Contents

Intention to submit document for the Work with US researchers BBSRC-NSF/BIO lead agency 2024 funding opportunity

Enabling Naturalistic, Long-Duration and Continual Neuroscience Experimentation with Advanced Machine Learning

October 17, 2024

1 Summary

Word limit: 2 A summary is not required for this section, please write 'N/A' in the textbox. Please still include a title for your project.

N/A

2 Core team

List the key members of your team and assign them roles from the following:

- project lead (PL)
- project co-lead (UK) (PcL)
- specialist
- professional enabling staff
- research and innovation associate
- technician
- researcher co-lead (RcL)

Only list one individual as project lead.

The core team section must only contain details of the UK applicants. The US applicant information should be listed in the 'US applicants' section.

Find out more about UKRI's core team roles in funding applications.

project lead (PL) Prof. Maneesh Sahani

project co-lead (UK) (PcL) Prof. Tiago Branco, Prof. Thomas Mrsic-Flogel

researcher co-lead (UK) (RcL) Dr. Joaquin Rapela, Dr. Dario Campagner

3 Application questions

3.1 Research theme

Word limit: 5 Please state the research theme you are applying under. Choose one of the following research themes:

1. biological informatics
2. understanding host-microbe interactions
3. synthetic cells and cellular systems
4. synthetic microbial communities

biological informatics

3.2 Vision

Word limit: 500

What are you hoping to achieve with your proposed work?

What the assessors are looking for in your response

Your vision should clearly address:

- one of the opportunity research themes (biological informatics, understanding host-microbe interactions, synthetic cells and cellular systems or synthetic microbial communities)
- the remit of the BBSRC and the NSF/BIO division associated with your chosen research theme

References may be included within this section, but this will count towards your word count.

Images are not required for this section.

3.2.1 Context

Conventional systems neuroscience experiments are typically short in duration and often place significant constraints on subject behavior to simplify data analysis. However, these restrictions may limit our ability to observe critical aspects of brain function and behavior that only manifest in more naturalistic and extended conditions.

At the Sainsbury Wellcome Centre (SWC) for Neural Circuits and Behaviour, we are pioneering Naturalistic, Long-Duration, and Continual (NaLoDuCo) foraging experiments in mice that span weeks to months. During these extended experiments, we collect high-resolution recordings of both behavioral and neural activity in naturalistic settings. In collaboration with the Gatsby Computational Neuroscience Unit (GCNU), we are developing novel analytical methods to interpret this new class of data.

This novel experimental approach will enable researchers to explore neural mechanisms underlying behavior over extended periods for the first time, offering the possibility of uncovering insights across a wide range of phenomena, including long-term behavioral adaptation, neural plasticity, and learning. The data generated from NaLoDuCo experiments represent an entirely new resource in neuroscience, with the potential to drive breakthroughs and discoveries that are beyond the reach of traditional experiments.

Our vision is to empower research centers worldwide to adopt this groundbreaking approach. However, the scale and complexity of the data generated pose significant challenges in data acquisition, visualisation, and analysis. In this proposal, we will address these challenges, developing and sharing openly the necessary expertise, hardware, and software to enable this transformative type of experimentation on a global scale.

3.2.2 Focus areas

Below, we outline the key focus areas we aim to address (Figure 3), along with their associated challenges. These challenges primarily revolve around the collection and analysis of continuously recorded, extremely large datasets–on the order of hundreds of terabytes–gathered from experiments spanning weeks to months.

While experiments in neuroscience that are naturalistic, long-duration, or continuous have been conducted in the past (e.g., Jhuang et al., 2010; Mao et al., 2021; Voloh et al., 2023), to the best of our knowledge, we are the first to integrate all three of these features in a single experimental paradigm. This combination introduces unprecedented complexities in data processing, as we aim to capture behavior and brain activity in their most ecologically valid, extended, and uninterrupted forms.

The focus areas of the proposed project are (Figure 3):

Data Collection & Management Efficiently gathering and organizing massive datasets over extended periods.

Data Sharing Providing easy access to large-scale datasets to researchers around the globe using cloud-based technologies.

Data Visualization Developing efficient web-based tools to visualize very large behavioral and neural datasets.

Spike Sorting Assigning spikes to neurons reliably, and tracking individual neurons across long-periods of time in real time.

Data Analysis Evaluating existing methods, and developing new ones, when necessary, to address key problems in behavioral and neural data analysis (Figure 2).

Inference-Driven Experimentation Creating a new type of experimentation driven by real-time behavioral and neural inferences.

These focus areas represent key technical and analytical challenges that, once addressed, will faciliate a transformative shift in neuroscience research.

3.3 Approach

Word limit: 500

How are you going to deliver your proposed work?

What the assessors are looking for in your response

Your approach should give an overview highlighting:

- a clear description of the objectives and methodology for the proposed work, including the contributions of the UK and US teams
- the potential outputs and outcomes of the proposed work

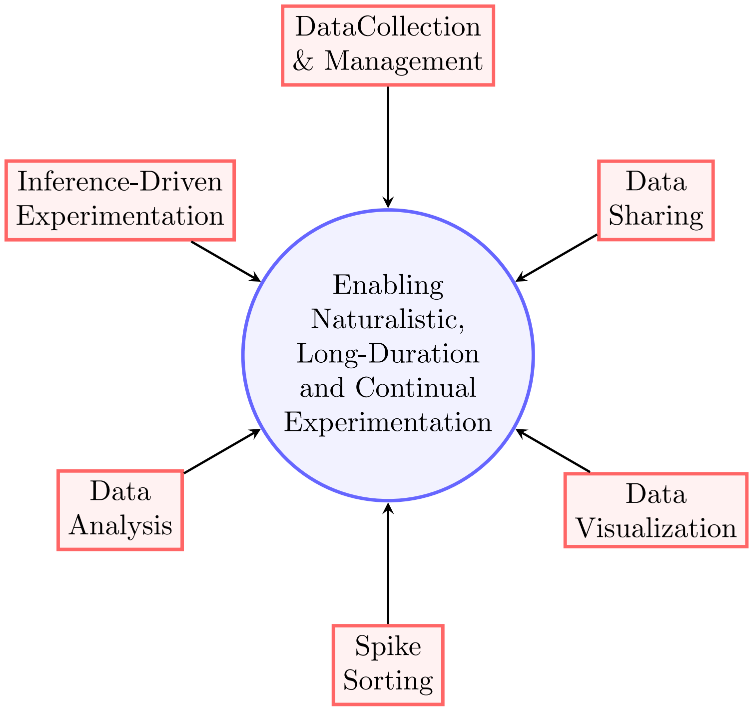
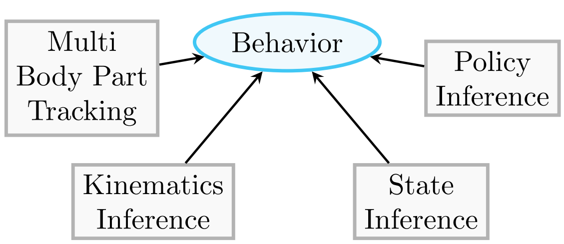
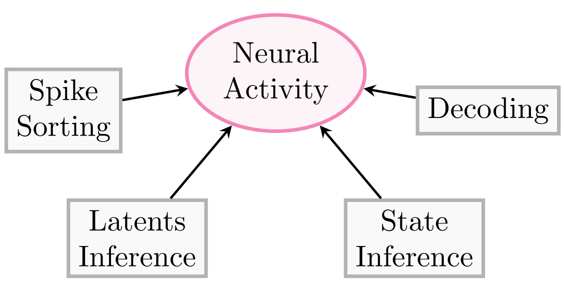


Figure 1: Project theme (blue) and focus areas (red).



(a)



(b)

Figure 2: Behavioral (a) and neural (b) data analysis problems to address.

References may be included within this section, but this will count towards your word count.

Images are not required for this section.

3.4 US applicants

Word limit: 200

Please provide the following details of the US applicants on this application:

1. name
2. institute
3. job title
4. role in project (for example, project lead or project co-lead)
5. email address

Please also indicate who the lead US applicant will be.

NSF will use this information to confirm applicant eligibility.

Please do not include details of US applicants in the 'Core team' section.

3.5 Resources

Word limit: 200

Please provide the following:

- overall estimates for costings and staffing full time equivalent (FTE) for both the UK and US components
- clear separation of UK and US costings, in pounds sterling and US dollars (USD) respectively

The overall budget should be below the maximum £2 million combined funder contribution

If there is more than one UK or US team associated with the application, please combine their estimates together.

A detailed calculation and breakdown of resources is not required at this stage, nor is a justification of costs.

The following is an example of how this might look.

UK Resources:

Total cost estimate: £600,000

Research council contribution: £480,000

0.2 FTE time, 1.0 FTE PDRA, 0.5 FTE technician

US Resources:

Total cost estimate: $300,000

1.0 FTE PDRA or 1.0 FTE doctoral researcher

Total funder contribution estimate:

£716,475 (£480,000 + £236,475 ($300,000 at exchange rate 0.79))

A More details about the vision

A.1 Context

Conventional systems neuroscience experiments are typically short in duration and often place significant constraints on subject behavior to simplify data analysis. However, these restrictions may limit our ability to observe critical aspects of brain function and behavior that only manifest in more naturalistic and extended conditions.

At the Sainsbury Wellcome Centre (SWC) for Neural Circuits and Behaviour, we are pioneering Naturalistic, Long-Duration, and Continual (NaLoDuCo) foraging experiments in mice that span weeks to months. During these extended experiments, we collect high-resolution recordings of both behavioral and neural activity in naturalistic settings. In collaboration with the Gatsby Computational Neuroscience Unit (GCNU), we are developing novel analytical methods to interpret this new class of data.

This novel experimental approach will enable researchers to explore neural mechanisms underlying behavior over extended periods for the first time, offering the possibility of uncovering insights across a wide range of phenomena, including long-term behavioral adaptation, neural plasticity, and learning. The data generated from NaLoDuCo experiments represent an entirely new resource in neuroscience, with the potential to drive breakthroughs and discoveries that are beyond the reach of traditional experiments.

Our vision is to empower research centers worldwide to adopt this groundbreaking approach. However, the scale and complexity of the data generated pose significant challenges in data acquisition, visualisation, and analysis. In this proposal, we will address these challenges, developing and sharing openly the necessary expertise, hardware, and software to enable this transformative type of experimentation on a global scale.

A.2 Focus areas and their challenges

Below, we outline the key focus areas we aim to address (Figure 3), along with their associated challenges. These challenges primarily revolve around the collection and analysis of continuously recorded, extremely large datasets–on the order of hundreds of terabytes–gathered from experiments spanning weeks to months.

While experiments in neuroscience that are naturalistic, long-duration, or continuous have been conducted in the past (e.g., Jhuang et al., 2010; Mao et al., 2021; Voloh et al., 2023), to the best of our knowledge, we are the first to integrate all three of these features in a single experimental paradigm. This combination introduces unprecedented complexities in data processing, as we aim to capture behavior and brain activity in their most ecologically valid, extended, and uninterrupted forms.

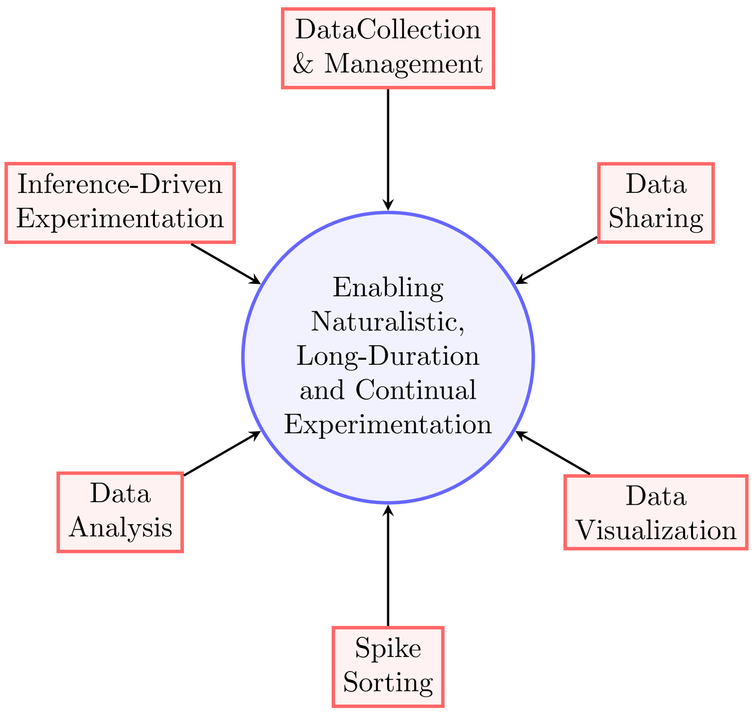


Figure 3: Project theme (blue) and focus areas (red).

A.2.1 Data acquisition and management

At the SWC we have already performed foraging experiments in mice continuously collecting behavioral and experimental data 24 hours a day for seven days. We will share openly the specifications of the hardware used to build these experiments (e.g., instructions for building large foraging arenas, video cameras specifications, electrophysiological recording hardware), as well as the software we used for experimental control, data quality control, data access and management.

The data acquisition and management software used in our project is already publically available in GitHub[1]. This software is already being used by scientists at the Allen Institue for Neural Dynamics and at Northwester University. We will substantially improve its documentation to simplify its usage by external users.

Challenges related to data acquisition and management include data indexing to allow fast access to very large amount of saved data, online quality control and alert systems to guarantee that anomalities in data collection are detected and corrected with minimal delay, and syncrhonization between multiple data streams.

A.2.2 Data dissemination

Datasets of the scale of hundreads of terabytes cannot be practically downloaded from data repositories. This is specially true for contiguous experiments where unique insights are extracted by characterizing full datasets, and not only parts of them. Therefore, we will store data in DANDI, which uses Amazon S3 buckets, and provide software in Amazon EC2 instances to visualize and analyze data on the cloud, avoiding costly data transfers. That is, the large dataset sizes of NaLoDuCo experiments make it impractical to distribute data to users and require to bring users to data. Fortunately, cloud technologies are now mature to allows this.

Importantly, if we distributed these very large datasets to users, only those in large research centers would have the computing power to process them. But, by deploying data and computing in the cloud, any person with Internet access around the world will be able to benefit from them. Storing large datasets in DANDI is free.

Dr. Ben Ditcher, founder of CatalystNeuro, has played a pivotal role in supporting the development and operations of the DANDI archive.

A.2.3 Data visualisation

Visualisations are essential for scientific discovery. For the proposed project visualisation present two major challenges. First, they need to display very large datasets at different temporal scales, from milliseconds to weeks and months. Second, as data and software will be deployed in the cloud, visualisation need

[1] https://github.com/SainsburyWellcomeCentre/aeon_mecha

to be web based. Standard visualization tools cannot display terabyte sized datasets. We will build custom web-based visualization tools to do this.

We have substantial experience building web-based visualization tools for neurophysiological data. Dr. Jeremy Magland is now developing Neurosift[2] a web-based visualizer for DANDI datasets.

A.2.4 Spike sorting

When electrodes are placed in the brain, they typically record spikes from multiple nearby neurons. Spike sorting attributes spikes to individual neurons.

Spike sorting is specially challenging for NaLoDuCo experiments. First, because these experiments require to track individual neurons of freely moving mice for weeks to months. Second, because spike sorting needs to be done online, to allow experiments driven by real-time machine learning inference, as described below.

Prof. Sahani pioneered the use of Bayesian inference methods for spike sorting (Sahani, 1999). Dr. Jeremy Magland has significantly advanced the field of spike sorting, particularly through his development of MountainSort[3] and his contributions to SpikeInterface[4].

A.2.5 Data analysis

Advanced data analysis methods are indispensable to extract meaning from NaLoDuCo experimental data. However, analyzing this data is challenging for at least three reasons. First, important insights will most probably come from the characterization of complete datasets, and not form subsets extracted from them. Conventional batch methods cannot be used with datasets of the size produced by NaLoDuCo experiments. For instance, for learning, batch linear regression cannot load into memory and invert a data matrix with high-resolution observations from a one-month-long experiment. Thus, **online methods** that can process infinite data steams become mandatory.

Second, a pervasive assumption in most ML algorithms is stationarity; i.e., the assumption that the statistics of data do not change over time. But in long-duration and continuous experiments this assumption is most often violated as, for example, the arousal of subjects changes. Hence, the analysis of data generated by these experiments requires **adaptive methods**.

Third, statistical algorithms consist of two key stages: learning (or trainning) and inference (or prediction). The learning stage identifies model parameters, and the inference stage uses the learned model to make predictions, or infer latent variables, from new unseen data. Frequently training is performed on a small subset of a dataset, and inference is done on the remaining data. However, since in long-duration and continual experiments behavior and neural activity are generall not stationary, it is not optimal to train models on data subsets and

[2] https://github.com/flatironinstitute/neurosift
[3] https://github.com/flatironinstitute/mountainsort5
[4] https://github.com/spikeinterface/spikeinterface

use them to make inferences on the remaining data, since the state of the animal at training and inference times may be different. To overcome this difficulty we will use **continual learning methods**.

We will evaluate methods to analyze different aspects of behavior and neural activity (Figure **??**). We will test how these methods process very large datasets, how they handle non-stationary data, and how feasible is to retrain them to adapt to changing conditions. We will adapt these methods so that they better address these challenges and, when needed, develop new ones. We will carefully report the outcomes of these evaluations so that researchers performing NaLoDuCo experimentation can choose the best methods that suit their needs.

A.2.6 Experiments driven by real-time machine learning inference

Small animal experiments are usually controlled by simple static rules or direct behavioral observations. Funded by a BBSRC award[5] we are developing software to allow a new type of experimental control based on statistical inferences made on behavioral and/or neural measurements.

For example, after inferring latent variables from neural activity and observing that one of these latents have crossed a threshold, we can deliver a reward (as done in learning to control a BCI; Clancy and Mrsic-Flogel, 2021), or perform an action (as done in motor imagery BCI; Lebedev and Nicolelis, 2006), or manipulate of neural activity (as done when studying the causal relation between a pattern of brain activity and behavior; Deisseroth, 2015). We propose to further develop the previous software and use it to test causal effects of neural activity patterns on foraging decisions using our NaLoDuCo foraging experiments.

Buidling experiments driven by real-time machine learning inferences brings at least two challenges. The first one is a machine learning problem, how to build fast inferences that can operate in real time. The second one is a neuroscience problem, how to identify neuroscience experiments suitable to real-time control, and then perform the experiment with real-time control. Fortunately at the Gatsby Unit we are experienced on building advanced machine learning algorithms to address the first challenge. And at the SWC we perform many sophisticated animal experiments that could benefit from real-time experimental control.

In summary, we are pioneering a new paradigm in neuroscience experimentation, driven by advanced inferential methods applied to rich behavioral and neural recordings. This innovative technology has the potential to transform the field, enabling experiments that were previously unimaginable. By leveraging these sophisticated inferences, we may unlock new dimensions of knowledge that could not be achieved through simpler, conventional approaches. This breakthrough could open doors to insights that redefine our understanding of brain-behavior relationships.

[5] https://gow.bbsrc.ukri.org/grants/AwardDetails.aspx?FundingReference=BB%2FW019132%2F1

References

Clancy, K. B. and Mrsic-Flogel, T. D. (2021). The sensory representation of causally controlled objects. *Neuron*, 109(4):677–689.

Deisseroth, K. (2015). Optogenetics: 10 years of microbial opsins in neuroscience. *Nature neuroscience*, 18(9):1213–1225.

Jhuang, H., Garrote, E., Yu, X., Khilnani, V., Poggio, T., Steele, A. D., and Serre, T. (2010). Automated home-cage behavioural phenotyping of mice. *Nature communications*, 1(1):68.

Lebedev, M. A. and Nicolelis, M. A. (2006). Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546.

Mao, D., Avila, E., Caziot, B., Laurens, J., Dickman, J. D., and Angelaki, D. E. (2021). Spatial modulation of hippocampal activity in freely moving macaques. *Neuron*, 109(21):3521–3534.

Sahani, M. (1999). *Latent variable models for neural data analysis*. California Institute of Technology.

Voloh, B., Maisson, D. J.-N., Cervera, R. L., Conover, I., Zambre, M., Hayden, B., and Zimmermann, J. (2023). Hierarchical action encoding in prefrontal cortex of freely moving macaques. *Cell reports*, 42(9).