A Dataset Documentation

A.1 Motivation

Q1. **For what purpose was the dataset created?**

A1. The purpose of this dataset is to provide a comprehensive benchmark for evaluating the accuracy, efficiency, and scalability of current and future multi-task, multi-session, and multi-subject models in large-scale scenarios. Currently, there is no benchmark dataset available for comparing these models. Additionally, this dataset aims to serve as an intermediate representation, bridging the gap between the metadata-rich and heterogeneous NWB/MATLAB files and machine learning algorithms. By unifying data from these diverse sources, this dataset is prepared and formatted for direct use in machine learning models.

Q2. **Who funded the creation of the dataset?**

A2. This work was supported by NIH RF1-DA056404 and the Portuguese Recovery and Resilience Plan (PPR), through project number 62, Center for Responsible AI, and the Portuguese national funds, through FCT - Fundação para a Ciência e a Tecnologia - in the context of the project UIDB/04443/2020.

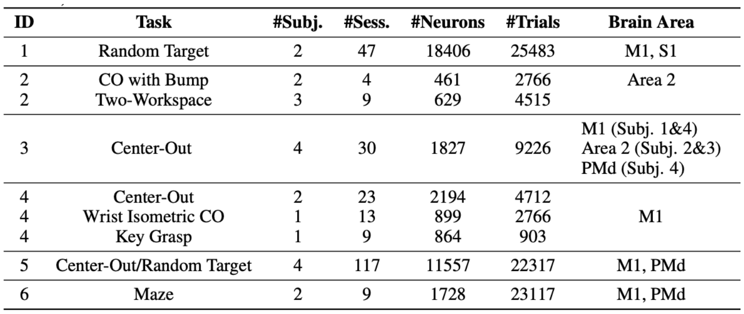
A.2 Composition

Q3. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

A3. The dataset consists of three types of data:

- Neurophysiological data
- Behavior covariates
- Event indications

Q4. **How many instances are there in total (of each type, if appropriate)?**



ID	Task	#Subj.	#Sess.	#Neurons	#Trials	Brain Area
1	Random Target	2	47	18406	25483	M1, S1
2	CO with Bump	2	4	461	2766	Area 2
2	Two-Workspace	3	9	629	4515	
3	Center-Out	4	30	1827	9226	M1 (Subj. 1&4) Area 2 (Subj. 2&3) PMd (Subj. 4)
4	Center-Out	2	23	2194	4712	M1
4	Wrist Isometric CO	1	13	899	2766	
4	Key Grasp	1	9	864	903	
5	Center-Out/Random Target	4	117	11557	22317	M1, PMd
6	Maze	2	9	1728	23117	M1, PMd

A4. Our dataset includes a comprehensive collection of instances across multiple categories:

- 19 subjects
- 261 sessions

These instances are aggregated from six public datasets.

Q5. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

A5. The dataset is not a sample from larger sets; it is a curated collection of six entire datasets. These datasets were chosen for their high quality, data consistency, robust experimental design, and extensive behavioral and neural records, as well as comprehensive metadata. While we have preserved all essential data necessary for machine learning pipelines, we limited the metadata to streamline the datasets. The original datasets were rich in metadata, but we retained only the essential elements. Additional metadata and detailed descriptions can be found on Kaggle (file descriptions).

Q6. **What data does each instance consist of?**

A6. Each instance in the dataset includes neurophysiological data, behavioral information, and event timings, detailed as follows:

Neurophysiological Data:

The columns for neurophysiological data are:

- **NeuronXX** (numeric): Represents single units for all datasets, except for dataset 1, where the first column per session corresponds to multi-units. Although we concatenated all the sessions, Neuron 1 in session 1 does not correspond to Neuron 1 in session 2. Each recorded population per session can be identified by dataset ID, animal, and session.

Behavioral Data:

The columns for behavioral data include:

- **target_dir** (numerical): Direction of the target in radians.
- **target_ID** (numerical): Identification of the target location. For example, in Center-Out tasks, there are 8 possible targets, each represented by an ID. **target_pos_x** and **target_pos_y** (numerical): Cartesian coordinates of the target position.
- **bump_dir** (numerical): Angle (in radians) of bump direction, if there was one. 0 radians is directly to the right, and $\pi/2$ radians is directly upward.
- **maze_num_target** (numerical): Number of targets (for the maze dataset).
- **maze_condition** (numerical): The set of 27 (or 108) mazes included was composed of 3 (or 12) "subsets". Each subset contained 3 related mazes. Each maze had 3 "versions": the 3-target with barrier, the 1-target with barriers, and the 1-target with no barriers. These 3 versions shared the same target positions. The 3-target and 1-target versions also shared the same barrier positions. In the 3-target version, exactly one target was accessible (for the maze dataset).
- **maze_num_barriers** (numerical): Number of barriers in the maze.
- **active_target_pos_x** (numerical): x position on screen of the target hit (for the maze dataset).
- **active_target_pos_y** (numerical): y position on screen of the target hit (for the maze dataset).
- **force_x** and **force_y** (numerical): Interface forces between the hand and the manipulandum handle, in Newtons.
- **hand_pos_x** and **hand_pos_y** or **cursor_pos_x** and **cursor_pos_y** or **finger_pos_x** and **finger_pos_y**(numerical): Velocity of hand, cursor, or finger.
  hand_vel_x and **hand_vel_y** or **cursor_vel_x** and **cursor_vel_y** or **finger_vel_x** and **finger_vel_y** (numerical): hand, cursor or finger velocity.

Events Data:

The columns for events data are:

- **EventTarget_Onset** (boolean): Indicates when the target is presented.
- **EventGo_cue** (boolean): Indicates when the go cue is presented.
- **EventBump** (boolean): Indicates when there is a bump (only for Center-Out with bump task).
- **EventMovement_start** (boolean): Indicates when the subject starts moving.
- **EventMovement_end** (boolean): Indicates when the subject stops moving.

Additional Information:

We provide comprehensive indexes to efficiently filter the data by:

- **datasetID**: Identifier for each dataset (1 to 6)
- **animal**: Identifier for each animal in the dataset
- **session**: Identifier for each session of a particular animal
- **trial_id**: Identifier for each trial within a session performed by an animal from a specific dataset

We also provide indexes to filter data for rewarded trials and task information:

- **result** (categorical): Indicates the trial outcome: Aborted (A), Incomplete (I), Failed (F), Rewarded (R)

  - **task** (categorical): Specifies the task name.

Q7. **Is there a label or target associated with each instance?**

A7. Yes, each instance can have associated labels or targets depending on the purpose of the model. For decoding models, all behavioral data covariates can be used as targets. For forecasting models, the data can be treated as a self-supervised learning task, using only the neurophysiological data.

Q8. **Is any information missing from individual instances?**

A8. There is no missing information from individual instances.

Q9. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

A9. Yes, relationships between individual instances are made explicit. Each row in the dataset corresponds to a specific time point, ensuring that data across different columns and types (neurophysiological, behavioral, and events) are synchronized temporally. This alignment allows for clear and precise analysis of how different data points relate to each other over time.

Q10. **Are there recommended data splits (e.g., training, development/validation, testing)?**

A10. Yes, we chronologically split the data into train, validation, and test sets, with a ratio of 7:1:2.

Q11. **Are there any errors, sources of noise, or redundancies in the dataset?**

A11. Yes, since the data originates from publicly available experiments with animals, there are likely to be sources of errors and noise in the dataset. Experimental variability, biological factors, and environmental influences can all contribute to these imperfections. Our role was to curate and preprocess this data, not to collect it, so these inherent issues may persist.

Q12. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

A12. The dataset is self-contained. No links to external resources.

Q13. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**

A13. There is no confidential data in this dataset.

Q14. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

A14. No.

Q15. **Does the dataset relate to people?**

A15. No.

Q16. **Does the dataset identify any subpopulations (e.g., by age, gender)?**

A16. No.

Q17. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

A17. No.

Q18. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

A18. No.

A.3 Collection Process

Q19. **How was the data associated with each instance acquired?**

A19. The data associated with each instance was acquired from a range of publicly available datasets, all focused on neurophysiological and behavioral experiments with nonhuman primates. These datasets include detailed recordings of neural activity, behavioral responses, and experimental conditions. The sources are as follows:

[1] O'Doherty, J. E., Cardoso, M. M. B., Makin, J. G., Sabes, P. N. (2017). Nonhuman Primate Reaching with Multichannel Sensorimotor Cortex Electrophysiology [Data set]. Zenodo. https://doi.org/10.5281/zenodo.583331

[2] Raeed H Chowdhury Joshua I Glaser Lee E Miller (2020) Dryad Digital Repository Data from: Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. https://doi.org/10.5061/dryad.nk98sf7q7

[3] Gallego-Carracedo, Cecilia et al. (2022). Local field potentials reflect cortical population dynamics in a region-specific and frequency-dependent manner [Dataset]. Dryad. https://doi.org/10.5061/dryad.xd2547dkt

[4] Ma, Xuan et al. (2023). Data from: Using adversarial networks to extend brain computer interface decoding accuracy over time [Dataset]. Dryad. https://doi.org/10.5061/dryad.cvdncjt7n

[5] Perich, Matthew G.; Miller, Lee E.; Azabou, Mehdi; Dyer, Eva L. (2024) Long-term recordings of motor and premotor cortical spiking activity during reaching in monkeys (Version draft) [Data set]. DANDI archive. https://doi.org/10.80507/dandi.123456/0.123456.1234

[6] Churchland, Mark; Cunningham, John P.; Kaufman, Matthew T.; Foster, Justin D.; Nuyujukian, Paul; Ryu, Stephen I.; Shenoy, Krishna V. (2024) Neural population dynamics during reaching (Version draft) [Data set]. DANDI archive. https://dandiarchive.org/dandiset/000070/draft

These datasets were chosen for their high quality, extensive metadata, and the depth of information they provide, making them ideal for various neural decoding and prediction model evaluations.

Q20. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

A20. Details about the data collection mechanisms and procedures can be found in the original papers cited earlier. These papers provide comprehensive descriptions of the hardware apparatus and other procedures used to collect the data.

Q21. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

A21. The dataset is not sampled; it comprises the entirety of the available data. Therefore, there is no specific sampling strategy involved.

Q22. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

A22. The data collection process involved no direct participation or compensation, as the dataset consists of publicly available data.

Q23. **Over what timeframe was the data collected?**

A23. The data were collected in 2024 over a time period spanning six months.

Q24. **Were any ethical review processes conducted (e.g., by an institutional review board)?**

A24. No, such processes are unnecessary in our case.

A.4 Preprocessing/cleaning/labeling

Q25. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

A25. Yes, several preprocessing steps were performed on the data:

- Velocity Computation: In datasets lacking velocity information, velocity was computed based on position data. This step ensured consistency and completeness of the behavioral data.
- Neural Spiking Data Transformation: The neural spiking data, often provided in spike times, was transformed into spike counts with a bin size matching the behavioral data's time resolution. This transformation facilitated analysis and modeling by aligning neural activity with behavioral events.
- Event Alignment: All events assigned by experimentalists were matched to the final dataset, ensuring temporal alignment with other data. This step helped consolidate all relevant information into a single coherent dataset, facilitating subsequent analysis.

These preprocessing steps helped ensure data consistency, completeness, and alignment, making the dataset ready for analysis and modeling.

Additionally, we retained only the attributes essential for machine learning algorithms. Information related to electrode names, waveforms, and other such details available in the original datasets is not included in the current dataset.

Q26. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

A26. Yes, the raw data is available in the original repositories where the data was sourced from. This ensures that the original, unprocessed data is preserved and accessible for any future analyses or unanticipated uses.

Q27. **Is the software used to preprocess/clean/label the instances available?**

A27. No.

A.5 Uses

Q28. **Has the dataset been used for any tasks already?**

A28. The dataset was used only to generate the results available in the paper.

Q29. **Is there a repository that links to any or all papers or systems that use the dataset?**

A29. There are still no applications of the presented datasets. We intend to keep track of its uses in the project GitHub repo.

Q30. **What (other) tasks could the dataset be used for?**

A30. The dataset can be used to train encoding and decoding models.

Q31. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

A31. We believe that our dataset will not encounter usage limit.

Q32. **Are there tasks for which the dataset should not be used?**

A32. No, users could use our dataset in any task as long as it does not violate laws.

A.6 Distribution

Q33. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

A33. Yes, the dataset is publicly accessible.

Q34. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**

A34. It will be distributed on Kaggle and Dandi.

Q35. **When will the dataset be distributed?**

A35. The dataset is publicly available as of today on Kaggle and Dandi. There are no plans in removing the dataset from public usage.

Q36. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

A36. The dataset is licensed under the Creative Commons CC BY-NC-ND 4.0 license.

Q37. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

A37. No.

Q38. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

A38. No.

A.7 Maintenance

Q39. **Who is supporting/hosting/maintaining the dataset?**

A39. The authors of the paper.

Q40. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

A40. Please contact this email address: carolina.filipe@research.fchampalimaud.org

Q41. **Is there an erratum?**

A41. No, there is no erratum as of yet. If necessary in the future, an erratum will be developed for the dataset, as well as for this document.

Q42. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

A42. There are no current plans on updating the current datasets. This can change in the future, either to introduce new variants to the dataset, or to correct any undetected bug.

Q43. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

A43. There are no applicable retention limits of the data.

Q44. **Will older versions of the dataset continue to be supported/hosted/maintained?**

A44. If any updates are published, previous versions will be available.

Q45. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

A45. There are no current mechanisms to contribute to the dataset. Novel ideas and variants of the dataset should be submitted via email to the authors or as an issue on GitHub.

Author Statement

The authors confirm the data in the NeuroTask is under the CC BY-NC 4.0 International License. The authors bear responsibility in case of violation of copyrights.