; the last two views are displayed separately for the left hemisphere (lh) and right hemisphere (rh). The results of the conducted statistical analyses are reported too, by marking with a white cross (panel a) or by color-highlighting (panel b) the regions that were significantly different ($p < 0.05$) between the two motor imagery conditions.

left-hand motor imagery, with the right hemisphere being characterized by an increased connectivity inflow in the right-hand imagery vs. left-hand imagery, and vice versa for the left hemisphere.

Also prior fMRI studies highlighted a lateralization of the connectivity pattern during right vs. left motor imagery[72,73]. Specifically, Zhang et al.[72] found a right-to-left connectivity pattern in left-hand motor imagery – especially emerging from the right supplementary motor area – and a left-to-right pattern in right-hand motor imagery – especially from the left supplementary motor and premotor areas. Moreover, Ogawa et al.[73] reported a connectivity outflow (out degree) stronger for right-hand vs. left-hand motor imagery in the right motor-related regions (right primary motor cortex, right inferior parietal cortex, right supplementary motor area, right dorsal premotor cortex), and stronger for left-hand vs. right-hand motor imagery in the left motor-related regions (left primary motor cortex, left inferior parietal cortex). In the same study, the authors reported a connectivity inflow (in degree) stronger for right-hand vs. left-hand motor imagery mainly in the right motor-related regions (right primary motor cortex, right inferior parietal cortex).

Motor imagery-induced changes in brain connectivity: results from graph theory

In the present study, a classic analysis approach based on graph theory measures (Figs. 3 and 4) revealed the same directionality pattern in the connectivity at the scalp level as in the prior EEG studies[18–20], by exploiting the same dataset (BNCI2014-001[51]). Indeed, we observed a significantly higher in degree and authority (i.e., inflow) at right (/left) electrode sites in the right-hand (/left-hand) motor imagery compared to the left-hand (/right-hand) motor imagery condition. This was obtained, not only within the entire frequency range, but also in alpha-band and beta-band, although with less statistical significance (i.e., no significant differences in alpha-band and fewer significant regions in beta-band). In the alpha-band, the same right-hemisphere regions exhibited also higher out degree and hubness (i.e., outflow) in right-hand vs. left-hand motor imagery, even though these results were not statistically significant. Interestingly, we observed the same modulation also at the cortex level utilizing a different dataset (Lee2019-MI[52]), with a wide statistical significance across the considered frequency

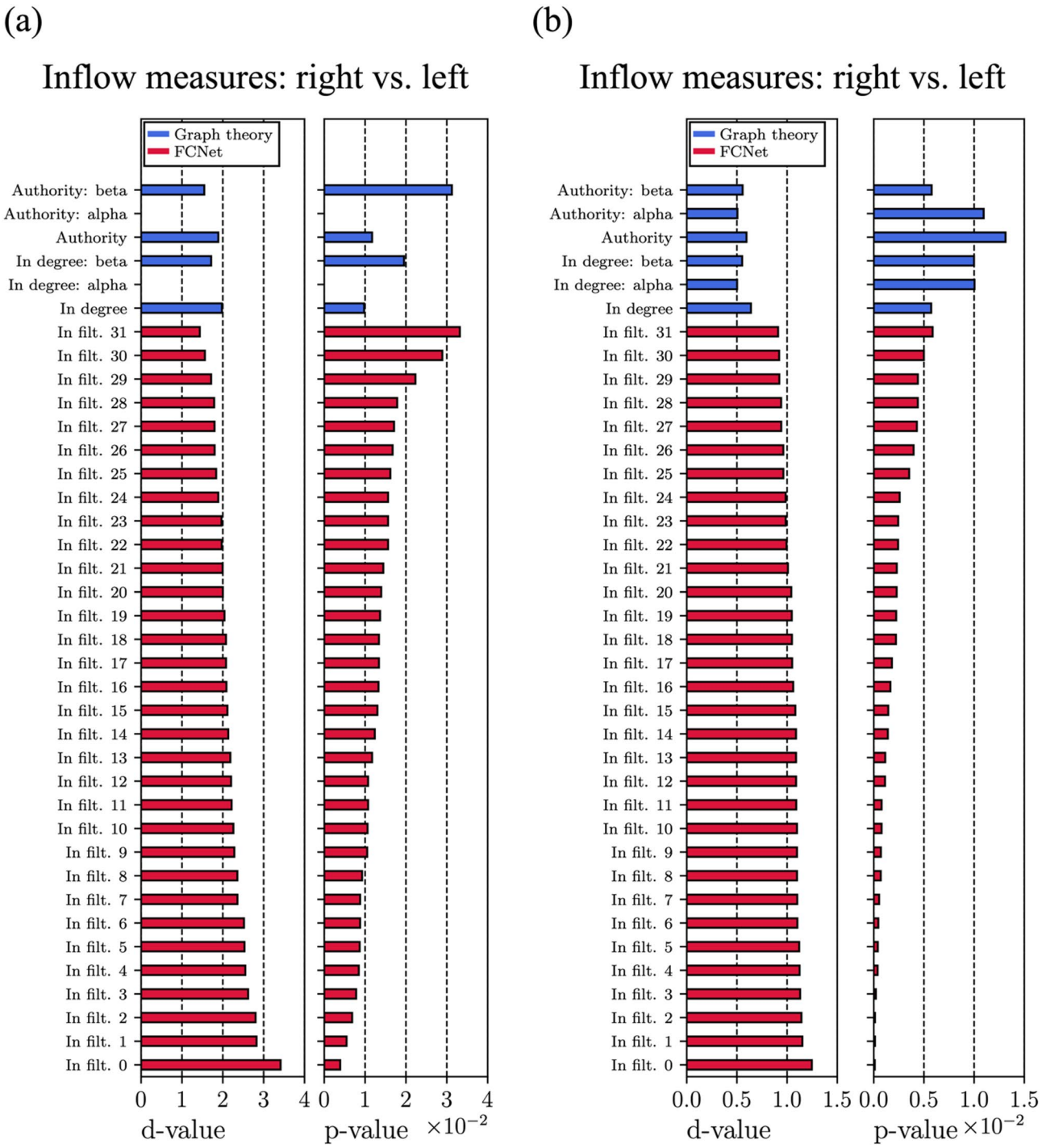

Fig. 8. Inflow measures: average absolute d-value and p-value across the significant brain regions. The figure displays the effect size and statistical significance while comparing the two motor imagery conditions utilizing inflow measures obtained with the classic approach and with the deep learning-enriched approach, separately as to scalp-level experiment (**a**) and cortex-level experiment (**b**). In the classic approach the measures were the in degree and the authority, computed across the entire frequency axis, within alpha-band and beta-band. In the deep learning-enriched approach the measures were the 32 FCNet-based inflow measures. For each measure, the absolute Cohen's d-value and the p-value are averaged across the brain regions showing statistically significant difference ($p < 0.05$) between the two conditions. To ease the readability, FCNet measures are sorted from the lowest to the highest effect size/statistical significance, separately in each plot.

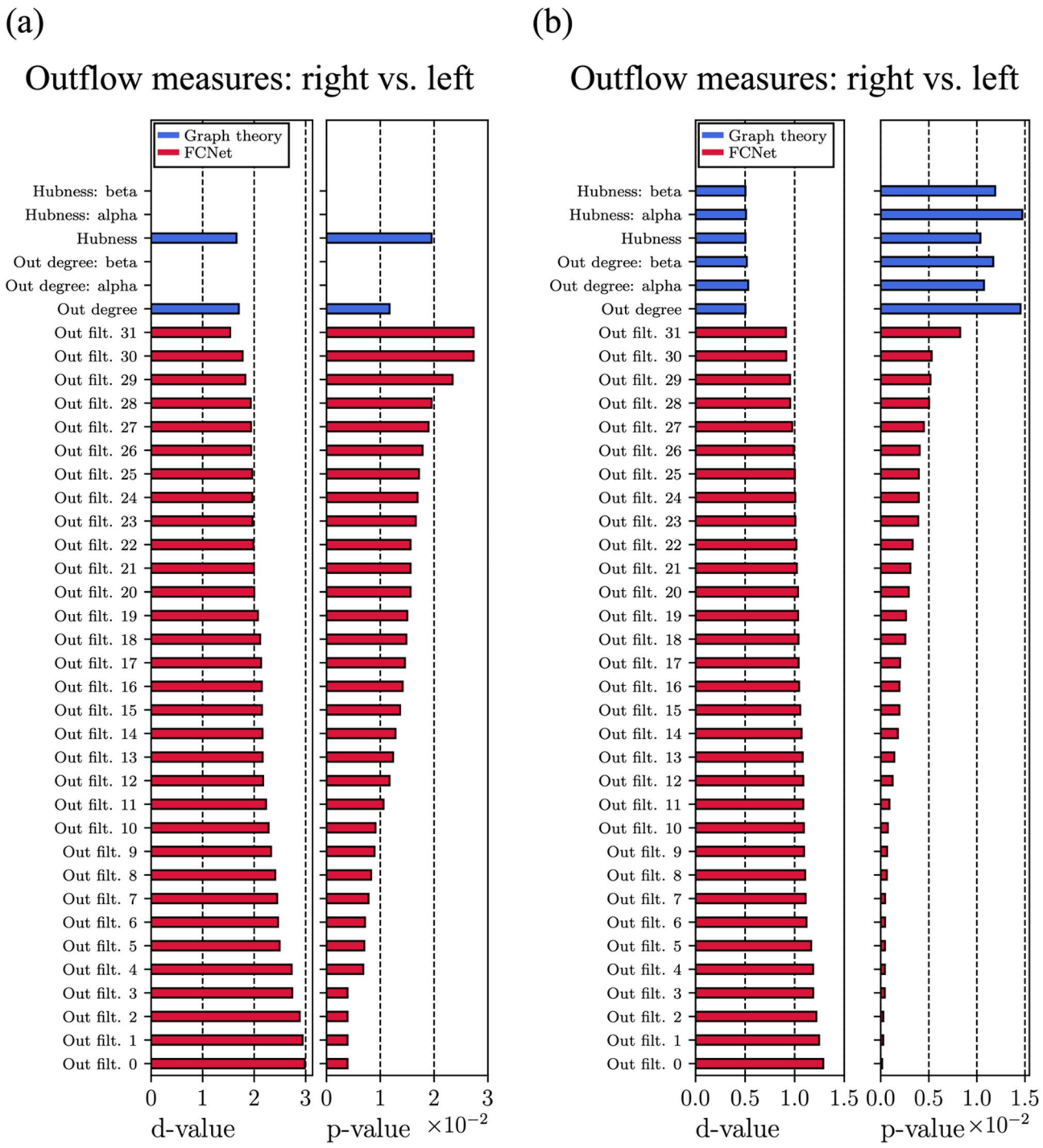

Fig. 9. Outflow measures: average absolute d-value and p-value across the significant brain regions. The figure displays the effect size and statistical significance while comparing the two motor imagery conditions utilizing outflow measures obtained with the classic approach and with the deep learning-enriched approach, separately as to scalp-level experiment (**a**) and cortex-level experiment (**b**). In the classic approach the measures were the out degree and the hubness, computed across the entire frequency axis, within alpha-band and beta-band. In the deep learning-enriched approach the measures were the 32 FCNet-based outflow measures. For each measure, the absolute Cohen's d-value and the p-value are averaged across the brain regions showing statistically significant difference ($p < 0.05$) between the two conditions. To ease the readability, FCNet measures are sorted from the lowest to the highest effect size/statistical significance, separately in each plot.

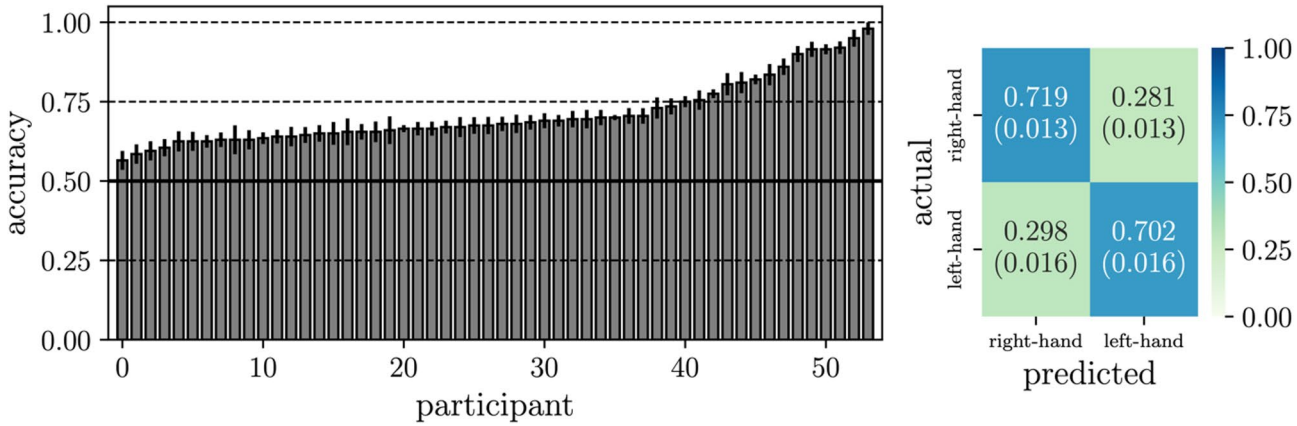

Fig. 10. FCNet decoding performance on a larger scalp-level dataset. See the caption of Fig. 5 for further details.

ranges (entire range, alpha-band, and beta-band), involving mainly the left and right motor-related cortices (pre-central and post-central gyri, parietal areas). The similar results across the two experiments were expected, as the data derived from two similar motor imagery tasks. It is worth highlighting that the slight differences observed – for example related to outflow measures – might be related to the lower number of participants included in the dataset used in the scalp-level analysis vs. cortex-level analysis (9 vs. 54), and to the processing operated for deriving cortex-level activity. Indeed, the reconstruction of cortical activity may have enhanced the task-relevant information from the EEG signals, leading to higher statistical results.

Motor imagery-induced changes in brain connectivity: results from the deep learning approach

As concerning the results from the analysis conducted with the proposed deep learning-enriched approach, the spectral relevance visualization (Fig. 6) confirmed that components in alpha-band primarily, and beta-band secondarily, contributed more to the discrimination between right-hand and left-hand motor imagery conditions. Additionally, when considering the FCNet-based inflow and outflow measures that resulted most relevant for the discrimination between right-hand and left-hand motor imagery (Fig. 7), we observed a similar pattern to the ones derived via graph theory analysis (Figs. 3 and 4). It is interesting to notice here that, while the outflow measures highlighted both the significant increased outflow in the right hemisphere and significant decreased outflow in the left hemisphere for right-hand vs. and left-hand motor imagery (Fig. 7 – bottom panels), inflow measures highlighted mainly the significant increased inflow occurring in the right hemisphere (Fig. 7 – upper panels). This result on the connectivity inflow suggests a stronger effect in the right hemisphere. Overall, by revealing the most relevant inflow and outflow measures for motor imagery, the deep learning approach matches the connectivity results found in past literature (see Sect. "Motor imagery-induced changes in brain connectivity: short literature overview"). Indeed, a hemispheric lateralization – reported in prior EEG and fMRI studies analyzing both connectivity patterns and centrality measures during motor imagery – emerges also in our results. The main involved regions were, at the scalp level, central electrode sites (as in prior EEG studies), and at the cortex level, the precentral gyrus, postcentral gyrus, parietal regions (e.g., inferior parietal cortex, supramarginal gyrus) and, to a less extent, the superior frontal gyrus and paracentral lobule. These ROIs included the primary motor cortex, supplementary motor area and inferior parietal cortex, in line with prior fMRI studies.

Finally, when comparing each inflow/outflow measure between motor imagery states in the performed statistical analysis (Figs. 8 and 9, and 11), all the considered measures exhibited a strong effect size and statistical significance in detecting right-hand vs. left-hand motor imagery differences, consistently across the scalp-level and cortex-level experiments. Thus, the proposed deep learning approach was able to extract inflow/outflow measure sensitive to changes between motor imagery states, like the classic approach. This is an expected result, as FCNet was trained to optimally distinguish the considered motor imagery states; thus, defining inflow/outflow measures based on FCNet features should reflect the network ability to optimally separate motor states. Remarkably, FCNet measures are based on the non-linear processing of brain functional connectivity operated by the network, by optimally combining the information in the causal spectra across the different frequencies (as operated in the spectral block of FCNet), and by optimally weighting the different brain regions in the computation of inflow and outflow measures (as operated in the spatial block of FCNet, in the learned convolutional kernels of the interpretable layers).

Overall, these results suggest that FCNet is useful for assessing the informativeness of brain functional networks related to cognitive states (in this case, right-hand motor imagery and left-hand motor imagery), quantifying the predictive power of frequency components of causal spectra (FCNet-based spectral relevance) and of connectivity inflows/outflows (FCNet-based connectivity inflow and outflow measures).

Finally, it is worth mentioning that these considerations are also supported by the high decoding accuracy scored by FCNet – quantifying the predictive capability of the entire learning system. FCNet architecture

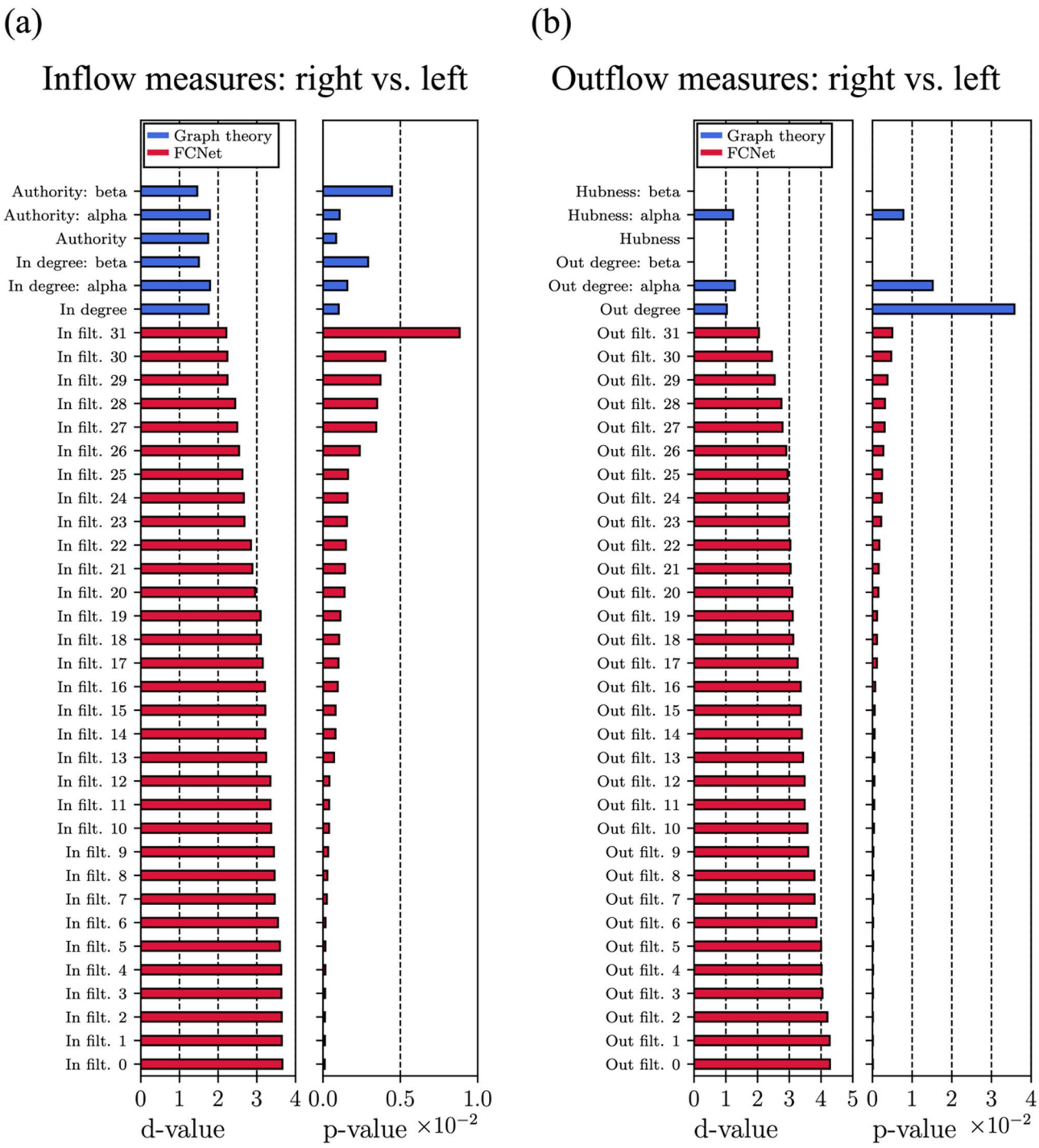

Fig. 11. Inflow and outflow measures on a larger scalp-level dataset: average absolute d-value and p-value across the significant brain regions. Inflow and outflow measures are reported in (**a**) and (**b**), respectively. See the caption of Figs. 8 and 9 for further details.

maximizes both interpretability and decoding performance. Indeed, the FCNet non-interpretable variants degraded both the decoding accuracy (Table 3) and the neurobiological coherence of explanations (Fig. 12), as they exhibited attenuations of spectral relevance in alpha and beta bands (i.e., relevance shifts away from alpha and beta bands). Moreover, the developed algorithm, not only provided useful insights about the modulations of oscillatory interactions in a data-driven way, but also resulted the most accurate neural decoder among different alternative neural decoders. Indeed, the benchmark analysis on decoding performance (Table 3) showed that the proposed decoder is significantly superior to both state-of-the-art machine learning and deep learning pipelines.

Decoding model	Input type	Scalp-level experiment		Cortex-level experiment	
		Accuracy (mean (SEM))	*P*-value	Accuracy (mean (SEM))	*P*-value
FCNet (proposed)	Functional connectivity	**0.780 (0.031)**	–	**0.754 (0.016)**	–
FCNet non-interpretable variant 1	Functional connectivity	0.686 (0.031)	**4.77·1e−3**	0.628 (0.015)	**3.15·1e−10**
FCNet non-interpretable variant 2	Functional connectivity	0.534 (0.014)	**4.77·1e−3**	0.530 (0.008)	**3.15·1e−10**
InOutDegrees+LDA[39,42]	Functional connectivity	0.637 (0.032)	**4.77·1e−3**	0.578 (0.012)	**3.15·1e−10**
InOutDegreesClust+LDA[39,42]	Functional connectivity	0.635 (0.026)	**4.77·1e−3**	0.575 (0.012)	**3.15·1e−10**
AuthHub+LDA[39,42]	Functional connectivity	0.638 (0.029)	**4.77·1e−3**	0.579 (0.012)	**3.15·1e−10**
AuthHubClust+LDA[39,42]	Functional connectivity	0.630 (0.032)	**4.77·1e−3**	0.573 (0.011)	**3.15·1e−10**
ShallowFBCSPNet[64]	Multi-variate time series	0.743 (0.053)	2.03·1e−1	0.735 (0.017)	**4.08·1e−2**
Deep4Net[64]	Multi-variate time series	0.557 (0.018)	**4.77·1e−3**	0.560 (0.013)	**3.78·1e−10**
EEGNet[36]	Multi-variate time series	0.664 (0.024)	**8.59·1e−3**	0.601 (0.016)	**4.15·1e−9**
EEGInception[68]	Multi-variate time series	0.567 (0.014)	**4.77·1e−3**	0.590 (0.017)	**8.07·1e−9**
EEGITNet[69]	Multi-variate time series	0.66 (0.024)	**4.77·1e−3**	0.573 (0.014)	**4.15·1e−9**

Table 3. FCNet decoding performance: comparison with FCNet variants and existing approaches. The accuracy of each learning system is reported in its mean value and standard error of the mean across participants (within brackets). Bold accuracy values mark the most accurate models. The p-values (corrected for multiple tests) obtained when comparing FCNet with other decoders are reported. Bold p-values mark significant comparisons ($p < 0.05$).

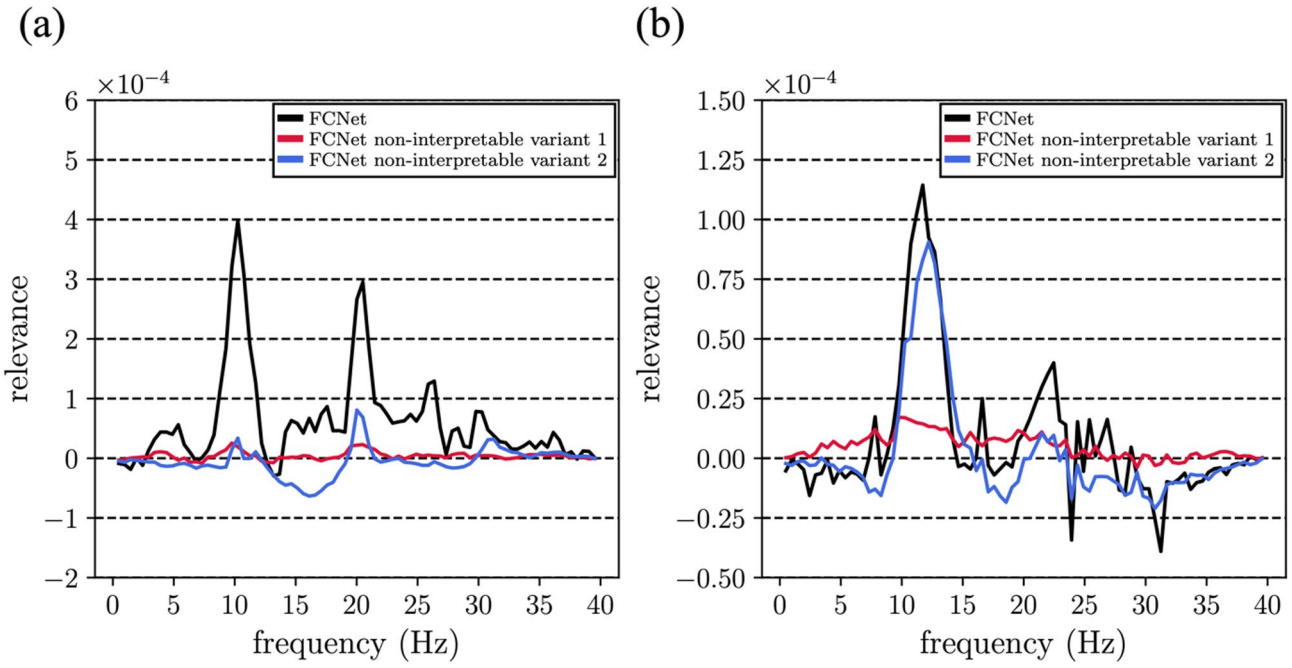

Fig. 12. Spectral relevance: FCNet vs. FCNet variants. The relevance attributed by FCNet (black lines) and its variants (red and blue lines) in the frequency domain is reported separately for the scalp-level experiment (**a**) and cortex-level experiment (**b**). Relevance patterns are reported as the mean value across participants.

Limitations and future directions

In this study, we presented how an analysis framework based on deep learning can be designed and used for analyzing oscillatory interactions in complex brain functional networks. Given the novelty of this methodology – used here for the first time with functional connectivity – this study meant to illustrate the novel analysis framework and to provide a first validation on real data recorded in a single cognitive task. Although preliminary, the obtained results are extremely promising and motivate to further enrich the validation of the proposed framework in the near feature by:

i. Performing simulations on synthetic data. By imposing a known connectivity pattern in simulated data, it would be possible to run a sensitivity analysis on processing choices, such as the choice for a specific reference value in the DeepLIFT explanation technique. Indeed, a different reference value may affect the

obtained explanations, and the default DeepLIFT reference value (zero) – widely used with neural time series[35–38] as well as in this study – could not be the most suitable one for analyzing EEG connectivity. Testing multiple references under a synthetic known ground truth, may help in the identification of a connectivity-oriented reference for DeepLIFT. Additionally, the use of simulated data with our interpretable and explainable framework may also help in assessing its trustworthiness. Indeed, by generating synthetic data with a known connectivity pattern and by testing our framework while progressively injecting more noise in the data, the performance and feature analyses would help identifying the threshold of FCNet accuracy ensuring that our framework reveals connectivity patterns matching the imposed ones. These analyses could also serve to test the beneficial effects of architectural changes, for example integrating attention mechanisms[74], in improving the predictive power of FCNet-based framework.

ii. Scaling up experiments on different cognitive tasks and recording modalities. Here, we used a single cognitive task (motor imagery) as exemplary task, characterized by a peculiar and well-characterized connectivity pattern. The application of our framework on other tasks characterized by a strong directionality (e.g., visuospatial attention tasks) would enhance the robustness of the proposed approach across different application scenarios, thus demonstrating a broader usefulness of our approach, strengthening the results provided in this preliminary analysis. Finally, the approach could also be tested on recording modalities different from EEG with higher spatial resolution, e.g., magnetoencephalography, to improve the quality of the source reconstruction in cortex-level experiments.

iii. Further increase the interpretability of the inflow/outflow computations in our interpretable CNN. Even though the CNN is realized to learn interpretable features related to connectivity inflow/outflow, the overall mathematical formulation that links the input connectivity to the learned inflow/outflow measures remains not clear. This represents a limitation compared to traditional graph theory-based measures. Future studies will further increase the interpretability in the CNN, for example introducing constraints in the feature learning operated by the network, to facilitate the revealing of the analytical formulation of the inflow/outflow measures.

Conclusions

In conclusion, here we proposed an interpretable convolutional neural network – termed FCNet – for decoding spectral functional connectivity. We exploited the network for assessing the predictability of brain functional networks for cognitive states, and the most informative frequency components and connectivity inflows/outflows related to the considered cognitive states. To this aim, we designed a data-driven analysis framework, by using the interpretable elements of FCNet for deriving data-driven inflow and outflow measures, and by using an explanation technique for revealing the most relevant frequency contents and connectivity inflows and outflows. Specifically, our approach relies on the computation of non-linear inflow and outflow measures derived by optimally combining the information of causal spectra across frequencies, and by optimally weighting the different brain regions in the computation of the inflow and outflow. We applied the proposed deep learning-enriched approach on real data recorded during a motor imagery task. Overall, our results suggest that the deep learning-enriched approach is able to reveal the most relevant frequencies and connectivity inflow/outflow of the oscillatory interactions, which are neurophysiologically plausible. Moreover, the inflow/outflow measures extracted with such approach are able to detect changes between brain states with high strength and significance, like measures derived from a more traditional approach. Neuroscientists could effectively exploit an analysis tool based on the knowledge embedded in FCNet to prospectively support data-driven investigations of brain oscillatory interactions during cognitive tasks.

Data availability

Public datasets have been used. Links: https://www.bbci.de/competition/iv/, https://gigadb.org/dataset/100542.

Received: 14 April 2025; Accepted: 25 August 2025
Published online: 03 October 2025

References

1. Varela, F., Lachaux, J. P., Rodriguez, E. & Martinerie, J. The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* **2**, 229–239. https://doi.org/10.1038/35067550 (2001).
2. Singer, W. Neuronal synchrony: A versatile code for the definition of relations? *Neuron* **24**, 49–65. https://doi.org/10.1016/S0896-6273(00)80821-1 (1999).
3. Siegel, M., Donner, T. H. & Engel, A. K. Spectral fingerprints of large-scale neuronal interactions. *Nat. Rev. Neurosci.* **13**, 121–134. https://doi.org/10.1038/nrn3137 (2012).
4. Womelsdorf, T. et al. Modulation of neuronal interactions through neuronal synchronization. *Science* **316**, 1609–1612. https://doi.org/10.1126/science.1139597 (2007).
5. Miraglia, F., Vecchio, F. & Rossini, P. M. Searching for signs of aging and dementia in EEG through network analysis. *Behav. Brain. Res.* **317**, 292–300. https://doi.org/10.1016/j.bbr.2016.09.057 (2017).
6. Magosso, E. & Borra, D. The strength of anticipated distractors shapes EEG alpha and theta oscillations in a working memory task. *NeuroImage* **300**, 120835. https://doi.org/10.1016/j.neuroimage.2024.120835 (2024).
7. Borra, D., Fantozzi, S., Bisi, M. C. & Magosso, E. Modulations of cortical power and connectivity in alpha and beta bands during the Preparation of reaching movements. *Sensors* **23**, 3530. https://doi.org/10.3390/s23073530 (2023).
8. Bastos, A. M. & Schoffelen, J. M. Review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* **9** https://doi.org/10.3389/fnsys.2015.00175 (2016).
9. Cao, J. et al. Brain functional and effective connectivity based on electroencephalography recordings: A review. *Hum. Brain. Mapp.* **43**, 860–879. https://doi.org/10.1002/hbm.25683 (2022).
10. Wang, H. E. et al. A systematic framework for functional connectivity measures. *Front. Neurosci.* **8** https://doi.org/10.3389/fnins.2014.00405 (2014).

11. Friston, K. J. Functional and effective connectivity: A review. *Brain Connect.* **1**, 13–36. https://doi.org/10.1089/brain.2011.0008 (2011).
12. Grech, R. et al. Review on solving the inverse problem in EEG source analysis. *J. Neuroeng. Rehabil.* **5**, 25. https://doi.org/10.1186/1743-0003-5-25 (2008).
13. Seth, A. K., Barrett, A. B. & Barnett, L. Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci.* **35**, 3293–3297. https://doi.org/10.1523/JNEUROSCI.4399-14.2015 (2015).
14. Granger, C. W. J. Investigating causal relations by econometric models and Cross-spectral methods. *Econometrica* **37**, 424. https://doi.org/10.2307/1912791 (1969).
15. Zeki, S. & Shipp, S. The functional logic of cortical connections. *Nature* **335**, 311–317. https://doi.org/10.1038/335311a0 (1988).
16. Stephan, K. E., Marshall, J. C., Penny, W. D., Friston, K. J. & Fink, G. R. Interhemispheric integration of visual processing during Task-Driven lateralization. *J. Neurosci.* **27**, 3512–3522. https://doi.org/10.1523/JNEUROSCI.4766-06.2007 (2007).
17. Frässle, S. et al. Mechanisms of hemispheric lateralization: asymmetric interhemispheric recruitment in the face perception network. *NeuroImage* **124**, 977–988. https://doi.org/10.1016/j.neuroimage.2015.09.055 (2016).
18. Hu, S. et al. Comparison analysis: Granger causality and new causality and their applications to motor imagery. *IEEE Trans. Neural Netw. Learn. Syst.* **27**, 1429–1444. https://doi.org/10.1109/TNNLS.2015.2441137 (2016).
19. Rathee, D., Cecotti, H. & Prasad, G. Estimation of Effective Fronto-Parietal connectivity during Motor Imagery using partial granger causality analysis. *Proceedings of the 2016 International Joint Conference on Neural Networks*. 2055–2062. https://doi.org/10.1109/IJCNN.2016.7727452 (2016).
20. Hamedi, M., Salleh, S. H. & Noor, A. M. Electroencephalographic motor imagery brain connectivity analysis for BCI: A review. *Neural Comput.* **28**, 999–1041. https://doi.org/10.1162/NECO_a_00838 (2016).
21. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198. https://doi.org/10.1038/nrn2575 (2009).
22. Sporns, O. Graph theory methods: applications in brain networks. *Dialog. Clin. Neurosci.* **20**, 111–121. https://doi.org/10.31887/DCNS.2018.20.2/osporns (2018).
23. Farahani, F. V., Karwowski, W. & Lighthall, N. R. Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Front. Neurosci.* **13**, 585. https://doi.org/10.3389/fnins.2019.00585 (2019).
24. Ursino, M. et al. Bottom-up vs. top-down connectivity imbalance in individuals with high-autistic traits: an electroencephalographic study. *Front. Syst. Neurosci.* **16**, 932128. https://doi.org/10.3389/fnsys.2022.932128 (2022).
25. Neuper, C., Wörtz, M. & Pfurtscheller, G. ERD/ERS patterns reflecting sensorimotor activation and deactivation. *Progress in Brain Research*. **159**, 211–222. https://doi.org/10.1016/S0079-6123(06)59014-4 (2006).
26. Pfurtscheller, G. & da Silva, F. H. L. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* **110**, 1842–1857. https://doi.org/10.1016/S1388-2457(99)00141-8 (1999).
27. Roy, Y. et al. Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* **16**, 051001. https://doi.org/10.1088/1741-2552/ab260c (2019).
28. Al-Saegh, A., Dawwd, S. A. & Abdul-Jabbar, J. M. Deep learning for motor imagery EEG-based classification: A review. *Biomed. Signal Process. Control.* **63**, 102172. https://doi.org/10.1016/j.bspc.2020.102172 (2021).
29. Ludwig, S. et al. EEGminer: discovering interpretable features of brain activity with learnable filters. *J. Neural Eng.* **21**, 036010. https://doi.org/10.1088/1741-2552/ad44d7 (2024).
30. Zhao, D., Tang, F., Si, B. & Feng, X. Learning joint space–time–frequency features for EEG decoding on small labeled data. *Neural Netw.* **114**, 67–77. https://doi.org/10.1016/j.neunet.2019.02.009 (2019).
31. Borra, D., Fantozzi, S. & Magosso, E. Interpretable and lightweight convolutional neural network for EEG decoding: application to movement execution and imagination. *Neural Netw.* **129**, 55–74. https://doi.org/10.1016/j.neunet.2020.05.032 (2020).
32. Borra, D., Bossi, F., Rivolta, D. & Magosso, E. Deep learning applied to EEG source-data reveals both ventral and dorsal visual stream involvement in holistic processing of social stimuli. *Sci. Rep.* **13**, 7365. https://doi.org/10.1038/s41598-023-34487-z (2023).
33. Borra, D. & Magosso, E. Deep learning-based EEG analysis: investigating P3 ERP components. *J. Integr. Neurosci.* **20**, 791–811. https://doi.org/10.31083/j.jin2004083 (2021).
34. Vahid, A., Mückschel, M., Stober, S., Stock, A. K. & Beste, C. Applying deep learning to single-trial EEG data provides evidence for complementary theories on action control. *Commun. Biol.* **3**, 112. https://doi.org/10.1038/s42003-020-0846-z (2020).
35. Sujatha Ravindran, A. & Contreras-Vidal, J. An empirical comparison of deep learning explainability approaches for EEG using simulated ground truth. *Sci. Rep.* **13**, 17709. https://doi.org/10.1038/s41598-023-43871-8 (2023).
36. Lawhern, V. J. et al. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **15**, 056013. https://doi.org/10.1088/1741-2552/aace8c (2018).
37. Gabeff, V. et al. Interpreting deep learning models for epileptic seizure detection on EEG signals. *Artif. Intell. Med.* **117**, 102084. https://doi.org/10.1016/j.artmed.2021.102084 (2021).
38. Raab, D., Theissler, A. & Spiliopoulou, M. XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series. *Neural Comput. Applic.* **35**, 10051–10068. https://doi.org/10.1007/s00521-022-07809-x (2023).
39. Wang, H. et al. Diverse feature blend based on Filter-Bank common Spatial pattern and brain functional connectivity for multiple motor imagery detection. *IEEE Access.* **8**, 155590–155601. https://doi.org/10.1109/ACCESS.2020.3018962 (2020).
40. Zheng, H. et al. Time-Frequency functional connectivity alterations in alzheimer's disease and frontotemporal dementia: an EEG analysis using machine learning. *Clin. Neurophysiol.* **170**, 110–119. https://doi.org/10.1016/j.clinph.2024.12.008 (2025).
41. Zhang, Y., Yan, G., Chang, W., Huang, W. & Yuan, Y. EEG-based multi-frequency band functional connectivity analysis and the application of spatio-temporal features in emotion recognition. *Biomed. Signal Process. Control.* **79**, 104157. https://doi.org/10.1016/j.bspc.2022.104157 (2023).
42. Zhang, R. et al. Using brain network features to increase the classification accuracy of MI-BCI inefficiency subject. *IEEE Access.* **7**, 74490–74499. https://doi.org/10.1109/ACCESS.2019.2917327 (2019).
43. Li, X., La, R., Wang, Y., Hu, B. & Zhang, X. A deep learning approach for mild depression recognition based on functional connectivity using electroencephalography. *Front. Neurosci.* **14** https://doi.org/10.3389/fnins.2020.00192 (2020).
44. Alves, C. L., Pineda, A. M., Roster, K., Thielemann, C. & Rodrigues, F. A. EEG functional connectivity and deep learning for automatic diagnosis of brain disorders: alzheimer's disease and schizophrenia. *J. Phys. Complex.* **3**, 025001. https://doi.org/10.1088/2632-072x/ac5f8d (2022).
45. Ellis, C. A., Miller, R. L. & Calhoun, V. D. Pairing explainable deep learning classification with clustering to uncover effects of schizophrenia upon whole brain functional network connectivity dynamics. *Neuroimage: Rep.* **3**, 100186. https://doi.org/10.1016/j.ynirp.2023.100186 (2023).
46. Shmueli, G. To explain or to predict? *Statist Sci.* **25** https://doi.org/10.1214/10-sts330 (2010).
47. Wagstaff, K., Cardie, C., Rogers, S. & Schrödl, S. Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*. 577–584. https://dl.acm.org/doi/10.5555/645530.655669 (2001).
48. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv.* https://doi.org/10.48550/ARXIV.1704.02685 (2017).
49. Abiri, R., Borhani, S., Sellers, E. W., Jiang, Y. & Zhao, X. A comprehensive review of EEG-based brain–computer interface paradigms. *J. Neural Eng.* **16**, 011001. https://doi.org/10.1088/1741-2552/aaf12e (2019).
50. Jayaram, V. & Barachant, A. MOABB: trustworthy algorithm benchmarking for BCIs. *J. Neural Eng.* **15**, 066011. https://doi.org/10.1088/1741-2552/aadea0 (2018).

51. Tangermann, M. et al. Review of the BCI competition IV. *Front. Neurosci.* **6** https://doi.org/10.3389/fnins.2012.00055 (2012).
52. Lee, M. H. et al. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience* **8**, giz002. https://doi.org/10.1093/gigascience/giz002 (2019).
53. Gramfort, A. MEG and EEG data analysis with MNE-Python, *Front. Neurosci.* **7** https://doi.org/10.3389/fnins.2013.00267. (2013).
54. Hallez, H. et al. Review on solving the forward problem in EEG source analysis. *J. Neuroeng. Rehabil.* **4**, 46. https://doi.org/10.1186/1743-0003-4-46 (2007).
55. Pascual-Marqui, R. D. Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: exact, zero error localization. *arXiv*. https://doi.org/10.48550/arXiv.0710.3341 (2007).
56. Desikan, R. S. et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021 (2006).
57. Hétu, S. et al. The neural network of motor imagery: an ALE meta-analysis. *Neurosci. Biobehavioral Reviews*. **37**, 930–949. https://doi.org/10.1016/j.neubiorev.2013.03.017 (2013).
58. Gallivan, J. P. & Culham, J. C. Neural coding within human brain areas involved in actions. *Curr. Opin. Neurobiol.* **33**, 141–149. https://doi.org/10.1016/j.conb.2015.03.012 (2015).
59. Srisrisawang, N. & Müller-Putz, G. R. Applying dimensionality reduction techniques in Source-Space electroencephalography via template and magnetic resonance Imaging-Derived head models to continuously Decode hand trajectories. *Front. Hum. Neurosci.* **16**, 830221. https://doi.org/10.3389/fnhum.2022.830221 (2022).
60. Ghumare, E. G., Schrooten, M., Vandenberghe, R. & Dupont, P. A Time-Varying connectivity analysis from distributed EEG sources: A simulation study. *Brain Topogr.* **31**, 721–737. https://doi.org/10.1007/s10548-018-0621-3 (2018).
61. Magosso, E., Ricci, G. & Ursino, M. Alpha and theta mechanisms operating in internal-external attention competition. *J. Integr. Neurosci.* **20**, 1. https://doi.org/10.31083/j.jin.2021.01.422 (2021).
62. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: A primer with examples, Hum. *Brain Mapp.* **15**, 1–25. https://doi.org/10.1002/hbm.1058 (2002).
63. Szucs, D. & Ioannidis, J. P. A. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15**, e2000797. https://doi.org/10.1371/journal.pbio.2000797 (2017).
64. Schirrmeister, R. T. et al. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain. Mapp.* **38**, 5391–5420. https://doi.org/10.1002/hbm.23730 (2017).
65. Borra, D., Magosso, E. & Ravanelli, M. A protocol for trustworthy EEG decoding with neural networks. *Neural Netw.* **182**, 106847. https://doi.org/10.1016/j.neunet.2024.106847 (2025).
66. An, J., Chen, X. & Wu, D. Algorithm contest of motor imagery BCI in the world robot contest 2022: A survey. *Brain Sci. Adv.* **9**, 166–181. https://doi.org/10.26599/BSA.2023.9050011 (2023).
67. Simões, M. et al. BCIAUT-P300: A Multi-Session and Multi-Subject benchmark dataset on autism for P300-Based Brain-Computer-Interfaces. *Front. Neurosci.* **14**, 568104. https://doi.org/10.3389/fnins.2020.568104 (2020).
68. Santamaría-Vázquez, E., Martínez-Cagigal, V., Vaquerizo-Villar, F. & Hornero, R. A novel deep convolutional neural network for assistive ERP-Based Brain-Computer interfaces. *IEEE Trans. Neural Syst. Rehabil Eng.* **28**, 2773–2782. https://doi.org/10.1109/TNSRE.2020.3048106 (2020).
69. Salami, A., Andreu-Perez, J. & Gillmeister, H. An explainable inception Temporal convolutional network for motor imagery classification. *IEEE Access.* **10**, 36672–36685. https://doi.org/10.1109/ACCESS.2022.3161489 (2022).
70. Pfurtscheller, G., Neuper, C., Ramoser, H. & Müller-Gerking, J. Visually guided motor imagery activates sensorimotor areas in humans. *Neurosci. Lett.* **269**, 153–156. https://doi.org/10.1016/S0304-3940(99)00452-8 (1999).
71. Pfurtscheller, G. & Neuper, C. Motor imagery activates primary sensorimotor area in humans. *Neurosci. Lett.* **239**, 65–68. https://doi.org/10.1016/S0304-3940(97)00889-6 (1997).
72. Zhang, T. et al. The Time-Varying network patterns in motor imagery revealed by adaptive directed transfer function analysis for fMRI. *IEEE Access.* **6**, 60339–60352. https://doi.org/10.1109/ACCESS.2018.2875492 (2018).
73. Ogawa, T., Shimobayashi, H., Hirayama, J. I. & Kawanabe, M. Asymmetric directed functional connectivity within the frontoparietal motor network during motor imagery and execution. *NeuroImage* **247**, 118794. https://doi.org/10.1016/j.neuroimage.2021.118794 (2022).
74. Vaswani, A. et al. Attention Is All You Need. *arXiv.* https://doi.org/10.48550/arXiv.1706.03762 (2017).

Acknowledgements

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.

Author contributions

D.B.: Conceptualization, Data curation, Methodology, Software, Formal analysis, Investigation, Project administration, Visualization, Writing - Original Draft, Writing - Review & Editing. E.M.: Validation, Methodology, Supervision, Resources, Writing - Review & Editing.

Funding

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006450001. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU – Project Title "Multisensory integration of locomotion-related visual and somatomotor signals – MulWALK", code PRIN2022-2022BK2NPS, CUP: J53D23010900006.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-17635-5.

Correspondence and requests for materials should be addressed to D.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.