Journal of Neural Engineering

PURPOSE-LED PUBLISHING™

PAPER • **OPEN ACCESS**

BIDSAlign: a library for automatic merging and preprocessing of multiple EEG repositories

To cite this article: Andrea Zanola *et al* 2024 *J. Neural Eng.* **21** 046050

View the article online for updates and enhancements.

You may also like

- Delayed administration of interleukin-4 coacervate alleviates the neurotoxic phenotype of astrocytes and promotes functional recovery after a contusion spinal cord injury
Manoj K Gottipati, Anthony R D'Amato, Jayant Saksena et al.
- Connectivity study on resting-state EEG between motor imagery BCI-literate and BCI-illiterate groups
Hanjin Park and Sung Chan Jun
- A preliminary study exploring the effects of transcutaneous spinal cord stimulation on spinal excitability and phantom limb pain in people with a transtibial amputation
Ashley N Dalrymple, Lee E Fisher and Douglas J Weber

This content was downloaded from IP address 153.109.130.52 on 20/03/2025 at 10:38

Journal of Neural Engineering

PAPER

BIDSAlign: a library for automatic merging and preprocessing of multiple EEG repositories

Andrea Zanola[1,2,*], **Federico Del Pup**[1,2,3], **Camillo Porcaro**[1,2,4,5,7] and **Manfredo Atzori**[1,2,6,7]

[1] Department of Neuroscience, University of Padua, Padua 35128, Italy
[2] Padua Neuroscience Center, Padua 35128, Italy
[3] Department of Information Engineering, University of Padua, Padua 35128, Italy
[4] Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, United Kingdom
[5] Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy
[6] Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), 3960 Sierre, Switzerland
[7] Both authors contributed equally to this work.
[*] Author to whom any correspondence should be addressed.

E-mail: andrea.zanola@studenti.unipd.it, federico.delpup@studenti.unipd.it, camillo.porcaro@unipd.it and manfredo.atzori@unipd.it

Keywords: EEG, preprocessing, BIDS, harmonization, public dataset, open source

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
9 May 2024

REVISED
17 July 2024

ACCEPTED FOR PUBLICATION
1 August 2024

PUBLISHED
20 August 2024

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Objective. This study aims to address the challenges associated with data-driven electroencephalography (EEG) data analysis by introducing a standardised library called *BIDSAlign*. This library efficiently processes and merges heterogeneous EEG datasets from different sources into a common standard template. The goal of this work is to create an environment that allows to preprocess public datasets in order to provide data for the effective training of deep learning (DL) architectures. *Approach.* The library can handle both Brain Imaging Data Structure (BIDS) and non-BIDS datasets, allowing the user to easily preprocess multiple public datasets. It unifies the EEG recordings acquired with different settings by defining a common pipeline and a specified channel template. An array of visualisation functions is provided inside the library, together with a user-friendly graphical user interface to assist non-expert users throughout the workflow. *Main results.* BIDSAlign enables the effective use of public EEG datasets, providing valuable medical insights, even for non-experts in the field. Results from applying the library to datasets from OpenNeuro demonstrate its ability to extract significant medical knowledge through an end-to-end workflow, facilitating group analysis, visual comparison and statistical testing. *Significance.* BIDSAlign solves the lack of large EEG datasets by aligning multiple datasets to a standard template. This unlocks the potential of public EEG data for training DL models. It paves the way to promising contributions based on DL to clinical and non-clinical EEG research, offering insights that can inform neurological disease diagnosis and treatment strategies.

1. Introduction

Electroencephalography (EEG) is one of the most widely used non-invasive diagnostic techniques in neuroscience, effectively employed to study a variety of diseases like epilepsy [1], Parkinson's [2] and Alzheimer's [3, 4]. However, EEG data have a low signal-to-noise ratio, since they are often corrupted by biological and non-biological artefacts that compromise the quality of the signal itself. Consequently, the proficient use of EEG recordings requires intense preprocessing, usually manually performed by an expert or trained medical personnel.

Over the years, extensive research has been focused on the problem of developing automated EEG preprocessing pipelines, necessary to overcome the aforementioned limitations. Among the others, the PREP pipeline by Bigdely-Shamlo *et al* 2015 [5] is one of the most used since it offers a standardised preprocessing and new methods for robust average re-referencing by iteratively detecting and interpolating bad channels. The Harvard Automated Processing

© 2024 The Author(s). Published by IOP Publishing Ltd

Pipeline for EEG (HAPPE), Gabard-Durnam *et al* 2018 [6] is a standardised pipeline that additionally offers artefact correction using independent component analysis (ICA), and automatic artefact rejection using the Multiple Artifact Rejection Algorithm (MARA) [7].

Things become even more complicated when the chosen pipeline employs ICA for artefacts rejection. Independently from the algorithm used for ICA, (e.g. fastICA [8], InfoMax based ICA [9]), and the specific parameters' values affecting the rejection, the automatic removal of EEG artefacts is still a challenging problem in which both algorithms and humans can have contradictory or even wrong classifications. Computational Testing for Automated Preprocessing toolbox [10] (Cowley and Korpela 2017) help the user with functions, allowing the comparison between different pipelines. Although efforts have been made to develop excellent and standardised pipelines progressively improved over the years, it seems that expert supervision is still necessary to check the quality results of those pipelines.

In this challenging context, Deep Learning (DL) has established itself as a powerful resource for the analysis of EEG data not only for its demonstrated superiority over conventional statistical techniques in discovering highly non-linear patterns embedded in the data [11], but also for its ability to efficiently work with minimally preprocessed data [12]. Nevertheless, DL architectures still require many, manually annotated training samples to achieve effective training. Despite the fact that very few populous EEG datasets exist, the availability of extensive datasets is limited due to costs, the availability of subjects for specific pathologies and laborious manual annotation processes that can only be performed by medical experts.

Notable exceptions to the above statement, for example, are the Healthy Brain Network (HBN) dataset [13], which stands out as one of the few publicly accessible large-scale EEG datasets, comprising the recording of 2951 children. Another exception is the Temple University Hospital (TUH) EEG Corpus [14], a collection of EEG recordings acquired since 2002 inside the TUH. Additionally, there is TDBRAIN [15] with 1274 psychiatric patients collected between 2001 and 2021. In contrast to such datasets, small EEG repositories are enormous [16]. Exploiting them could represent the turning point of DL-based EEG data analysis, which is why novel unsupervised solutions have been recently investigated.

Recent advancements in DL-based EEG analysis have shown how merging datasets acquired in different experimental settings can potentially improve model performance and enhance its generalizability [17]. However, while considerable efforts have been devoted on creating automatic pipelines and preprocessing multiple datasets simultaneously [18], more work needs to be done on their aggregation. Indeed none of the previously mentioned EEG pipelines address the challenging problem of integrate the heterogeneity of the EEG data.

The complexity of the EEG signal, compounded by its high variability, arises from several sources. These include the acquisition protocol, such as the channel system, the number and names of channels. Additionally, the storage modality, like dataset architecture and file extension, contributes to this variability. As a result, addressing the aggregation problem presents significant challenges.

Contributions: This work proposes *BIDSAlign* a new library designed to preprocess and efficiently merge multiple datasets. It can handle large amounts of data and standardise them to a common template. The library is addressed to both medical and deep-learning researchers and non-expert users. This work offers several functionalities useful for rapid and highly customisable preprocessing, taking the user from data loading, to result visualisation. Thanks to the complete and highly specific set of preprocessing and harmonisation functions, BIDSAlign not only allows to process data for DL studies based on proprietary data, but it also allows to integrate multiple data sources by harmonising them, laying the foundations for large studies based on both proprietary and public data.

paper structure: The rest of the paper is organised as follows. Section 2 provides a background on acquisition protocols and automatic preprocessing of EEG data. Section 3 illustrates the standardisation library's main functionalities and details its implementation. Results regarding the dataset selection, the library coverage and usability are presented in section 4 and discussed in section 5. Finally, the paper's conclusions are in section 6.

2. Background on EEG preprocessing

2.1. Data formats and channel systems

EEG data formats and channel systems are inherently heterogeneous due to the existence of multiple EEG machine manufacturers, each employing its own acquisition protocols and file formats for data storage. For instance, BioSemi® uses *.bdf* file extension, while BrainVision® uses the triple (*.vhdr, .vmrk, .eeg*) file extensions. EEGLAB [19], one of the most used MATLAB® toolboxes for EEG preprocessing, uses the couple *.set, .fdt* as preferred file extensions. These three, together with the European data format (*.edf*) [20], are the only ones allowed in the Brain Imaging Data Structure (BIDS) [21] storage system. BIDS, initially designed for storing MRI and fMRI data, has expanded [22] to include EEG, MEG (magnetoencephalography) and iEEG (intracranial EEG) providing a standardised format for data storage.

Furthermore, the diversity of EEG channel systems introduces variability in data collection, as different systems are utilized for various purposes and users may select differing numbers of channels from EEG caps. Since the pioneering work of Jasper in 1958 [23], international standards have been established for electrode placement on the scalp. The original 10–20 system introduced by Jasper *et al* featured 19 electrodes and two mastoid electrodes. Subsequent contributions by Nuwer [24], Chatrian *et al* [25], and Klem *et al* [26] led to the development of the 10–10 international system, also known as the extended 10–20 system, which has become widely adopted. However, even within the 10–10 system, the number of channels can vary, ranging from 63 to 78. In 2001, Oostenveld and Praamstra [27] introduced the 10–5 international system, significantly increasing the density of channels on the scalp and ushering in the era of High-Density EEG (HDEEG), allowing for the placement of up to 345 channels.

Channel systems, such as the Geodesic Sensor Net [28], deviate from the international standard, leading to differences in channel names and positions. Different electrode placements make the standardisation of the EEG data even more difficult since conversion tables are needed to move from one system to another. The correspondence between two channels of two different systems is made in terms of the distance between the electrodes, because positions do not coincide except for some particular ones. The following technical note explains an example of such a procedure applied to the Hydro-Cel Geodesic Sensor Net (HCGSN) channel system [29].

2.2. Automatic preprocessing of EEG data

Although great variability comes from how data is acquired and saved, software like EEGLAB is capable of successfully importing the majority of data types available in the literature. Aside from fundamental and common operations like resampling and filtering, ICA is another common operation performed on data. Many algorithms have been developed for automatic IC rejection, including some based on statistical thresholding [30, 31]. Algorithms like MARA [7] instead are based on supervised machine learning algorithms that perform the 'reject' vs. 'accept' binary classification of the ICs. ICLabel [32] instead is based on Convolutional Neural Networks and not only is it capable of classifying ICs as artefact vs not-artefact, but also recognising the type, classifying the artefacts between seven classes: brain, muscle, eye, heart, line noise, channel noise and 'other', meaning an undefined class. It is worth mentioning that in the current implementation, MARA needs EEG structures with channel labels as in 10–10 standard system; thus, other names make MARA fail.

Automatic bad channels and artefactual EEG segments identification is another important step in an EEG pipeline. The most straightforward solutions mark a trial or a time-segment as bad if the peak-to-peak amplitude in any sensor exceeds a certain fixed threshold, which is usually manually set after visual inspection. This, however, can be unfeasible for large datasets. Modern algorithms instead calculate more advanced trial statistics, such as FASTER [33], that reject trials based on a threshold on the z-score of the trial variance, its amplitude range and other metrics. However, these methods still need to be fully automated as rejection thresholds must be fixed or manually tuned. For single-trial analysis, an outlier detection algorithm is better suited. Random Sample Consensus [34] algorithm is implemented as part of the PREP pipeline, which annotates outlier sensors as bad. As mentioned in Pedroni *et al* [35], computationally less expensive and faster alternatives than the bad channel identification algorithms offered by PREP, are the three bad channel identification algorithms of the EEGLAB function *clean_rawdata*, which respectively search for flat, low-correlated and highly noise-corrupted channels. Artifact Subspace Reconstruction (ASR) [36] instead rejects artefacts based on their component variance. It accomplishes this by initially identifying segments of data considered 'clean' to establish a reference, followed by the application of principle component analysis to eliminate artifacts in relation to this reference. In particular, this method does not identify bad channels, but detects 'bad' time segments in the data and attempts to interpolate them based on the rest of the EEG signal. Both *clean_rawdata* with the respective three functions and ASR are used and implemented inside the Automagic pipeline [35]. It is worth to mention that *clean_rawdata* has been deprecated, and all the aforementioned functions inside it, together with *clean_asr*, are inside the EEGLAB function called *clean_artifacts*.

2.3. Scalable EEG pipelines

Building upon the foundation of EEG preprocessing methodologies, recent efforts have focused on enhancing scalability. In PREP, the user can import an EEG file in EEGLAB and run the preprocessing through the graphical user interface (GUI); otherwise the relative command line function accepts as input the EEGLAB's EEG structure. In HAPPE, the user specifies the folder and file format in which data are stored; then, the pipeline is applied for each of the present file. Delorme *et al* 2021 [18], created a bids-Matlab-tools plug-in for EEGLAB, capable of importing and exporting BIDS studies [21]. This tool is important both because it allows the user to specify the whole dataset and not the single files anymore, and on account of the alignment to the BIDS format.

Bids-matlab-tool is relevant because of the increasing importance of repositories like OpenNeuro [37], which facilitate data sharing between researches and standardise the storage according to the BIDS format. Automagic [35] internally uses the EEGLAB's function *pop_importbids* from bids-matlab-tool to read input datasets organised in BIDS format. Thus, it represents a good example of a preprocessing pipeline capable of scaling well to large amounts of data. As mentioned in their work, by default Automagic uses the PREP pipeline and it uses ICLabel for automatic IC rejection.

2.4. EEG data preprocessing for DL

Although the previous section shows the effort in the research communities to develop well-performing EEG pipelines, the literature presents conflicting views on whether preprocessing EEG data is beneficial or unnecessary for training DL architectures [38]. Based on the work of Roy *et al* [12], of over 154 studies reviewed, 111 performed preprocessing, 23 did not while other works did not provide such information. It was also found that only 36 studies rejected or corrected artifacts in their data. Such numbers highlight that, although DL could handle raw EEG data, a minimal amount of preprocessing is typically performed. This usually includes the following steps:

(i) **Baseline correction**: Various approaches exist for baseline correction. The traditional way is subtracting the mean of a baseline period from every time point of the baseline and post-stimulus interval. In other words, the average voltage values of each electrode are calculated within a time interval (usually the pre-stimulus interval). Then, this average is subtracted from that time interval of the signal. When it comes to non-epoched data, the average voltage values of each electrode are calculated on the whole recording.
(ii) **Resampling** to a lower sampling rate, most of the times between 125–128 Hz or 250–256 Hz, depending on the experimental setting;
(iii) **Filtering** with a band-pass filter to remove all signal components outside the investigated EEG bands; band-pass filtering is adopted to remove the DC component, typically with a low-frequency pass band of 0.1, 0.5 or 1 Hz;
(iv) **Re-referencing** to a common reference (average), to a group of channels (mastoids) or to a specific channel (e.g. Cz).

While the aforementioned steps are standard and widely adopted by the community, these steps does not correct the signal for bad time portions or bad channels, so further steps could include:

- **ICA** decomposition that allows for the rejection of artefacts affecting the recording;
- **Artifact Subpace Reconstruction** (ASR) [36] cleans heavily corrupted channels or EEG segments, allowing both interpolation and removal of the data;
- **Interpolation** of missing data, generally performed with interpolator from the PCHIP family (e.g. spp-chip, cubic spline, makima).

ICA decomposition and rejection, even if they can definitely increase the quality of the EEG signal, are typically not done in a big-data scenario, due to the computational time required for such operations. Other preprocessing steps include data re-scaling, signal partition with a specified overlap, channel selection, and transformation to another view (e.g. 2D spectra, EEG topography maps). However, such steps are not designed to improve the quality of the EEG signal but rather to adapt the input data to the design choices of the DL model. In conclusion, more efforts should be made by the DL research community to evaluate the effectiveness of the preprocessing of EEG data at different levels, on DL architectures.

3. Methods

3.1. Overview

BIDSAlign has been developed to offer a preprocessing tool for multiple EEG datasets focusing on the unification to a common template. The software can be fully run by command line or with a GUI and, therefore, does not require any programming knowledge. It comes with a blank example script of how to use the software, together with an interactive guided example to increase the library's usability; a detailed guide for the GUI is included as well.

BIDSAlign runs in MATLAB® and can be installed by simply downloading its source code from GitHub[8]. The library is built on top of EEGLAB, its main dependency, and requires the installation of other EEGLAB plugins to properly run some preprocessing steps, namely: *BIOSIG*, *FileIO* and *bva-io* for the data import; *FIRfilt* for the signal filtering; *FastICA*, *DIPFIT*, *ICLabel* and *MARA* for the IC analysis and rejection; *CleanRawData* for the ASR; *eegstats* for the visualization of individual alpha frequency (IAF) corrected results. *FIRfilt*, *ICLabel* and *CleanRawData* plugins are distributed along with the EEGLAB code, and imported by default. Plugins can be automatically downloaded and installed within the library, ensuring compatibility and facilitating use for non-experts. BIDSAlign was written and tested in MATLAB 2023b and EEGLAB 2023.0, ensuring compatibility with the latest versions of its main dependencies.

[8] https://github.com/MedMaxLab/BIDSAlign.

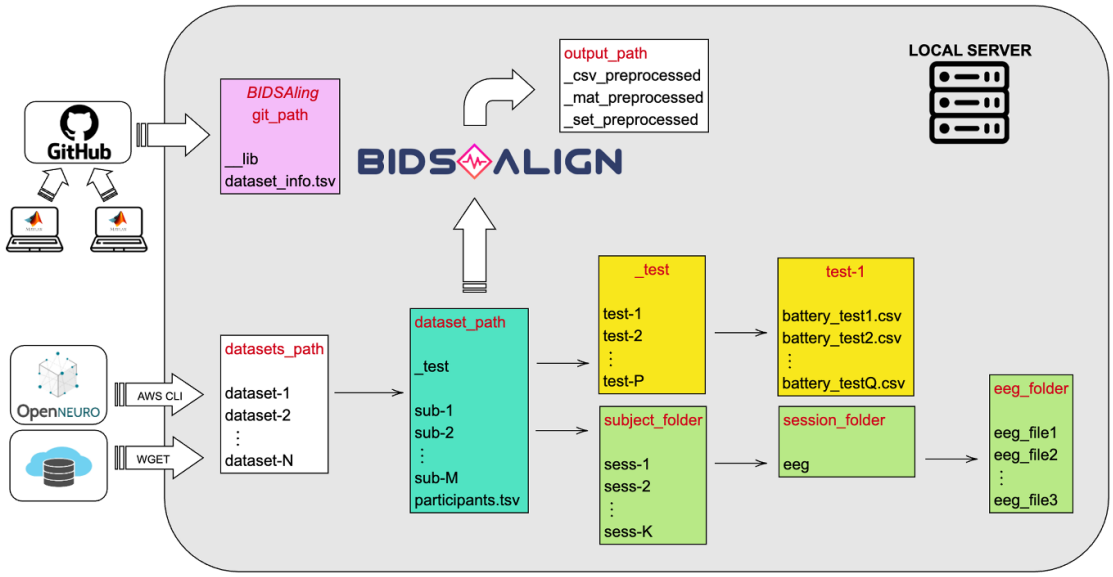

Figure 1. Folder structure visualized from the root path up to the file path. The employed workflow in the case of distributed computing resources is presented. On the top-left, multiple users can contribute to the development of the library via Git, and a clone of the repository is put inside the server. On the bottom-left, different modalities can be used to get a public dataset using either AWS-CLI command lines or the unix command wget. On the bottom-right, how a dataset is structured based on the BIDS format; 'sub' stays for subject, 'sess' for session. BIDSAlign takes the files to be preprocessed, performs the preprocessing and the output is stored inside 'output_path' (top-right).

BIDSAlign is completely open source. The community can contribute to its development by reporting bugs, suggesting new features, or implementing new functionalities. Furthermore, the code base is structured to be easy to interpret, allowing the experienced users to readily modify BIDSAlign to their specific needs.

3.2. Library structure

In facilitating EEG dataset preprocessing, the library adheres to a specific folder structure, aligning data with the BIDS format while also integrating demographic, medical, and diagnostic information in the processed file. The library is built over three nested loops reflecting the folder structure expected for an EEG dataset, saved in the following way:

dataset/subjects/sessions/eeg/objects.

The latest version of BIDS (v1.9.0) is considered in this work; for more information about please see the EEG modality specific section [39]. If the dataset is not saved according to the BIDS format, the user must change the dataset structure, with a provided function called *create_dataset_architecture*, before the launch of the preprocessing. This step is not only necessary for the correct functioning of the code, but also helps the alignment of the dataset to the BIDS format. It is worth mentioning that even datasets saved in BIDS, but with other templates should be changed as well. BIDSAlign can handle and consequently change the following dataset architectures:

- Dataset/objects,
- Subjects/objects,
- Subjects/eeg/objects,
- Subjects/sessions/objects,
- Subjects/sessions/folder/objects.

The overall workflow and folder organization that the user should follow is shown in figure 1. All the demographic and medical information should be stored inside the participant file placed inside the dataset's path. If other diagnostic tests, in form of CSV files are associated with the subjects, as happens in the MPI LEMON dataset [40], an auxiliary folder can be created, here called '_test'. Inside this folder, many others can be created, one for every inventory or battery of tests ('test-1', 'test-2', ...); for each of these inventory multiple CSV files ('battery_test1', 'battery_test2', ...) can be saved as well. The library merges together these information and those inside the participant file.

3.3. Work Modalities

The library provides versatile preprocessing options, including single file, single dataset and all dataset processing modes, while accommodating various selection criteria.

- **Single file**: The user specifies the filename and filepath of the file to be preprocessed and indicates from which dataset it comes from.
- **Single dataset**: The user specifies the dataset name to be preprocessed.

- **All datasets**: All datasets specified in a table file with name dataset_info will be preprocessed, in parallel or in a sequential manner, depending on the computing resources at disposal.

Considering the single dataset modality, the user has different ways to select a group of files respect the whole dataset.

In datasets containing distinct patient groups, users have the option to preprocess each group individually, making group-based selection the primary method for choosing subjects. Here the user can decide to select a sub-group of subjects, based on a specific feature and corresponding value, present in the participant file (e.g feature: 'Group', value: 'A' , Alzheimer).

The second option at disposal is to select subjects, sessions and objects by name. Label-based selection indeed can be performed, since BIDS expects folder names in the following format, 'sub-<label>', 'ses-<label>', 'task-<label>', so the user can specify these labels (e.g. 'task-eyesclosed'). In this way, only files or folders with that label name will be considered.

Finally, the last option is number based, where the user can decide to select a sub-group of subjects, sessions or objects based on the absolute index (e.g. from subject 5 to subject 10; e.g. only the first session). These three selection modalities, possibly combined together, allow the user to make an even finer-grained selection of the files without specifying the filenames. In those cases, the criteria are applied in the same order as described here. In conclusion, the library can be utilized in two ways: it can automatically preprocess an EEG recording and additionally align it to a specified template. While the first case is medical-oriented, the latter is for deep-learning researchers, targeting a large pool of possible users in this field.

3.4. Inputs

BIDSAlign expects information related to the dataset as a file named *dataset_info*. This file contains all the important information that is known *a priori* or that has to be specified by the user, specifically the dataset name, the folder name in which the dataset is stored, the reference used and the sampling rate. Indeed, BIDS expects a JSON file for every EEG recording containing such information. Thus the role of the dataset info file is central when the dataset is not stored in the BIDS format and the JSON file is not present. Additional user-provided information are the EEG file extension and the channel system; others that will be discussed more in detail in section 3.5, are:

- Channel location filename: If given, it can be a specific file that stores a channel location to use for every record, '*loaded*' if the EEG file already incorporates the channel coordinates, empty if no channel location was provided, or '*default*' if the standard one from EEGLAB should be used.
- Channel to remove: Here, you can select all the non-brain channels or other *a priori*-known bad channels; these channels will be removed and never inserted again.
- Nose direction: Useful if the nose direction of the coordinates system is different from the standard used by EEGLAB, which is '+X'.

3.5. Unification to a common template

Unification to a common template refers to the procedure of taking different EEG recordings and step-by-step integrating the intrinsic heterogeneity, producing as result, EEG data that are comparable. This unification is achieved at different levels during the entire course of the preprocessing and the following sub-sections aim to demonstrate that. In particular, 3.5.1–3.5.3 shows how channels and in general channel systems are aligned to the same common template discussed in 3.5.5. Re-referencing, i.e. the process of changing the reference channel of the EEG data, is another highly-variable factor in EEG data and will be addressed in sub-section 3.5.4.

3.5.1. Channels names

Channel names are saved in several ways. First of all, some datasets present old channels that have been renamed, such as T6 to T8, T5 to P8, T4 to P7 and T3 to T7. Other times, names are written with extra symbols or with different combinations of upper and lower case characters (e.g. FCZ, Fcz, fcz, FCz, FCz., FCz..); all these cases are harmonised to the uppercase (e.g. FCZ). In other cases, words are added to the channel names, for example, TP9 LEFT EAR to TP9, TP10 RIGHT EAR to TP10. Finally, in HydroCel Geodesic Sensor Net (HCGSN) channel systems, the electrode Cz is renamed E1129 or E2257 according to the system used.

3.5.2. Channels systems

The function *save_to_template* allows users to unify the data to a common template. Despite BIDSAlign provides a template that considers the IFCN 10-10 standard (see 3.5.5), the user can create its own template of desired channels written in any desired order. This function compares the channels present in the channel location of the file and the template, thus the following scenarios could arise:

- If the loaded EEG file has some missing channels with respect to the ones indicated in the template, interpolation is performed. This is done by passing to the function *pop_interp* a channel location with the missing channels and it automatically finds and interpolates them;

- If the loaded EEG file has more channels than the ones indicated in the template, the library takes only what required to fill the latter;
- If the loaded EEG file is recorded with the HCGSN129 or HCGSN257 channel system, the library uses a conversion file to determine which channels are needed.

The conversion file is a two-column matrix, where the first is the channel names according to the given template and the other is the corresponding channel names according to another channel system. This is useful for handling more channel systems at once. Only HCGSN has been considered in this work, so the conversion file was edited accordingly the technical note [29].

3.5.3. Channel original location

Differences in handling channel locations in EEG datasets, particularly between BIDS and non-BIDS formats, are here highlighted, along with the preprocessing steps contingent upon channel location information availability. In BIDS format, the *channels.tsv* file is recommended, but not strictly required; even if recommended, the only required fields are the channel names, type (e.g. EEG, EOG) and unit of measurement (e.g. mV, uV, nV). While not strictly mandatory, the *electrodes.tsv* file should include channel names and cartesian coordinates if provided. Additionally, discrepancies may arise in channel label names between the EEG channel location structure, *channels.tsv*, and *electrodes.tsv*. The hypothesis behind this is that the order in which the channels are stored in the EEG file, is the same as the order in which they appear in the *channels.tsv*, thus EEG channel labels are renamed accordingly, if this file is present.

The best case scenario is when the channel location is already loaded in the EEG file. In this case, the user can provide such information inside the 'channel location filename' field of the *dataset_info* file, by setting its value to 'loaded'. In this way, no external file will ever be imported. This, however, often does not happen, and only labels are loaded in the channel location structure. If the dataset does not come in BIDS format and there is only a single channel location for every file, the user can set the corresponding name in *dataset_info*; this channel location should be stored in the dataset path (see figure 1). If the field is left empty, the library will try to find an *electrodes.tsv* file at every level of the BIDS architecture. In cases where no channel location is loaded, if the channel.tsv is present, the library will automatically create a channel location, taking the coordinates from standard files taken from EEGLAB. Finally, if both the channel location and the channel file are not provided, the preprocessing will be stopped, returning an error; figure 2 shows a schematic summary.

It is worth stressing that while cartesian coordinates are mandatory in the *electrodes.tsv*, one of the most used interpolation method is the spherical one. Thus, a conversion from cartesian to spherical coordinates needs to be performed. The EEGLAB function *readloc* automatically tries to convert the coordinates, failing in some cases. A rotation of the axis is performed, if a different noise direction respects the standard '+X' is used; then there will be a centering of the coordinates. If coordinates have not been converted, preprocessing will halt to address this issue.

3.5.4. Data re-referencing

Re-referencing stands as a crucial stage in an automatic preprocessing pipeline, capable of eliminating noise that pervades all channels and altering the form and scalp distribution of ERP waveforms, thus impacting interpretation. The EEG reference used for the entire dataset can be declared inside the dataset info file and will be assigned to the corresponding field of the EEGLAB's data structure, 'EEG.ref'. If the reference field is left empty in *dataset_info* and a JSON file is present, the reference declared inside the latter will be used instead. However, if both scenarios fail, the reference within the EEG structure will be retained; this procedure is due to implementation nuances in the EEGLAB import functions such as *pop_loadbv*, *pop_loadset*, and *pop_biosig*. Once the data reference is recognized, the possible cases that would arise when re-referencing are summarized in table 1.

3.5.5. Template

If the user wants to store the EEG data in a *.mat* file, the library allows the possibility to save the EEG data in a 2D or 3D-object. While the 2D matrix format (channels $\times$ time) is the usual way to store EEG data, the 3D tensor format (channels $\times$ channels $\times$ time) rearranges the EEG channels in a way resembling how they are placed on the head. DL architectures like 3D Convolutional Neural Networks can exploit the local arrangement of the electrodes in order to perform the assigned task correctly. The channels taken into account in the template and the corresponding arrangement in the tensor template are shown in figure 3. Considering the selected electrodes, the channels are organised in a 9×9 table, with the appropriate zero padding. The proposed disposition reflects the electrodes placements on the skull, where the only arbitrary positions are those of O1-PO3, O2-PO4 and symmetrically FP1-AF3, FP2-AF4.

3.6. Preprocessing

The preprocessing of the EEG data is at the core of BIDSAlign, thus specific details are outlined in subsequent sections, ensuring comprehensive understanding of the several steps and parameter values. All the inputs provided in *dataset_info* and parameters of the preprocessing are imported into the function *preprocess_dataset*, whose objective is to prepare

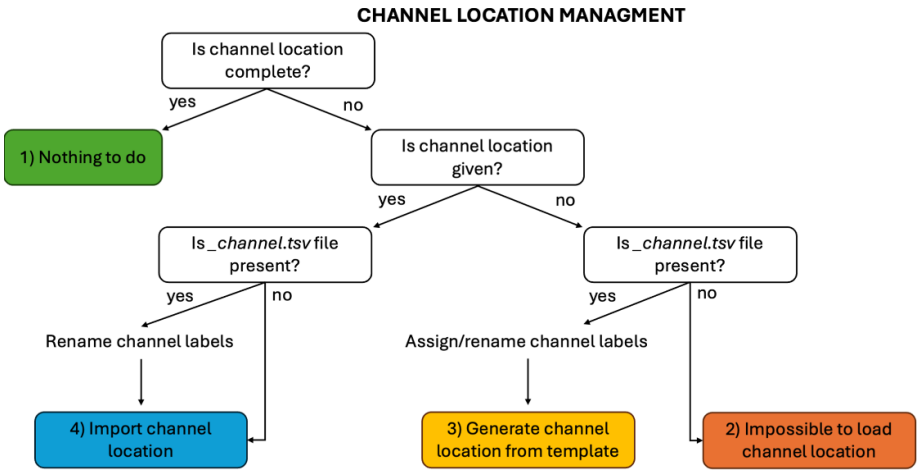

Figure 2. Management of the possible cases, when importing a channel location. If the EEG structure imported already has the electrodes coordinates then there is nothing to do (case 1). If the channel location is not given and there is not the names of the electrodes is impossible to load a channel location (case 2). In cases where at least the names of the channels are present then standard coordinates from the template are used (case 3). If instead the channel location is given, then BIDSAlign imports the channel location, eventually by renaming the channels in the EEG structure (case 4).

Table 1. A table showing all the possible re-reference combinations. Different scenarios arise if the electrode of the old reference is present inside the data or not, and if the electrode of the new desired reference was recorded too. The old reference is the current dataset reference, while the new reference is the new desired one. Channel S is the old channel reference, channel T is the new desired channel reference, and 'common' refers to the common average reference.

				New reference	
				Channel T	Common
Old reference	Common		Channel T is recorded	Re-reference to channel T	Nothing to do
			Channel T is not recorded	Error: channel T not present	
	Channel T		Channel T is recorded	Nothing to do	Common re-reference plus (if required) interpolate channel T and/or channel S
			Channel T is not recorded	Interpolating channel T will give a vector of zeros	
	Channel S	Channel S is recorded	Channel T is recorded	Re-reference to channel T	
			Channel T is not recorded	Error: channel T not present	
		Channel S is not recorded	Channel T is recorded	Re-reference to channel T plus (if required) interpolate channel T and/or channel S	
			Channel T is not recorded	Error: channel T not present	

everything needed for single EEG file processing. The preprocess is done in three steps.

(i) *Extract_filenames*: Extracts if present, the correct *channels.tsv*, *electrodes.tsv*, *events.tsv* and *eeg.json* files, based on the selected object filename.
(ii) *Preprocess_single_file*: Preprocess the data using a combination of custom and defined EEGLAB functions.
(iii) *Save_data_totemplate*: If required, this function uniforms all the EEG data according to the given template.

The preprocessing steps implemented for the second function are described in the following subsections. The EEG history is updated for each preprocessing step in order to store a logfile with all the performed operations and the sets of parameters adopted during the preprocessing; EEG history already

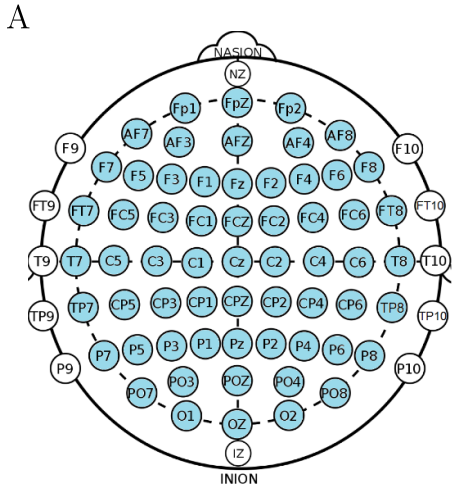

Figure 3. (Panel (A) The template used in this study correspond to the International Federation of Clinical Neurophysiology (IFCN) 10–10 standard containing 61 channels. (Panel (B) Arrangement of the 61 selected channels in the tensor template, which is the IFCN template put in a squared matrix, 9×9. Zero padding was put in positions with no electrodes. (Panel (C) show the arrangement for some of the 61 selected channels in the matrix template, which is a vector of length 61.

included in the original/raw file is maintained. All preprocessing steps mentioned in the following paragraphs except 3.6.1, can be disabled, giving free-choice to the user to design his own pipeline. While preprocessing steps can be disabled, they are always performed in the described order. Future developments of the library will allow the customization of the preprocessing steps order. Small pipelines with few preprocessing steps are more suitable for real-time scenario; an analysis of the computational time required by BIDSAlign, for progressively more richer pipelines are shown in the supplementary material.

3.6.1. Data import

BIDSAlign, with the function *import_data*, can load most of the common EEG file types, since it relies on EEGLAB functions exclusively. However, the library was tested on the BIDS format (requiring the use of one between, *.vhdr, .set, .bdf* and *.edf*). Thus, if the EEG format differs from the BIDS format extensions (e.g. Neuroscan CNT *.cnt*, EGI MFF *.mff* and eXimia EEG *.nxe*) the preprocessing will be stopped. Even though the four formats listed above are the ones supported by BIDSAlign, other EEG formats can be imported with dedicated EEGLAB plugins. Since the library is open-source, the authors encourage other researchers to contribute to BIDSAlign by expanding the formats supported by BIDSAlign.

3.6.2. Channels and time sections removal

If NaNs are present in the data for any reason, time sections containing at least one NaN in any channel are removed. Afterwards, the function *pop_select* is used to remove both specific channels and time portions. Regarding the channels, the user can specify in the dataset info which channel to remove, and it is useful if there is *a priori* knowledge of which of them is considered bad. Channels here specified are removed and never inserted during the preprocessing in the *.set* file; if these channels are required by the template, they will be interpolated and kept for the *.mat* file instead. Then, the user can remove the initial and final sections of the signal since they are often contaminated by artefacts, for example due to the charge or discharge of the amplifier.

3.6.3. Baseline removal

With the function *pop_rmbase* the user can remove the DC-offset in the data, by subtracting the channel-wise mean from each data point.

3.6.4. Resampling and filtering

Resampling and filtering operations, are done with the functions *pop_resample* and *pop_eegfiltnew*. The original sampling frequency is always checked because if it is not an integer value, the sampling frequency declared in dataset info or read in the JSON file is used instead.

3.6.5. ICA and rejection

EEGLAB function *pop_runica* allows to perform ICA decomposition with different methods; those supported by BIDSAlign are 'runica', using the InfoMax method and 'fastica'. ICA can be done both in the pipeline for automatic artifact rejection and at the end to save the IC decomposition of the signal in the final EEG structure. If ICs are used for artifact handling, MARA and ICLabel are both supported as the two main methods for automatic artefacts classification. Once identified, BIDSAlign allows to either reject or correct artifact components. If the user choose to correct them, the wavelet enhanced Independent Component (wICA) algorithm for artifact suppression is used [41]. This method applies a wavelet thresholding to the demixed independent

components affected by characteristic artefacts, aiming at recovering the neural activity and preserving a greater amount of data.

3.6.6. Notch filtering
Notch-Filtering operations are done with the function *pop_eegfiltnew*. The notch-frequency can be taken from dataset info, read from the JSON file or set by the user; the bandwidth of the transition can also be set. Since ICA is able to isolate line-noise components, and ICLabel can specifically detect this type of noise, the Notch Filtering step was placed after ICA.

3.6.7. Automatic channel correction and removal
The function *clean_artifacts* is used for automatic bad-channel and bad-time windows identification. Internally it calls functions such as *clean_flatlines*, *clean_drifts*, *clean_channels*, *clean_asr* and *clean_windows*. The first one ensures that data contains no flat-lined channels; the second consists of the application of a high pass filter, but it is not used since filtering is pre-applied. *clean_channels* instead ensures that the data contains no channels that record only noise for extended periods of time. These three functions are used in the first step of cleaning the data and they can be applied with the function *prepstep_1ASR*. In this way, bad channels for the corresponding criteria are removed, without changing the recording length.

A second cleaning step of the data is achieved with the function *prepstep_2ASR*, which mainly uses *clean_asr* function. This second step is activated only when the previous step returns and EEG data, with an absolute maximum value in μV, above a certain threshold defined by the user. The function ensures that data contains no events that have abnormally strong power; by setting a value for 'BurstCriterion', *clean_asr* will be used, thus bursts are repaired by ASR; on the contrary, the final EEG data will not have such repairs and the bad time portions of the signal are removed.

3.6.8. Interpolation, re-referencing and saving
As the final steps, the bad-channels removed by ASR are interpolated using the function *pop_interp*; then, re-reference is applied as described before (see table 1) with the function *pop_reref*. Finally, if the preprocessed *.set* file is required by the user, the *pop_saveset* function will be used to store it inside the folder '_set_preprocessed' (see figure 1), within a sub-folder named with the following convention:

$$\{\text{dataset}\}\{\text{group}\}_\{\text{pipeline}\}.$$

For example, dataset 'ds004504' [42], group Alzheimer 'A', pipeline applied only filtering 'FILT', thus the folder name will be 'ds004504A_FILT'. This convention is used for results visualisation as explained in section 3.8.

3.7. Output of the library
The library, if required by the user, outputs three files all named as *W_J_K_I*: a *.mat*, a *.csv* and a *.set* file; this name composed of four numbers, reflects the BIDS structure, in particular:

$$\text{dataset}_{(W)} > \text{subjects}_{(J)} > \text{sessions}_{(K)} > \text{eeg} > \text{objects}_{(I)}.$$

The numbers in the filename correspond to the positions of the files and folders, ordered alphabetically. If the subjects folders are named as 'sub-00x' and the subjects in the participant file are correctly ordered, the subject number '00x' and the subject ID (W) coincide. The *.mat* file serves as a container for the preprocessed EEG data. It can either store a N-dimensional numeric array or a structured variable, which includes the following fields:

- **Data**: The preprocessed EEG data either in matrix (2D) or in tensor (3D) format, saved according to the given template;
- **File path**: The file path of the preprocessed EEG recording;
- **Pad file**: A binary list, that indicates with the value one if a channel was interpolated, otherwise zero;
- **Template**: The provided template;
- **Subj info**: A structure containing subject information copied from the participant file attached to every dataset. If other diagnostical test are saved, the library import also these information. The library always performs a double check between the name of the folder and the subject ID, saved in the participants' file;
- **Group label**: A short indication from which group the subject comes from (e.g. group Alzheimer thus group label 'A');
- **Pipeline label**: A short indication of which pipeline was applied (e.g. for a pipeline with only filtering, the pipeline label could be 'FILT');

The *W_J_K_I.mat* file can be imported into deep-learning frameworks such as PyTorch [43], for training DL architectures. The *W_J_K_I.csv* file instead is a converted version of the field 'EEG.event' in the EEGLAB's EEG structure; the conversion was done using the EEGLAB function *eeg_eventtable*. This file could be useful if the user would like to divide the signal into windows matching a precise length, due to the experimental design. Finally, the *W_J_K_I.set* file is the preprocessed EEGLAB's EEG structure.

3.8. Visualisation
Building upon the preprocessing capabilities discussed in the previous sub-sections, the library offers diverse visualization options to aid in the analysis and interpretation of EEG data. By naming the folders of the set files with the convention explained in 3.6.8, data are stored and ready to be visualised.

The library offers three types of visualisation: group based, pipeline based and single file based.

In group-based analysis, the user can compare multiple group of subjects (e.g. healthy/pathological) or the same subjects performing different tasks (e.g. eyes closed/open) (see section 3.3). Pipeline-based analysis is instead used when different pipelines are applied to the same group and the user wants to compare differences in the data. In the last case, the user can visualise the properties of one specific EEG file.

Three main plots are returned by the function *group_visualization*: a plot of the group-averaged power spectral density (PSD) mediated over a set of channels, a boxplot and a topoplot showing respectively the average power and the group-averaged scalp topography, calculated for each of the five EEG bands.

Regarding the first plot, the PSD is calculated following the common practice of using the function *pwelch*, 50% overlap, hanning window, NFFT = 512; data can be normalised by z-scoring the channels. The second figure comprises five boxplots, covering the following EEG bands (see [44]): $\delta : [0.1, 4]$ Hz, $\theta : [4, 8]$ Hz, $\alpha : [8, 13]$ Hz, $\beta : [13, 30]$ Hz, $\gamma : [30, 45]$ Hz. On the boxplots, the central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers. Since data are often not-normally distributed, violin plots could be more useful and they are put on the right side of the boxplots. Violin plots have been created using the normal kernel smoothing function estimate for univariate data. If the IAF [45] is required, an additional boxplot is plotted and the group alpha frequencies are compared; the IAF is calculated with the function restingIAF [46, 47], which is inside the EEGLAB plug-in eegstats. If the IAF is calculated, frequency bands are consequently corrected as done by Saggar *et al* in [44].

Finally, average topographies are calculated for each frequency band; if only two groups or two pipelines are compared, a non-parametric permutation t-test is automatically done on the topographies, in order to find significantly different brain areas. This was implemented with the function *bramila_ttest2_np*, taken from [48, 49]. False Discovery Rate, step-up Benjamini–Hochberg method [50] was used to control for individual testing of each group and frequency band, giving corrected p-values for multiple comparisons (20 000 permutations). This procedure uses permutations of group labels to independently estimate the null distribution of each EEG channel, without assuming it to be identical across channels. However, the underlying and verified assumption is that EEG channels correspond to the same locations across all subjects. This procedure was also employed to generate reliable p-values for comparing the different groups across each band depicted in the boxplots.

On top of that, the library comes with other visualization tools. For example, BIDSAlign supports Evoked Response Potential (ERP) visualisation, allowing the user to evaluate both average-over-trials single-subject ERP and grand average ERP for a whole group or condition. Moreover, different channels and different pipelines can be inspected simultaneously; finally, scalp topographies can also be evaluated in time, for a series of defined time steps.

In conclusion, template comparison helps the user to understand and visualise the effects of the conversion between two channel systems. This is done not only by looking at the channel spatial correspondence between the channel systems, but also by inspecting how scalp topographies change by taking a sub-sample of channels.

4. Results

This section's main goal is to demonstrate BIDSAlign's potential, showing how the library can import and preprocess data from heterogeneous sources and produce meaningful medical results that are easily and understandably visualised.

4.1. Selected datasets

In order to demonstrate its usefulness and usability, BIDSAlign was tested on six publicly available datasets, selected according to the following criteria, common EEG analytics: first, datasets must be open access and directly available; second, they must include acquisitions from more than 20 subjects with recordings having at least 19 channels, with data sampled at least at 250 Hz; third, datasets must correspond to different channel systems, focusing on exemplary diseases such as Parkinson's and Alzheimer's. Considering point two, the choice of 19 channels, was performed following the guidelines of the American Clinical Neurophysiology Society [51], which state that the minimum number of electrodes for performing clinical electroencephalography is 21 (19 and A1–A2 used for contra-lateral referencing of all EEG electrodes). Secondly the choice of having a sampling frequency of 250 Hz is due to sampling considerations. Considering the lower-gamma band range ($\gamma_L : [30, 45]$ Hz) and taking into consideration the conservative rule of multiplying by four the maximum frequency to consider, one obtains a minimum sampling rate of 180 Hz. As the majority of datasets considered possess a native sampling rate that is a multiple of 250 Hz and exceeds the previously calculated value, 250 Hz was selected as the preferred value. The datasets selected by the criteria mentioned above, are reported in table 2; they were chosen from the OpenNeuro [37] platform, which collects and shares directly downloadable open datasets. In the following sub-sections, a brief description of the selected datasets is given.

4.1.1. ds004148

This dataset named 'A test-retest resting and cognitive state EEG dataset' [52], contains resting (eyes

Table 2. Free available datasets found and selected for results visualization. Datasets selected cover all the channel systems handled by BIDSAlign, which are the 10–20, 10–10, 10–5, HCGSN-129 and HCGSN-257, with variable number of channels. 'Reference' and 'Sampling Rate' describes two important properties of the acquisition protocol. 'Number of Subjects', 'Groups' and 'Tasks Recorded', show the great variability of the recordings, which comprise rest and tasks, for both healthy and pathological conditions. 'Steps Already Done', shows which preprocessing operations have been done by the authors on the dataset. Acronyms: HCGSN stays for Hydro-Cel Geodesic Sensor Net; EO stays for Eyes Open, EC stays for Eyes Closed; MSIT stays for Multi-Source Interference Task, SRT stays for Simple Reaction Time; CMS stays for Common Mode Sense active electrode, while DRL stays for Driven Right Leg passive electrode.

Dataset	Channel system	Number of channels	Reference	Sampling rate [Hz]	Number of subjects	Groups	Tasks recorded	Steps already done
Test Retest Rest [52]	10-10	64	FCz	500	60	Healthy	Rest EO/EC, memory, music, subtraction	raw
EEG-Alzheimer [42]	10-20	19	A1-A2	500	88	Healthy Alzheimer	Rest EC	filter[0.5,60] Hz
UC S.Diego [53, 54]	10-10	41	CMS-DRL	512	31	Healthy Parkinson	Rest EO	raw
HBN EO EC [13]	HCGSN	129	Cz	500	2951	Healthy	Rest EO/EC	raw
HDEEG 2 [55]	HCGSN	257	Cz	1000	23	Healthy	Rest EO, visual, auditory naming and memory	filter[3, 45] Hz
Nencki Symfonia [56]	10-5	127	FCz	1000	42	Healthy	Rest EO, MSIT, 3-stimuli visual oddbal, SRT	filter< 280 Hz

closed, eyes open) and cognitive state (subtraction, music, memory) EEG recordings for 60 young healthy subjects (M = 20.0, SD = 1.9 years). The results (figure 4) show the differences between the two conditions, for the same 60 subjects, where only the first session was considered.

4.1.2. ds004504

This dataset named 'A dataset of EEG recordings from: Alzheimer's disease, Frontotemporal dementia and Healthy subjects' [42], contains EEG resting state eyes closed recordings. The results (figure 5) show the differences between the Alzheimer (A, 36 subjects) and Control (C, 29 subjects) groups. These two groups are age-matched, where for the Alzheimer group the mean age is M = 66.4, SD = 7.9 years, while for the control group is M = 67.9, SD = 5.4 years.

4.1.3. ds002778

This dataset named 'UC San Diego Resting State EEG Data from Patients with Parkinson's Disease' [53], contains rest eyes open EEG recordings. The results (figure 6) show the differences between the Parkinson (P, 15 subjects) and Control (C, 16 subjects) groups. For Parkinson's patients, only session 'OFF' was considered ('POFF'), in which patients discontinued medication use at least 12 hours before the session. The results (figure 6) show the differences between the two groups and how IAF correction and normalisation affects the scalp topographies. These two groups are age-matched, where for the Parkinson group the mean age is M = 63.3, SD = 8.2 years, while for the control group is M = 63.5, SD = 9.7 years.

4.1.4. ds004621

This dataset named 'The Nencki-Symfonia EEG/ERP dataset' [56], contains EEG recordings for a variety of tasks: (i) an extended multi-source interference task, MSIT+; (ii) a 3-stimuli oddball task; (iii) a control, simple reaction task, SRT; and (iv) a resting-state eyes-open protocol. The results (figure 7) for the 42 subjects, show the comparisons between two pipelines during the 3-stimuli oddball task.

4.1.5. ds003421

This dataset, named 'HD-EEGtask(Dataset 2)' [55], contains EEG recordings during four conditions, (i) resting eyes open, (ii) visual and (iii) auditory naming and (iv) memory. Results are reported for the first subject ('sub-01') and for the resting-state condition. The results (figure 8), shows the comparison between the channel location of different channel systems and how they affects the scalp topographies.

4.1.6. ds004186

This dataset, named 'HBN EO/EC task' [13], contains recordings of children's eyes-open and eyes-closed EEG. Eyes-open lasted for 20 s, while eyes closed for 40 s. Results are reported for the first subject

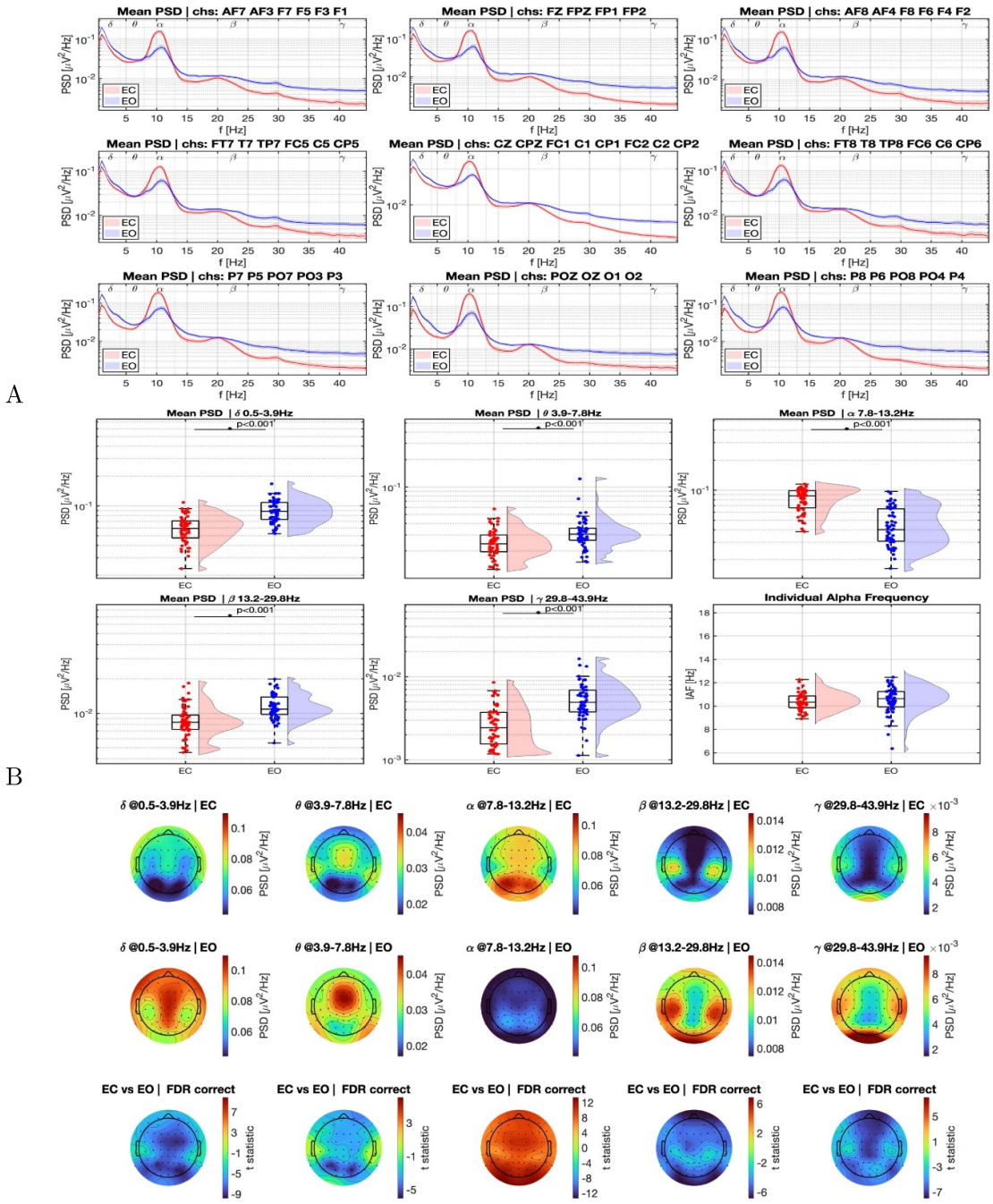

Figure 4. ds004148 results. Panel (A) shows the mean power spectral density (PSD) for Eyes Closed (EC) and Eyes Open (EO) groups, with lighter bands showing one standard deviation across subjects. Panel (B) shows the mean PSD for each band and the Individual Alpha Frequency (IAF) for both groups. Panel (C) shows mean group topographies in the first two rows and False Discovery Rate (FDR)-corrected t-statistics for group comparison in the last row, with color bars for PSD and t-statistics. The Step-up Benjamini–Hochberg method was used for FDR correction.

('sub-NDARAA075AMK') and for the first recording ('run-1'). Results figure 9 shows how different pipelines affects the EEG recording.

4.2. Library validation: group analysis

This section demonstrates the effects of preprocessing on the data and shows how BIDSAlign performs group-analysis retrieving meaningful results. The datasets chosen analysed in this section are ds004148 [52], ds004504 [42], ds002778 [53] and ds004621 [57]. The processing pipeline considered follows the steps described in section 3.6. The data are resampled to 250 Hz, filtered between 1–45 Hz, the first 5 s and last 5 s are always removed, in order to avoid potential initial and final divergences in the recording due to technical issues like switching on-off the EEG amplifier. IC decomposition is done with fastICA, calculating no more than 30 components, with hyperbolic tangent as the non-linearity method. Automatic IC rejection was done with ICLabel and both first

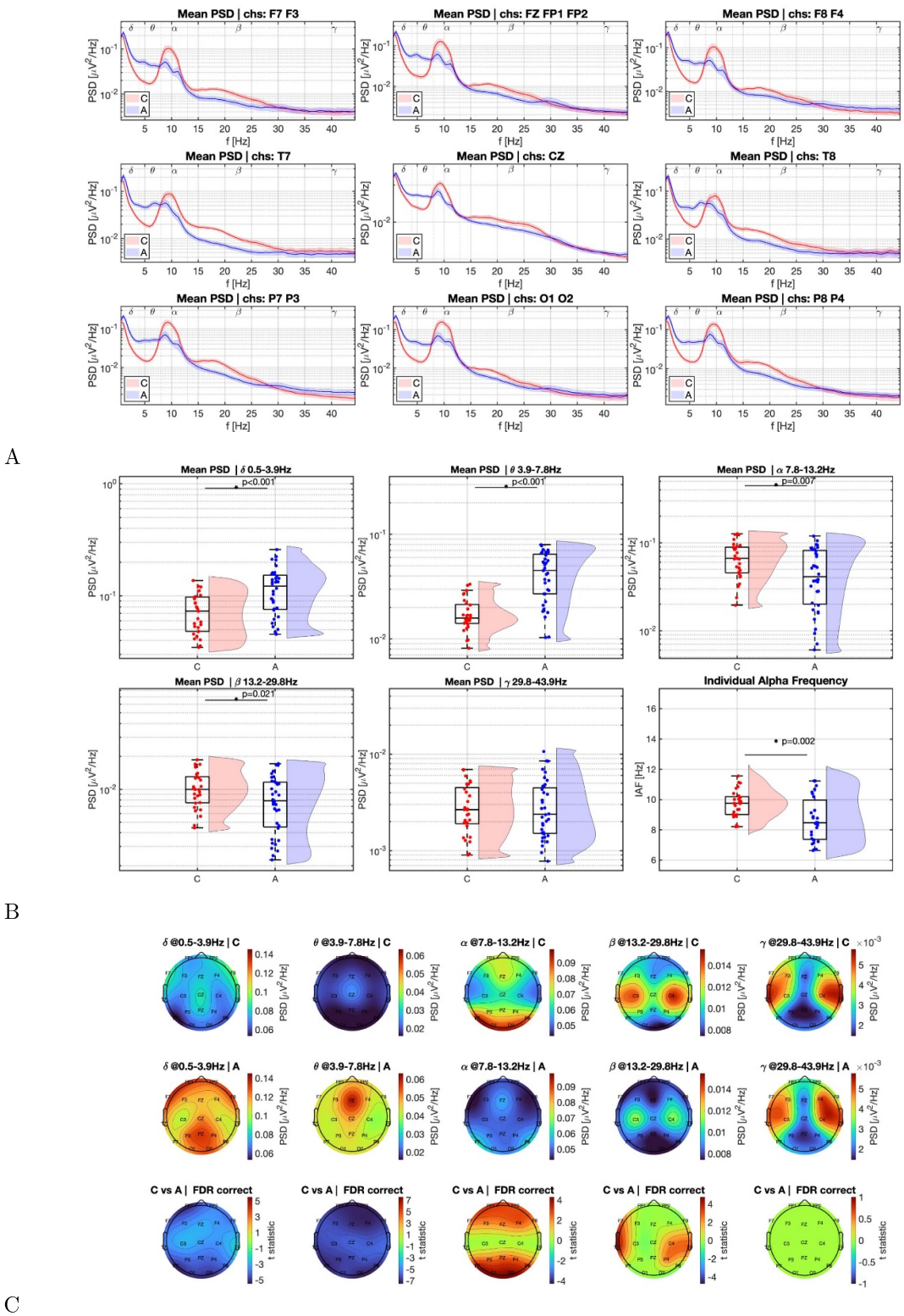

Figure 5. ds004504 results. Panel (A) shows the mean power spectral density (PSD) for Control (C) and Alzheimer (A) groups, with lighter bands showing one standard deviation across subjects. Panel (B) shows the mean PSD for each band and the Individual Alpha Frequency (IAF) for both groups. Panel (C) shows mean group topographies in the first two rows and False Discovery Rate (FDR)-corrected t-statistics for group comparison in the last row, with color bars for PSD and t-statistics. The Step-up Benjamini–Hochberg method was used for FDR correction.

and second-step ASR were performed, with default parameters and burst removal; the second ASR in which artefacts are removed is activated whenever one of the channels has an absolute voltage value above the voltage threshold of 250 μV. Spherical interpolation was applied to the missing channels and the

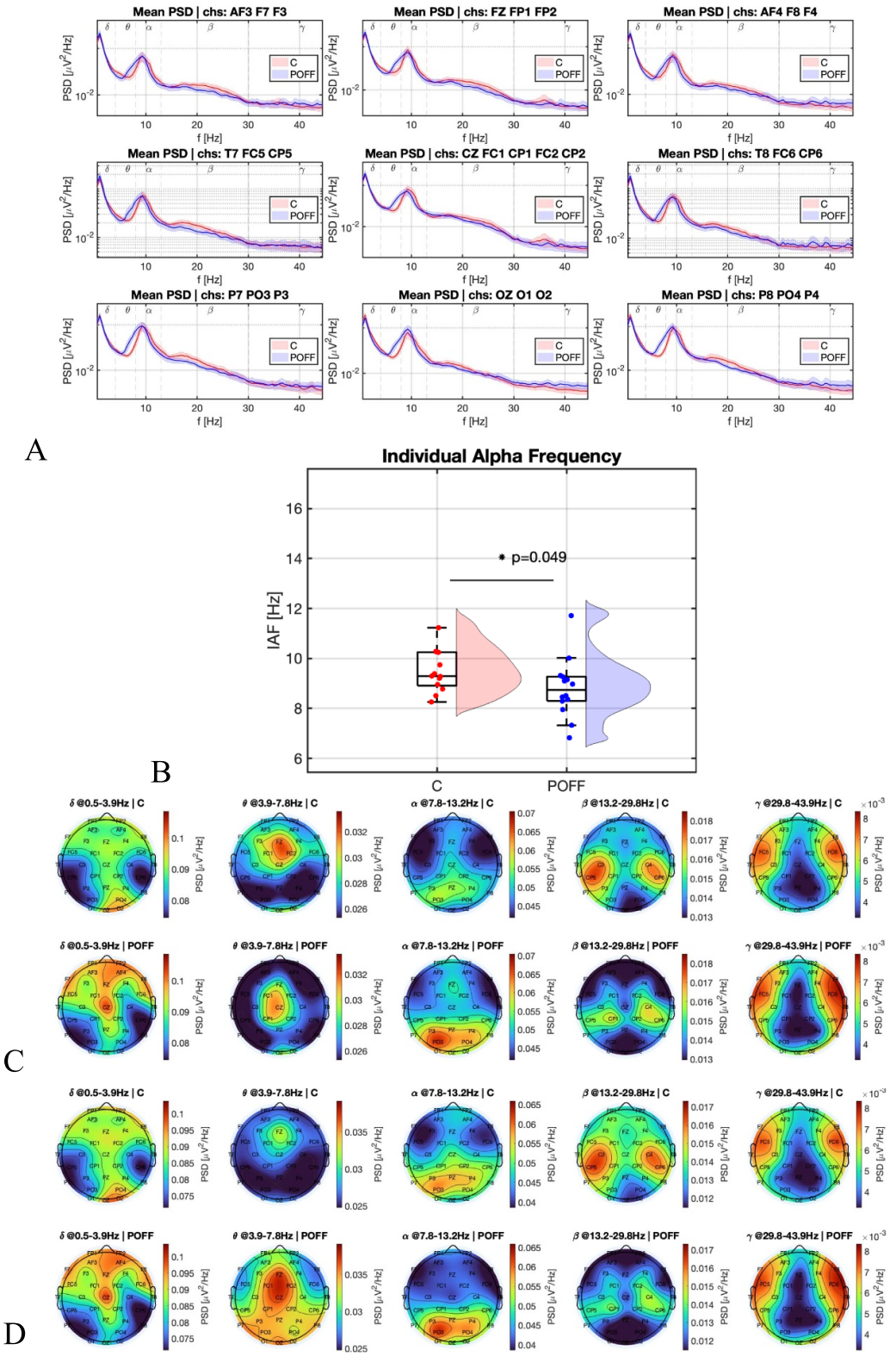

Figure 6. ds002778 results. Panel (A), shows the mean power spectral density (PSD) for the channels considered, for red group Control (C) and blue group Parkinson without medication (POFF). Panel (B), shows the difference in the Individual Alpha Frequency (IAF) between the two groups. Panels (C) and (D), show the average-group topographies, with colour bars representing the PSD, where data have been normalised. In panel (C) topographies are IAF corrected, while in panel (D) not.

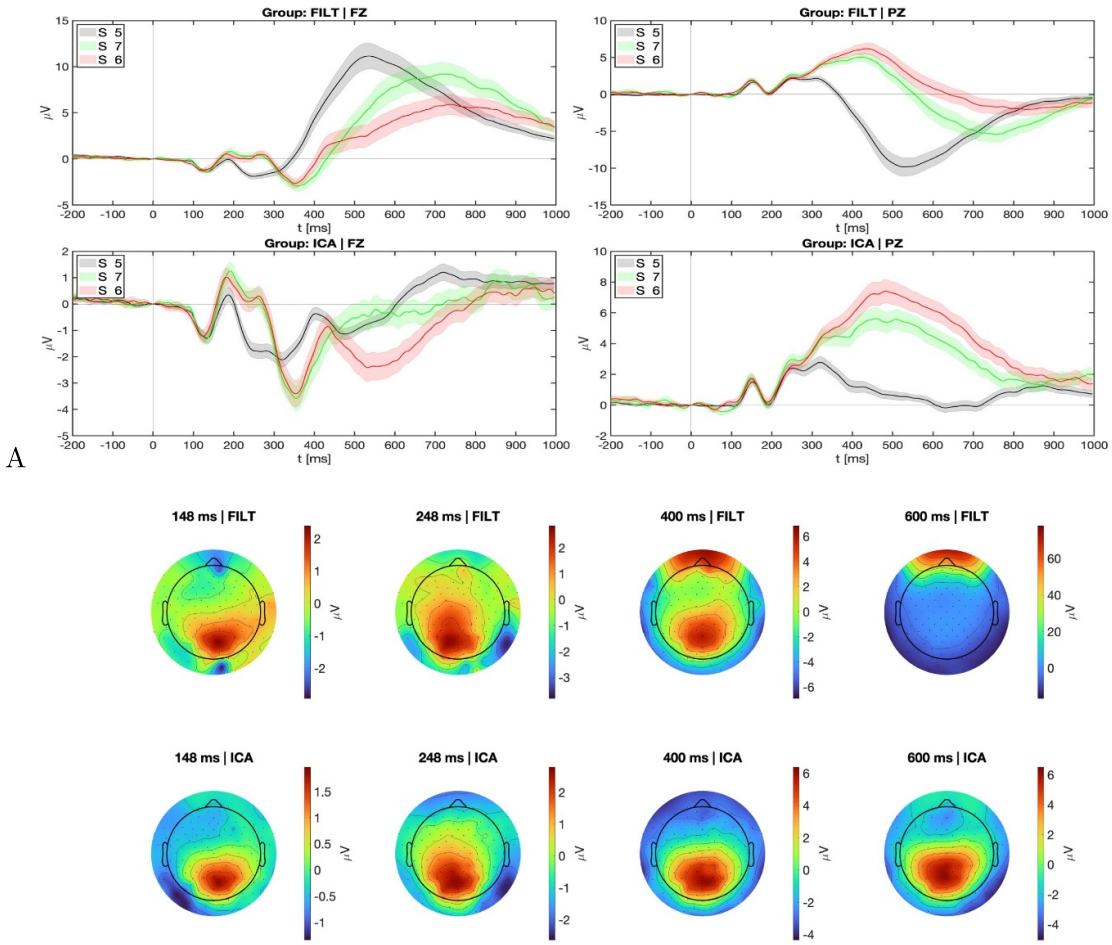

Figure 7. ds004621 results. (Panel (A) Grand average ERP image for channels FZ (left) and PZ (right), with two pipelines. The 'FILT' pipeline comprises filtering and re-referencing to average, while the 'ICA' pipeline adds ICA as an auxiliary step, where the components have been removed automatically by ICLabel; they are respectively at the top and bottom of the image. In black line there are events S5 (standard stimulus), in green events S7 (target stimulus) and in red events S8 (deviant stimulus). Bands in lighter colours represent one standard deviation across subjects. Panel (B), shows the grand average ERP scalp topography at various time points, for data preprocessed with the 'FILT' pipeline (top) and the 'ICA pipeline' (down).

data were average re-referenced. A non-parametric permutation *t*-test shows statistical significant results with *p*-value threshold *a priori* determined of 0.05.

4.2.1. ds004148

The power spectrum density (PSD, figure 4, panel (A)) was calculated on the following ROIs: frontal-left (AF7, AF3, F7, F5, F3 and F1), fronto-central (FZ, FPZ, FP1 and FP2), frontal-right (AF8, AF4, F8, F6, F4 and F2), temporal-left (FT7, T7, TP7, FC5, C5, CP5), central (CZ, CPZ, FC1, C1, CP1, FC2, C2, CP2), temporal-right (FT8, T8, TP8, FC6, C6, CP6), parietal-left (P7, P5, P3, PO7, PO3), occipital (POZ,OZ,O1,O2) and parietal-right (P8, P6, P4, PO8, PO4), with the red line representing the mean ROI's channels for the Eyes Closed (EC) condition and the blue line representing the same for the Eyes Open (EO). Bands in lighter colours represent one standard deviation across subjects.

Statistical analysis differences in the delta, theta, alpha, beta, gamma band and the estimated IAF are shown in figure 4, panel (B); differences are found to be significant with $p < 0.001$, for all the EEG features except for the IAF. For the EC group, the median IAF is 10.3 Hz (25th–75th percentile: 9.9 Hz–10.9 Hz), while for the EO group the median IAF is 10.6 Hz (25th–75th percentile: 9.9 Hz–11.2 Hz).

The PSD topographic maps (first two rows) plus the statistical differences between the two conditions (EC vs EO, third row) in each band are shown in figure 4, panel (C). Statistically significant ROIs were found in each band; indeed, in the delta band, there is a significant bilateral increase in activation in the frontal area for the EC condition relative to the EO condition. The theta band shows a significant decrease in occipital activation for the EC condition together with an increase in activation in the frontal-central area for the EO condition. In the alpha band instead, the activation in the occipital area for the EC

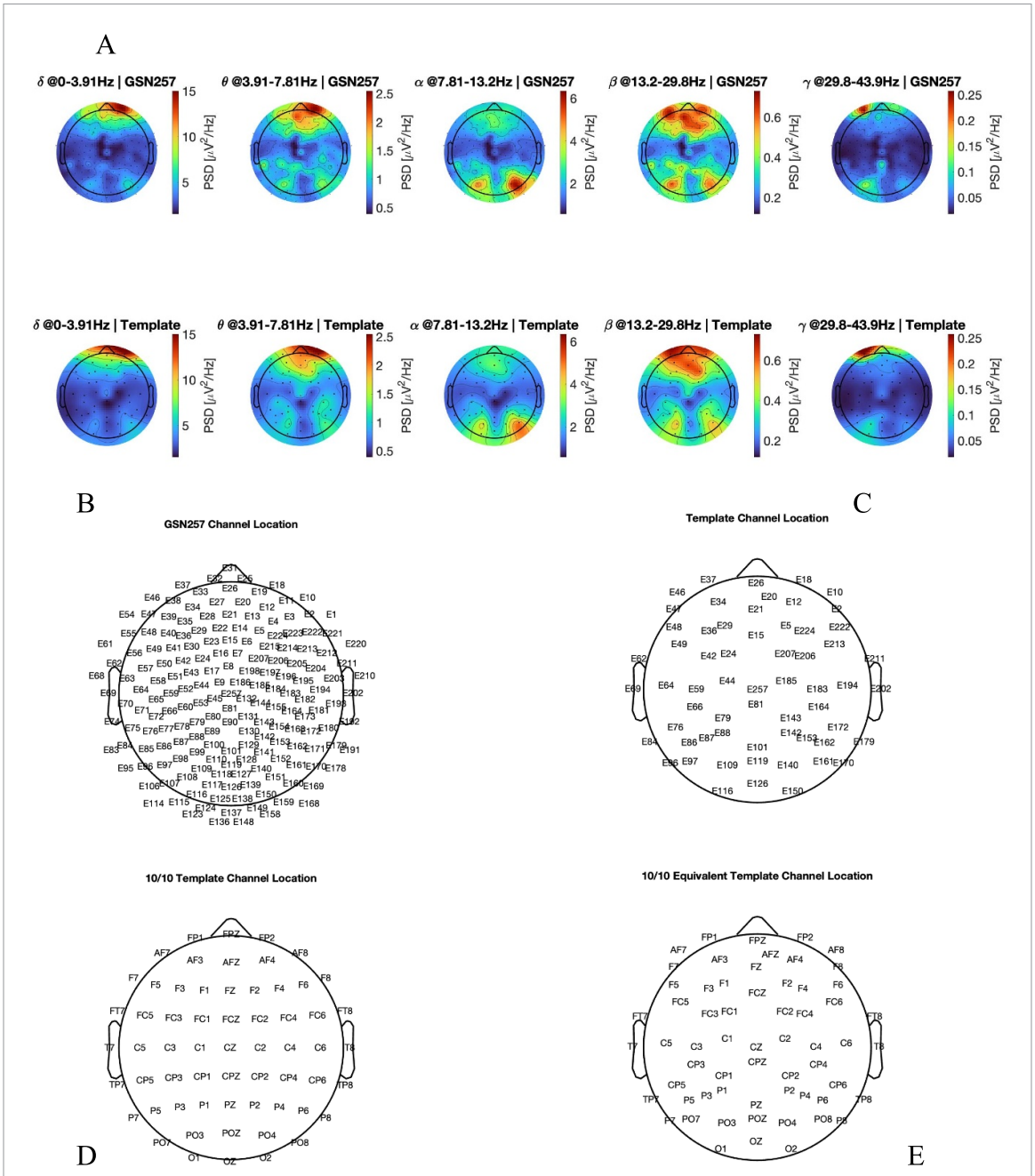

Figure 8. ds003421 results. (Panel (A) The top row shows PSD topographies for the five EEG bands, considering all the channels from the HCGSN-257 channel system. The bottom row shows the same, considering only the 61 channels from the chosen template (figure 3). Panels (B)–(C)–(D)–(E) show the differences between the channel locations of the HCGSN257 and the chosen template. In more detail, Panel (B) shows the electrode location in the HCGSN257 system, Panel (C) shows the channels specified in the template. Panel (E) shows the same as panel (C) but with the 10–10 respective names, while Panel (D) shows the 10–10 standard channel location.

condition was found to be significantly greater respect the EO condition.

4.2.2. ds004504

The PSD (figure 5, panel (A)) was calculated on the following ROIs: frontal-left (F7 and F3), frontal-central (FZ, FP1 and FP2), frontal-right (F8 and F4), temporal-left (T7), central (CZ), temporal-right (T8), parietal-left (P7, P3), occipital (O1 and O2) and parietal-right (P8 and P4), with the red line representing the mean of the ROI's channels for the healthy Control (C) group and the blue line representing the Alzheimer (A) group. Bands in lighter colours represent one standard deviation across subjects.

Statistical analysis differences in the delta, theta, alpha, beta, gamma band and the estimated IAF are shown in figure 5, panel (B); differences are found to be significant with $p < 0.001$, for all the EEG features except for the gamma band. For the C group the median IAF is 9.8 Hz (25th–75th percentile: 9.0 Hz–10.2 Hz), while for the A group the median IAF is 8.5 Hz (25th–75th percentile: 7.4 Hz–10.0 Hz).

The PSD topographic maps (first two rows) plus the statistical differences between the two groups (C

Table 3. Different pipelines used and reported in figure 9. '2ASRc' indicates that the automatically identified bad-time windows are corrected instead of been removed as done in '2ASR'.

Preprocessing step	FILT	ICA	1ASR	ICA1ASR	2ASR	2ASRc
Non EEG-channels removal	×	×	×	×	×	×
Time segments removal						
Baseline removal	×	×	×	×	×	×
Resampling						
Filtering	×	×	×	×	×	×
Independent component analysis		×		×		
Automatic component rejection		×		×		
wICA artifact suppression						
Line noise removal						
Bad-channel removal			×	×		
Bad-time windows removal					×	
Bad-time windows correction						×
Spherical interpolation	×	×	×	×	×	×
Re-reference	×	×	×	×	×	×

vs A, third row) in each band are shown in figure 5, panel (C). Statistically significant ROIs were found in each band; for both the delta and the theta bands, the Alzheimer group shows higher activation equivalently in all the scalp. In the alpha band instead, the activation in the occipital and frontal area for the A group was found to be significantly greater respect the C group, while the central area was not found to be statistically significant. There are no significant differences in the gamma band demonstrated by two equivalent scalp topographies for the two groups.

4.2.3. ds002778

The PSD (figure 6, Panel (A)) was calculated on the following ROIs: frontal-left (AF3, F7 and F3), frontal-central (FZ, FP1 and FP2), frontal-right (AF4, F8 and F4), temporal-left (T7, FC5 and CP5), central (CZ, CP1, CP2, FC and FC2), temporal-right (T8, FC6 and CP6), parietal-left (P7, P3 and PO3), occipital (OZ, O1 and O2) and parietal-right (P8, P4 and PO4), with the red line representing the mean of the ROI's channels for the healthy Control (C) group and the blue line representing the Parkinson (POFF) group ('OFF', because without medication). Bands in lighter colours represent one standard deviation across subjects.

Statistical analysis differences for the estimated IAF is shown in figure 6, panel (B). Differences are found to be significant with $p = 0.049$, where for the C group the median IAF is 9.3 Hz (25th–75th percentile: 8.9 Hz–10.3 Hz), while for the POFF group the median IAF is 8.7 Hz (25th–75th percentile: 8.3 Hz–9.3 Hz).

In panels C and D instead there are the topographies for the delta, theta, alpha, beta and gamma bands IAF corrected and non-IAF corrected respectively. Delta, beta and gamma bands are less affected by the IAF correction showing insignificant differences. A decrease in activation in the occipital lobe in the theta band is evident for the 'POFF' group comparing panels (C) and (D), as well as an increase in activation in the frontal area. The control group instead shows an increase in the occipital area in the alpha band. In the theta and alpha band, there are differences in the scalp topographies as described before, however statistically significant ROIs were not found, either in IAF and non-IAF comparisons or in the between-group comparison.

4.2.4. ds004621

The grand average ERP (figure 7, panel (A)) was calculated for channels FZ and PZ, with the black line representing the mean ERP for the standard stimulus (S5), the green line representing the same for the deviant stimulus (S7) and the red line representing the same for the target stimulus (S6). Bands in lighter colours represent one standard deviation across subjects. The two upper plots show the three stimuli for each channel with data preprocessed with the 'FILT' pipeline, while the two bottom plots show the same but with data processed with the 'ICA' pipeline. The 'FILT' pipeline comprises filtering between [0.5, 40] Hz, re-referencing to average and segmenting data into segments from 200 ms before stimulus presentation to 1000 ms after stimulus presentation as done in the original paper [57]. The 'ICA' pipeline adds ICA as an auxiliary step, where the components have been removed automatically by ICLabel. The results of the 'ICA' pipeline and those shown in [57] (see figure 5 in their paper), highly agree at visual inspection.

Figure 7, panel (B) shows the scalp activation in μV at four time steps, where in the first raw data have been processed with the 'FILT' pipeline, while in the second with the 'ICA' pipeline. Both figures agree in showing how the 'ICA' pipeline effectively removes the blinks evident in figure 7, panel (B) at 400 ms and 600 ms. The time steps were selected to catch the scalp activity of the visual stimulus V1 (100 ms),

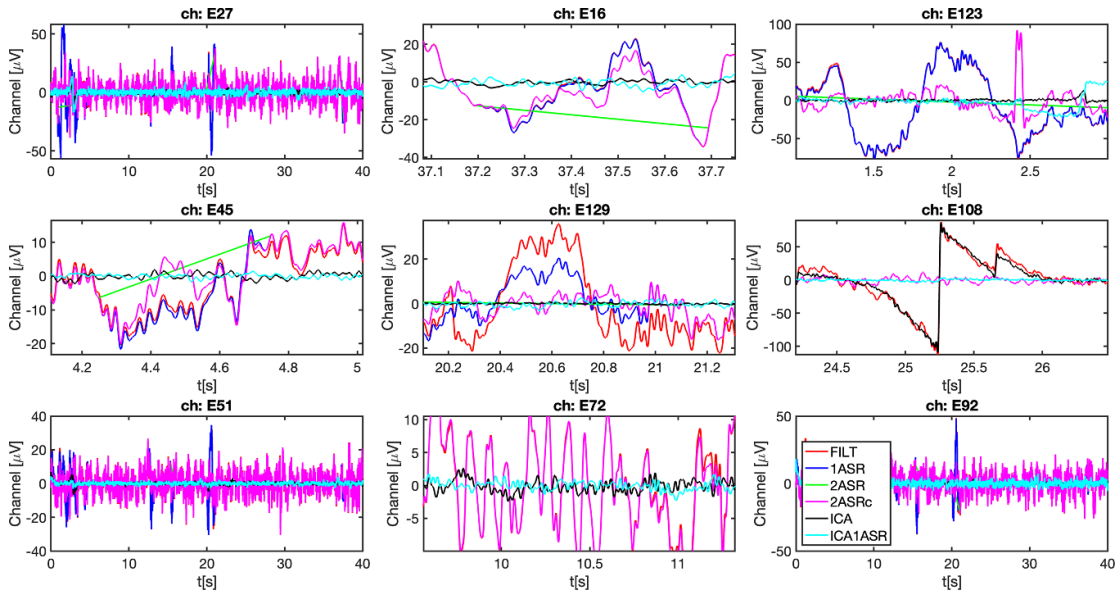

Figure 9. ds004186 results. It is reported the EEG signal for some channels (E27-F5; E16-AFZ; E123-F6; E45-T7; E6-FCZ; E108-T8; E51-P5; E72-POZ; E92-P4). The pipelines applied (see table 3) have the following colours: in red ('FILT'), the signal is filtered with baseline removal (these two operations are applied to all the other pipelines). The line in dark blue ('1ASR') shows the signal with automatic bad-channel removal; in green ('2ASR') instead the signal with automatic bad-channel removal and bad time segments removal; the line in magenta ('2ASRc') shows the signal with automatic bad-channel removal and bad time segments correction. The line in black ('ICA') shows the signal with automatic ICs removal; finally in cyan ('ICA1ASR') instead the signal with both automatic IC rejection and bad-channel removal.

N2b (200–250 ms), P3a (300–350 ms) and P3b (300–500 ms) and blink (600 ms).

4.3. Library validation: subject analysis

The results of this section aim to show the possible visualisations that BIDSAlign can create for single subject analysis, demonstrating how they can be useful and meaningful for the user. All the results are shown for the first subject, for the first session if multiple sessions are present, and for the first EEG file if there are multiple trials.

4.3.1. ds003421

The results in figure 8, since this dataset has EEG recordings with 257 channels, show how the scalp topography changes, moving from the original channel template (HCGSN-257) to the converted 10–10 system with the chosen template (see figure 3). The data in this case was filtered between $[1, 45]$ Hz and automatic IC rejection was applied.

Panel (A) shows the PSD for the delta, theta, alpha, beta and gamma bands for the original channel system (top) and the converted one with the chosen template (bottom). The topographies are more detailed in the original channel system due to the highest number of channels, but the areas and the intensity of the activation are maintained in the new channel system; the legends have the same ranges for an easier comparison.

Panels E and D show the effects of the electrodes placement due to the conversion file [29]. Despite the differences in electrode placement and symmetry between the two channel systems (HCGSN and the 10–10 international standard), the HCGSN-257 electrodes selected according to the template (Panel (E)) cover the scalp well, positioned approximately correctly relative to the actual 10–10 system. In [29] the reader can find more information on the arc length (in cm) between an electrode in one channel system and its equivalent in the other.

4.3.2. ds004186

Since the files of this datasets are short (40 s) and difficult to preprocess as mentioned in the dataset description, this section aims to shows how different pipelines effect the EEG recording and how BIDSAlign can help to visualize it. In figure 9 is reported the temporal EEG signal for some channels (E27-F5, E16-AFZ, E123-F6, E45-T7, E6-FCZ, E108-T8, E51-P5, E72-POZ and E92-P4). Six pipelines have been tested, with different combination of preprocessing steps; the original sampling rate (500 Hz) was maintained.

Filtering was done between 1–45 Hz and re-referencing was done to the common average. ICA was performed with fastICA considering all the components; ICLabel was used to automatic IC rejection. The second ASR, in which time segments are removed, is activated whenever one channel has an absolute voltage value above the voltage threshold of 250 μV.

Different parts of the signal have been selected and zoomed in details, in order to inspect the signal and get insightful considerations. In the subplots of

channels E45(T7) and E16(AFZ), straight green lines show where the ASR algorithm identifies bad time portions of the signal. In these time windows, the user can compare red and blue lines ('FILT' and '1ASR' pipelines) with magenta line ('2ASR' pipeline) showing how bad time segments are corrected by the ASR algorithm. While in channels E45(T7) and E16(AFZ) magenta signal closely follows the red and blue signals indicating an ineffective repairing, in channel E129(CZ) the corrected channel is indeed better since closer to a biological signal respect the red and blue ones. Moreover, while in channel E108(T8) both '2ASRc' and 'ICA1ASR' pipelines performs equally well in eliminating the divergence at 25.25 s, channels E72(POZ) shows how the pipelines with ICA step ('ICA' and 'ICA1ASR') have a definitely smaller range of values respect pipelines with non-ICA step (e.g. 'FILT' and '2ASRc'). In conclusion figure 9 shows the ineffectiveness of filtering and ICA decomposition alone for this EEG recording while demonstrating the great effectiveness of the ASR algorithm in finding bad portions of the signal and up to some extent the ability to correct them.

5. Discussion

In the last decade, greater attention has been paid to brain electrophysiological substrates in healthy controls [58, 59] and in neurodegenerative diseases in both rest conditions [3, 4] and during specific tasks [60, 61]. The growing interest in using electrophysiological techniques such as EEG compared to other neuroimaging techniques such as fMRI and PET is mainly due to the low cost of EEG, widespread availability in clinical settings, non-invasiveness and portability, which makes this unique tool of its kind to be used at the patient's bedside [62, 63]. Furthermore, the proliferation of electrophysiological markers developed over the past decade [62, 63], along with the growing availability of EEG datasets, further supports this idea.

5.1. BIDSAlign

Even-though data production is enormous in the medical field, large datasets properly acquired, archived, shared and ready-to-use are few but fortunately growing in number. Platforms like EEGLAB and tools like bids-Matlab-tool, lend a hand in this direction, since help and encourage researchers to share data, in a defined and standard format. However, even if the number of such datasets is increasing year after year, their size is typically small; for example, on the OpenNeuro website, there are 189 EEG datasets, 75 have less than 20 subjects and 146 have less then 50 subjects. There are, of course, big datasets for example MPI-LEMON [40] (215 subjects) and TDBRAIN [15] (1200 subjects) but both of them unfortunately not stored in BIDS format. BIDSAlign helps researchers in this direction, because it was designed to handle both BIDS and non-BIDS datasets and needs a minimal amount of user-provided information.

If one starts to explore the large population of datasets, as mentioned before, many challenges could arise in the effective use of them. Authors have encountered cases of datasets saved in the *.vhdr* format, in which the sampling rate was a non-integer number, or pointing to the wrong *.eeg file*; there were some other cases in which the *_electrodes.tsv* file had channel names totally different with respect the usual GSN channel names that are composed of 'E' and the number of the channel. In other cases, the channel labels inside the EEG structure were anonymized, specified simply by 'A' and the channel number; moreover, they were different with respect to the channel names present in the channel location; some other datasets do not provide the channel location in any form. Finally the EEG reference is another problematic aspect of public datasets. Sometimes, the channel reference is present in the EEG structure but in others it is not, or even reported wrongly. Other times it was automatically added and indicated as 'average reference' when it was not. All these cases mine the use of such data from the user side, discouraging the use of public datasets. BIDSAlign manages all the cases described above, unlocking these datasets and making them available for scientific research.

One of the main aims of BIDSAlign is to guide DL researchers from data import, to preprocessing and result visualisation. The structure of the library was designed entirely around harmonisation to a common template, which is essential to carry out DL on multiple datasets. Data with different references, different sampling rates and different handling of the artefacts can cause respectively biases, under-representation of the subjects and an anomalous divergent gradient for windows with abnormal power. Thus, a visual inspection of the data used for training architectures, before the actual training, is a suggested good practice. Pipelines and group comparison, single file inspection and visualisation and group topographies help, in a concise and meaningful way, DL researchers gain insight into the data their DL architecture will be trained on. BIDSAlign indeed can fulfill this task, with appropriate visualisation functions, and for more expert users, the source code can be easily modified with the preferred style.

The library, together with all the aforementioned objectives, touches on another important aspect, which is the reproducibility of the results in neuroscience. Certainly, BIDSAlign offers a convenient means of preprocessing raw data from scientific studies that have made their data publicly available, enabling straightforward comparisons. While the primary source for reproducibility remains the original code used by the authors, there are instances where this code may not be publicly

accessible. Additionally, independently reproducing results demonstrates the robustness of the obtained findings.

Most importantly, in terms of the reproducibility of the results in the DL field, it is extremely important to try to keep all the possible confounds under control, integrating the various sources of heterogeneity. That is why works like BIDSAlign and selfEEG [64] are important: if the first as seen in this work, uniforms the preprocessing and the template, and thus the data, the latter uniforms how data are split, loaded and handled in a DL environment (e.g. PyTorch [43]). In this way differences between classifiers (for example pathology classifiers), are due to the effective improvement of the state of the art, and not to different preprocessing or due to wrong splits of the data. Finally, comparison between classifiers should not be done by looking at the mean test accuracy, that is only one of the many metrics that can be used, but for example employing statistical tests like the McNemar's test as reported in [65]).

5.2. Group analysis

In this part of the paper, based on BIDSAlign, we have investigated six available datasets, as reported in table 2.

For the first dataset, here indicated as ds004148 [66], we have replicated the main results obtained in the original publication session 1 with a clear increase in the alpha peak for the EC with respect to the EO as physiologically expected. Other independent studies that used this dataset found similar results [67, 68]. In addition, differences in the IAF were tested since it is a consolidated electrophysiological marker of inter-individual differences in cognitive function [47], ageing [69] and many neurodegenerative pathologies [61]. How IAF changes with ageing, can be viewed in the supplementary material. Moreover, IAF provides an empirical basis for defining individualised frequency bands [47]. Here, no difference between EO and EC was found was found as well, as physiologically expected.

For the second dataset, named ds004504, we have extended the results obtained in the original article for Alzheimer's vs. the healthy control [70]. A higher delta and theta band in Alzheimer's with respect to control was found, as also reported in the literature [3, 71, 72]. There are also differences in the IAF, with a lower alpha peak for Alzheimer's with respect to the Control group, as also well reported in the literature [3, 71, 72]. Other independent studies that used this dataset found similar results [73, 74].

Instead, for the third dataset, ds002778, we did not find a difference in the beta band relative to the original paper [75]. The reason might be the normalisation applied to our data. However, in the original work, a difference in PSD was found in almost all the frequency bands, which might be due to the difference in power between the EEG recorded for the healthy controls concerning Parkinson's. This idea is supported by the fact that no differences between Parkinson's with medication ON (PON) and Parkinson's with medication OFF (POFF) were observed, at the spectral level. Other independent studies that used this dataset found similar results [76]. In our preprocessed data, the beta band decreases for the POFF concerning the control group (C), even if no significant value was reached (see figure 6). Interestingly, the topography map between the C and POFF seems to be not physiological (see supplementary material figure 2, panel (C)) compared to the normalised one (see figure 6, panel (C)). The same patterns is shown for POFF vs. PON non-normalised (see figure 2, Panel (B)) and normalised (see figure 2, panel (A)) data.

The dataset ds004621 [57] allowed to verify the automation capabilities of BIDSAlign. In fact, matched ERP results were obtained with respect to the original paper [57] in a fully automatic fashion compared to the semi-supervised one used. This result strongly supports the performance of BIDSAlign tool on one side and demonstrates the robustness of the results obtained on the other.

Finally, two additional datasets ds003421 [55] and ds004186 [13] show how BIDSAlign is well-equipped to handle databases with no standard electrode positions and very noisy data, such as the one recorded in children.

6. Conclusion

This paper presents the library *BIDSAlign*, which has the objective of helping the research community preprocess EEG data and standardise heterogeneous EEG datasets, particularly for DL applications. On a broader level, it provides a one-fits-all solution for preprocessing EEG data, since the user can decide to preprocess one file, a portion of a dataset, an entire one, or even more datasets in one single run. Preprocessing can be easily applied to subjects split by age, pathology, medication or even the values of diagnostic scales. For the DL community, BIDSAlign offers a way, to collect and integrate heterogeneous EEG datasets, extract data usable for training DL architectures, unify them to a common template, and integrate sources of heterogeneity. The effectiveness of the library was demonstrated with six datasets, in which for four of them, medically meaningful insights were retrieved in a fully-unsupervised way, from data

loading to result visualisation. Moving forward, continued contributions and adaptations to the library will further extend its utility, ensuring its relevance in evolving biomedical data storage practices.

Data availability statement

No new data were created or analysed in this study.

Funding

This work was supported by the STARS@UNIPD funding program of the University of Padova, Italy, through the project: MEDMAX. This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101137074—HEREDITARY.

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Code availability

Code is available at https://github.com/MedMaxLab/BIDSAlign.

Authors' contributions

Conceptualization from AZ, CP and MA. The library was coded by AZ and FDP. Dataset search and selection was done by AZ. Data curation and analysis was performed by AZ and CP. Medical interpretation was performed by CP. The first draft of the manuscript was written by AZ and all authors commented and edited previous versions of the manuscript. All authors read and approved the final manuscript.

ORCID iDs

Andrea Zanola https://orcid.org/0000-0001-6973-8634
Federico Del Pup https://orcid.org/0009-0004-0698-962X
Camillo Porcaro https://orcid.org/0000-0003-4847-163X
Manfredo Atzori https://orcid.org/0000-0001-5397-2063

References

[1] Tecchio F, Cottone C, Porcaro C, Cancelli A, Di Lazzaro V and Assenza G 2018 Brain functional connectivity changes after transcranial direct current stimulation in epileptic patients *Front. Neural Circuits* **12** 44
[2] Averna A, Coelli S, Ferrara R, Cerutti S, Priori A and Bianchi A M 2023 Entropy and fractal analysis of brain-related neurophysiological signals in Alzheimer's and Parkinson's disease *J. Neural Eng.* **20** 051001
[3] Porcaro C, Vecchio F, Miraglia F, Zito G and Rossini P M 2022 Dynamics of the 'cognitive' brain wave P3B at rest for Alzheimer dementia prediction in mild cognitive impairment *Int. J. Neural Syst.* **32** 2250022
[4] Smits F M, Porcaro C, Cottone C, Cancelli A, Rossini P M and Tecchio F 2016 Electroencephalographic fractal dimension in healthy ageing and Alzheimer's disease *PLoS One* **11** 1–16
[5] Bigdely-Shamlo N, Mullen T, Kothe C, Su K M and Robbins K A 2015 The PREP pipeline: standardized preprocessing for large-scale EEG analysis *Front. Neuroinform.* **9** 16
[6] Gabard-Durnam L, Méndez Leal A, Wilkinson C and Levin A 2018 The harvard automated processing pipeline for electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data *Front. Neurosci.* **12** 97
[7] Winkler I, Haufe S and Tangermann M 2011 Automatic classification of artifactual ICA-components for artifact removal in EEG signals *Behav. Brain. Funct.* **7** 30
[8] Hyvärinen A and Oja E 2000 Independent component analysis: algorithms and applications *Neural Netw.* **13** 411–30
[9] Bell A J and Sejnowski T J 1995 An information-maximization approach to blind separation and blind deconvolution *Neural Comput.* **7** 1129–59
[10] Cowley B U and Korpela J 2018 Computational testing for automated preprocessing 2: practical demonstration of a system for scientific data-processing workflow management for high-volume EEG *Front. Neurosci.* **12** 236
[11] Rafiei M H, Gauthier L V, Adeli H and Takabi D 2022 Self-supervised learning for electroencephalography *IEEE Trans. Neural Netw. Learn. Syst.* **35** 1457–71
[12] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk T H and Faubert J 2019 Deep learning-based electroencephalography analysis: a systematic review *J. Neural Eng.* **16** 051001
[13] Alexander L M, Escalera J, Ai L, Andreotti C, Febre K, Mangone A, Vega-Potler N, Langer N, Alexander A and Kovacs M 2022 HBN EO/EC task *OpenNeuro* (https://doi.org/10.18112/openneuro.ds004186.v2.0.0)
[14] Obeid I and Picone J 2016 The temple university hospital EEG data corpus *Front. Neurosci.* **10** 196
[15] Dijk H, van Wingen G, Denys D, Olbrich S, Ruth R and Arns M 2022 The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database *Sci. Data* **9** 333
[16] Silva C P D, Tedesco S and O'Flynn B 2024 EEG datasets for healthcare: a scoping review *IEEE Access* **12** 39186–203
[17] Mohsenvand M N, Izadi M R and Maes P 2020 Contrastive representation learning for electroencephalogram classification (*Proceedings of the Machine Learning for Health NeurIPS Workshop) Proc. Machine Learning Research* vol 136 ed E Alsentzer, M B A McDermott, F Falck, S K Sarkar, S Roy and S L Hyland (PMLR) pp 238–53
[18] Delorme A, Truong D, Martinez-Cancino R, Pernet C, Sivagnanam S, Yoshimoto K, Poldrack R, Majumdar A and Makeig S 2021 *Tools for Importing and Evaluating BIDS-EEG Formatted Data 2021 10th Int. IEEE/EMBS Conf. on Neural Engineering (NER)* pp 210–3
[19] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
[20] Kemp B, Värri A, Rosa A C, Nielsen K D and Gade J 1992 A simple format for exchange of digitized polygraphic recordings *Electroencephalogr. Clin. Neurophysiol.* **82** 391–3
[21] Gorgolewski K J *et al* 2016 The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments *Sci. data* **3** 1–9
[22] Pernet C R, Appelhoff S, Gorgolewski K J, Flandin G, Phillips C, Delorme A and Oostenveld R 2019 EEG-BIDS, an extension to the brain imaging data structure for electroencephalography *Scientific Data* **6** 103

[23] Jasper H H 1958 Report of the committee on methods of clinical examination in electroencephalography *Electroencephalogr. Clin. Neurophy.* **10** 371–5
[24] Nuwer M R 1987 Recording electrode site nomenclature *J. Clin. Neurophys.* **4** 121–34
[25] Chatrian G E, Lettich E and Nelson P L 1988 Modified nomenclature for the '10%' electrode system *J. Clin. Neurophysiol.* **5** 183–6
[26] Klem G H, Lüders H O, Jasper H H and Elger C 1999 Ten-twenty electrode system of the international federation *Electroencephalogr. Clin. Neurophys.* **52** 3–6
[27] Oostenveld R and Praamstra P 2001 The five percent electrode system for high-resolution EEG and ERP measurements *Clin. Neurophys.* **112** 713–9
[28] Tucker D M 1993 Spatial sampling of head electrical fields: the geodesic sensor net *Electroencephalogr. Clin. Neurophysiol.* **87** 154–63
[29] Luu P and Ferree T 2000 *Determination of the Geodesic Sensor Nets' Average Electrode Positions and Their 10–10 International Equivalents Technical Note Electrical Geodesics* (Inc. Eugene)
[30] Junghöfer M, Elbert T, Tucker D M and Rockstroh B 2000 Statistical control of artifacts in dense array EEG/MEG studies *Psychophysiology* **37** 523–32
[31] Barbati G, Porcaro C, Zappasodi F, Rossini P M and Tecchio F 2004 Optimization of an independent component analysis approach for artifact identification and removal in magnetoencephalographic signals *Clin. Neurophys.* **115** 1220–32
[32] Pion-Tonachini L, Kreutz-Delgado K and Makeig S 2019 ICLabel: an automated electroencephalographic independent component classifier, dataset and website *NeuroImage* **198** 181–97
[33] Nolan H, Whelan R and Reilly R 2010 FASTER: fully automated statistical thresholding for eeg artifact rejection *J. Neurosci. Methods* **192** 152–62
[34] Fischler M A and Bolles R C 1981 Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography *Commun. ACM* **24** 381–95
[35] Pedroni A, Bahreini A and Langer N 2019 Automagic: standardized preprocessing of big EEG data *NeuroImage* **200** 460–73
[36] Mullen T R, Kothe C A E, Chi Y M, Ojeda A, Kerth T, Makeig S, Jung T P and Cauwenberghs G 2015 Real-time neuroimaging and cognitive monitoring using wearable dry EEG *IEEE Trans. Biomed. Eng.* **62** 2553–67
[37] Markiewicz C J *et al* 2021 The openneuro resource for sharing of neuroscience data *eLife* **10** e71774
[38] Alzahab N A, Apollonio L, Di Iorio A, Alshalak M, Iarlori S, Ferracuti F, Monteriú A and Porcaro C 2021 Hybrid deep learning (hDL)-based brain-computer interface (BCI) systems: a systematic review *Brain Sci.* **11** 75
[39] Oliver-Taylor A, Li A, Thomas A, Flinker A, Wagner A S, Karakuzu A, Nikolaidis A, Lazari A, de la Vega A and Giacomel A 2023 BIDS-specification (https://doi.org/10.5281/zenodo.10175846)
[40] Babayan A *et al* 2019 A mind-brain-body dataset of MRI, EEG, cognition, emotion and peripheral physiology in young and old adults *Sci. Data* **6** 1–21
[41] Castellanos N P and Makarov V A 2006 Recovering eeg brain signals: artifact suppression with wavelet enhanced independent component analysis *J. Neurosci. Methods* **158** 300–12
[42] Miltiadous A *et al* 2024 A dataset of EEG recordings from: Alzheimer's disease, frontotemporal dementia and healthy subjects *OpenNeuro* (https://doi.org/10.18112/openneuro.ds004504.v1.0.7)
[43] Paszke A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.)
[44] Saggar M *et al* 2012 Intensive training induces longitudinal changes in meditation state-related EEG oscillatory activity *Front. Hum. Neurosci.* **6** 256
[45] Klimesch W 1999 EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis *Brain Res. Rev.* **29** 169–95
[46] Corcoran A, Alday P and Bornkessel-Schlesewsky I 2019 corcorana/restingIAF: restingIAF v1.0.3 (https://doi.org/10.5281/zenodo.2575868)
[47] Corcoran A W, Alday P M, Schlesewsky M and Bornkessel-Schlesewsky I 2018 Toward a reliable, automated method of individual alpha frequency (IAF) quantification *Psychophysiology* **55** e13064
[48] Glerean E 2016 hfASDmodules: reorganization of functionally connected subnetworks in autism *Zenodo* (https://doi.org/10.5281/zenodo.44657)
[49] Glerean E *et al* 2016 Reorganization of functionally connected brain subnetworks in high-functioning autism *Hum. Brain Mapp.* **37** 1066–79
[50] Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* B **57** 289–300
[51] Sinha S R, Sullivan L, Sabau D, San-Juan D, Dombrowski K E, Halford J J, Hani A J, Drislane F W and Stecker M M 2016 American clinical neurophysiology society guideline 1: minimum technical requirements for performing clinical electroencephalography *J. Clin. Neurophysiol.* **33** 303–7
[52] Wang Y, Duan W, Dong D, Ding L and Lei X 2022 A test-retest resting and cognitive state EEG dataset *OpenNeuro* (https://doi.org/10.18112/openneuro.ds004148.v1.0.1)
[53] Rockhill A P, Jackson N, George J, Aron A and Swann N C 2021 UC San Diego resting state EEG data from patients with parkinson's disease *OpenNeuro* (https://doi.org/10.18112/openneuro.ds002778.v1.0.5)
[54] George J S, Strunk J, Mak-McCully R, Houser M, Poizner H and Aron A R 2013 Dopaminergic therapy in Parkinson's disease decreases cortical beta band coherence in the resting state and increases cortical beta band power during executive control *NeuroImage: Clin.* **3** 261–70
[55] Mheich A, Dufor O, Yassine S, Kabbara A, Biraben A, Wendling F and Hassan M 2021 HD-EEG for tracking sub-second brain dynamics during cognitive tasks *Sci. Data* **8** 32
[56] Dzianok P, Antonova I, Wojciechowski J, Dreszer J and Kublik E 2023 The Nencki-Symfonia EEG/ERP dataset (https://doi.org/10.18112/openneuro.ds004621.v1.0.1)
[57] Dzianok P, Antonova I, Wojciechowski J, Dreszer J and Kublik E 2022 The Nencki-Symfonia electroencephalography/event-related potential dataset: multiple cognitive tasks and resting-state data collected in a sample of healthy adults *GigaScience* **11** giac015
[58] Liu Q, Ganzetti M, Wenderoth N and Mantini D 2018 Detecting large-scale brain networks using EEG: impact of electrode density, head modeling and source localization *Front. Neuroinform.* **12** 4631–43
[59] Porcaro C, Cottone C, Cancelli A, Salustri C and Tecchio F 2018 Functional semi-blind source separation identifies primary motor area without active motor execution *Int. J. Neural Syst.* **28** 1750047
[60] Porcaro C, Mayhew S D and Bagshaw A P 2021 Role of the ipsilateral primary motor cortex in the visuo-motor network during fine contractions and accurate performance *Int. J. Neural Syst.* **31** 2150011
[61] Porcaro C, Balsters J H, Mantini D, Robertson I H and Wenderoth N 2019 P3b amplitude as a signature of cognitive decline in the older population: an EEG study enhanced by functional source separation *NeuroImage* **184** 535–46
[62] Comanducci A *et al* 2020 Clinical and advanced neurophysiology in the prognostic and diagnostic evaluation of disorders of consciousness: review of an IFCN-endorsed expert group *Clin. Neurophysiol.* **131** 2736–65

[63] Porcaro C, Nemirovsky I E, Riganello F, Mansour Z, Cerasa A, Tonin P, Stojanoski B and Soddu A 2022 Diagnostic developments in differentiating unresponsive wakefulness syndrome and the minimally conscious state *Front. Neurol.* **12** 778951
[64] Del Pup F, Zanola A, Tshimanga L F, Mazzon P E and Atzori M 2024 SelfEEG: a python library for self-supervised learning in electroencephalography *J. Open Source Softw.* **9** 6224
[65] Dietterich T G 1998 Approximate statistical tests for comparing supervised classification learning algorithms *Neural Comput.* **10** 1895–923
[66] Wang Y, Wei D, Dong D, Ding L and Lei X 2022 A test-retest resting and cognitive state EEG dataset during multiple subject-driven states *Sci. Data* **9** 566
[67] McKeown D J, Finley A J, Kelley N J, Cavanagh J F, Keage H A D, Baumann O, Schinazi V R, Moustafa A A and Angus D J 2023 Test-retest reliability of spectral parameterization by 1/f characterization using SpecParam *Cerebral Cortex* **34** bhad482
[68] Li P, Wang C, Li M, Xuan X, Zhou B and Li H 2023 Flexible silver/carbon nanotube-graphene oxide-polydimethylsiloxane electrode patch for electroencephalography language *Adv. Intell. Syst.* **5** 2300018
[69] Turner C *et al* 2023 Developmental changes in individual alpha frequency: recording EEG data during public engagement events *Imaging Neurosci.* **1** 1–14
[70] Miltiadous A, Tzimourta K D, Giannakeas N, Tsipouras M G, Afrantou T, Ioannidis P and Tzallas A T 2021 Alzheimer's Disease and Frontotemporal Dementia: a robust classification method of EEG signals and a comparison of validation methods *Diagnostics* **11** 1437
[71] Dauwels J, Vialatte F and Cichocki A 2010 Diagnosis of Alzheimer's disease from EEG signals: where are we standing? *Curr. Alzheimer Res.* **7** 487–505
[72] Rossini P M, Rossi S, Babiloni C and Polich J 2007 Clinical neurophysiology of aging brain: from normal aging to neurodegeneration *Prog. Neurobiol.* **83** 375–400
[73] Wang Z, Liu A, Yu J, Wang P, Bi Y, Xue S, Zhang J, Guo H and Zhang W 2024 The effect of aperiodic components in distinguishing alzheimer's disease from frontotemporal dementia *GeroScience* **46** 751–68
[74] Zheng X *et al* 2023 Diagnosis of alzheimer's disease via resting-state eeg: integration of spectrum, complexity and synchronization signal features *Front. Aging Neurosci.* **15**
[75] Jackson N, Cole S, Voytek B and Swann N 2019 Characteristics of waveform shape in Parkinson's disease detected with scalp electroencephalography *eNeuro* **6** 3
[76] Anjum M F, Dasgupta S, Mudumbai R, Singh A, Cavanagh J F and Narayanan N S 2020 Linear predictive coding distinguishes spectral eeg features of parkinson's disease *Parkinsonism Relat. Disord.* **79** 79–85