Mesmerize: a dynamically adaptable user-friendly analysis platform for 2D & 3D calcium imaging data.

Kushal Kolar[*], Daniel Dondorp, Jordi Cornelis Zwiggelaar[#], Jørgen Høyer[#], Marios Chatzigeorgiou[*]

Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgt. 55, 5006 Bergen, Norway

*email: kushalkolar@gmail.com, Marios.Chatzigeorgiou@uib.no

# contributed equally

Abstract

Calcium imaging is an increasingly valuable technique for understanding neural circuits, neuroethology, and cellular mechanisms. The analysis of calcium imaging data presents challenges in image processing, data organization, analysis, and accessibility. Tools have been created to address these problems independently, however a comprehensive user-friendly package does not exist. Here we present "Mesmerize", an efficient, expandable and user-friendly analysis platform, which uses a Findable, Accessible, Interoperable and Reproducible (FAIR) system to encapsulate the entire analysis process, from raw data to interactive visualizations for publication. Mesmerize provides a user-friendly graphical interface to state-of-the-art analysis methods for signal extraction & downstream analysis. We demonstrate the broad scientific scope of Mesmerize's applications by analyzing neuronal datasets from mouse and a volumetric zebrafish dataset. We also applied contemporary time-series analysis techniques to analyze a novel dataset comprising neuronal, epidermal, and migratory mesenchymal cells of the protochordate *Ciona intestinalis*.

Introduction

Large-scale calcium imaging of neuronal activity in populated brain regions, or entire animals, has become an indispensable technique in neuroscience research. The analysis of calcium imaging datasets presents significant challenges in the domains of image preprocessing, signal extraction, dataset organization, downstream analysis, and visualizations. As a result, analysis of calcium imaging data requires computational expertise that are rather uncustomary among biologists. Numerous state of the art packages, such as the Caiman library[1], Suite2p[2],SIMA[3], EZCalcium[4] and ImageJ[5] provide users with a myriad of options for image pre-processing and ROI/signal extraction. Workflow management tools for neurophysiological analysis, such as DataJoint[6] and NWB[7], provide programmers with tools for dataset organization. Users with computational training often incorporate these tools using custom written scripts or spreadsheets. In contrast, biomedical scientists with little or no programming experience would immensely benefit from a user-friendly platform to organize, analyze, visualize, and share 2D and 3D calcium imaging data.

An important attribute of such a platform would be the ability to seamlessly incorporate cutting edge tools that will readily address current and future technical challenges. The immense growth we have seen over the last decade in new imaging technologies combined with the ever-increasing palette of genetically encoded indicators have fueled an increase in the temporal and spatial resolution of the acquired data sets. Calcium imaging is not only a workhorse technique for monitoring brain-wide activity, but it is becoming increasingly popular in the dissection of developmental and physiological processes at the level of entire embryos or organs. These types of information rich datasets are characterized by the presence of large populations of morphologically and functionally diverse, tightly packed, cells that exhibit diverse activity profiles, making downstream processing challenging. In particular, the analysis of 2D and 3D calcium imaging datasets poses significant technical hurdles across multiple domains including those of image preprocessing, signal extraction, dataset organization, downstream analysis, and visualization.

One of the greatest challenges that modern biomedical research faces is compliance with FAIR data (Findable, Accessible, Interoperable and Reusable) principles, which aim to set new and robust standards in terms of reproducibility and data sharing. However, even some of the most advanced analyses pipelines rely on custom written scripts and spreadsheets, without a standardized system to organize and functionally link raw imaging data, analysis procedures and visualizations[8,9]. This greatly impedes the reproducibility of the work even when the raw data are available[8–10]. State of the art project management tools, such as OMERO[11], Biaflows[12], Cytomine[13], OpenBIS[14] and KNIME[15]are geared towards cell biology and histological analysis, and are not suited for neurophysiological or calcium imaging analysis (Table 1). Most crucially, none of these tools support the rich and comprehensive annotations necessary for most experiments in the field of neuroscience. For example, the analysis of neurophysiological experiments often requires temporal mapping of complex combinations of stimuli and behavioral annotations that directly correspond to the imaging data (Table 1). There are also experimental scenarios where the cells or regions of interest (ROIs) additionally require a combination of annotation tags (text/numerical labels) describing features such as the cell type, morphology, or identity, which can be mapped back to the corresponding cell(s) or ROI(s). Finally, for publication, authors have to produce figures integrating all of the above (i.e. the calcium imaging data, the annotations and the downstream analysis) to effectively and coherently convey the biological findings. While there are many tools for producing basic static visualizations, there is an urgent need for a software platform that can produce *interactive* visualizations where the imaging data and analysis history of every datapoint can be instantly retrieved[8,9,16]. Interactive and traceable visualizations have various applications, such as quality control[8], reproducibility[9,16,17], and allowing for a better understanding of experiments and underlying the biology[8].

From the examination of the tools currently available for calcium imaging analysis and bio-imaging project management (Table 1), we demonstrate that there is currently no tool that provides a comprehensive suite of features necessary for calcium imaging analysis and project management, i.e. image processing, ROI extraction, project organization, downstream analysis and interactive visualizations. To address these challenges, we created Mesmerize – a free and open source comprehensive platform that encapsulates these requirements within a reproducible system. The Mesmerize platform also provides graphical user interfaces (GUI) for the analysis and visualization of 2D and 3D datasets, thereby allowing biomedical scientists to create FAIR (Findable, Accessible, Interoperable and Reusable) datasets[10,18] within a flexible system that can be adopted by a wide variety of researchers who work on diverse biological problems. Mesmerize is not a pipeline, but rather a highly modular platform that presents users with many options along each step of their specific user-defined calcium imaging analysis workflow. Consequently, this flexible design allows developers to easily add new or customized modules for image processing, analysis, and visualization. In summary, the ability to create modular and adaptable workflows grants Mesmerize a very broad scope of applicability across a variety of labs in various fields of neuroscience. For example, it may be used to study whole-brain dynamics, sensory-motor integration systems, or activity defects in disease models. Beyond neuroscience, Mesmerize has the potential to be transformative in the hands of developmental biologists and physiologists interested in mapping embryonic and post-embryonic calcium dynamics of specific tissues/organs or entire embryos. Mesmerize lets users create and dynamically curate an unlimited number of categorical labels that map to entire imaging sessions, single ROIs, and temporal periods. This rich and complex annotation capability goes beyond standard neurobiological annotations such as behavioral correlates or sensory stimuli and can be extended to developmental stages, shared gene expression patterns, morphological and phenotypic cell type descriptors, and subcellular compartments to a name a few. This flexibility means that Mesmerize is broadly suitable for cell biologists, developmental biologists and other specialties beyond neuroscience. In scenarios where the

114 analysis workflows require further tailoring, Mesmerize can serve as a blueprint for future platforms
115 that seek to encapsulate data analysis, project organization and interactive traceable visualizations in
116 other fields.

117 As introduced above, calcium imaging analysis usually requires the following components 1)
118 pre-processing & ROI/signal extraction 2) data annotation and organization 3) downstream analysis
119 and 4) visualization. Mesmerize provides end-users with extensive graphical interfaces for each of
120 these components to analyze their 2D and 3D datasets. Users with basic Python or scripting skills can
121 utilize the API to implement more customized or complex analysis. We have built the graphical
122 interfaces using the Qt framework due to its maturity and extensive developer community. All data
123 structures are well-documented and built using pandas DataFrames[19] and numpy arrays[20,21], both
124 highly prevalent and mature libraries. These features make Mesmerize a highly accessible platform,
125 allowing users to easily integrate Mesmerize into their analysis workflows, or develop new customized
126 modules.

127 **Mesmerize Platform**

128 ***Rich Data Annotation***

129 The first step of any calcium imaging analysis workflow requires a system for users to explore
130 their imaging data and perform ROI extraction. We demonstrate that Mesmerize works with both 2D
131 and 3D datasets from a broad set of model organisms, such as mice, zebrafish, and *Ciona intestinalis*
132 (Fig 1a). These datasets can be visualized using the Mesmerize Viewer, which provides GUI front-ends
133 (based on pyqtgraph) and API interfaces for various signal extraction modules (Fig 1b). Importantly,
134 the Viewer also facilitates extensive *in-place* annotation of experimental information (Fig 1c-e), such
135 as:

136 1. Cell identities, morphology, or any other tags that map to individual cells/ROIs (Fig 1e)
137 2. Temporal mapping, such as stimulus or behavioral periods (Fig 1d)
138 3. Data that map to entire recordings, such as an animal's genotype, age, strain etc. (not shown)

139 These annotations may be performed through the GUI, or automated through the simple scripting
140 interface. Mesmerize's unique support for customizable annotations makes it broadly applicable for
141 diverse range of researchers and distinguishes it from other calcium imaging and image analysis tools
142 (Table 1). The highly versatile annotation functions within Mesmerize enable scientists to efficiently
143 curate and analyze complex datasets that are emerging from the use of multiplexed imaging combining
144 several cell-specific promoters that express Genetically Encoded Calcium Indicators (GECIs). For
145 example, researchers can perform a cohort of experiments that utilize tens of GCaMP promoters,
146 multiple combinations of optogenetic and/or chemogenetic lines, multiple UAS-GAL4 systems,
147 multiple drugs etc. in one efficient, organized and reproducible system. To illustrate this capacity of
148 Mesmerize, we leverage a powerful emerging model organism, the protochordate *Ciona intestinalis*.
149 The Ciona dataset analyzed here includes annotations for seven different GCaMP6s promoters, eight
150 anatomical regions, and twenty-one cell types (Supplementary table 1 & 2).

151 ***ROI Extraction***

152 Graphical front-ends help users explore imaging data, perform pre-processing, and signal
153 extraction. They help facilitate efficient workflows for advanced users, and are necessary for users
154 without extensive programming experience. From a user's perspective these front-ends, which we call
155 *Viewer Modules*, interact with the Mesmerize Viewer in a manner similar to the various components
156 within ImageJ and its plugins. This familiarity in the user-end design will allow Mesmerize to be easily
157 adopted by more biologists, and broaden the reach of cutting-edge packages, (such as the CaImAn
158 library[1]) allowing them to perform more accurate and in-depth analysis.

159 By default, *Viewer Modules* are provided for NoRMCorr[22], CNMF(E)[23–25], NuSeT[26], as well as
160 importers for Suite2p[2] outputs and ImageJ[5] ROIs (Fig 1b). These front-ends encompass a very broad
161 variety of user-options for motion correction and signal extraction from both 2D and 3D calcium
162 imaging datasets. Many Viewer Modules are used in conjunction with the Mesmerize Batch Manager

163 which streamlines the exploration of parameter space and data organization for these computationally
164 intensive tasks.

165 ROI extraction and image processing are not limited to the default options that we provide,
166 these Viewer Modules can be expanded, customized and created by users with modest programming
167 experience. We provide an API and scripting interfaces, which allows ROIs to be extracted from any
168 other custom technique which the user may desire. This flexibility allows scientists to conveniently
169 integrate and combine their favorite pre-processing or ROI extraction technique into their analysis
170 workflow. For example we created a simple API[27] to a deep-learning approach for cellular
171 segmentation using the NuSeT[26] network, which is useful for the segmentation of recordings using
172 nuclear-localized GCaMP. The NuSeT method can be used through a GUI that can be expanded to
173 include additional deep-learning segmentation approaches from this rapidly evolving field in the future.
174 Furthermore, the binary masks produced by the NuSeT Viewer Module can be used for seeding
175 CNMF(E)[23,25], thereby allowing these two cutting-edge tools to be combined in manner that would be
176 non-trivial for users without extensive programming experience. In summary, these features
177 demonstrate how Mesmerize can be a powerful platform for complex integration and interoperability
178 between multiple state of the art analysis tools for both end-users and developers.

179 ***Project Organization***

180 Current software platforms for bio-image dataset organization are not suited for handling
181 calcium imaging data (Table 1). Mesmerize packages all data associated with an imaging sample, i.e.
182 extracted signals, annotations etc, into a *Project Sample* (Fig 1f). A collection of *Project Samples*
183 constitute a *Project Dataset*, which can be explored and filtered in a user-friendly manner to create
184 experimental groups using the *Project Browser* (Fig 1g). Project Samples can be modified throughout
185 the course of a project. Therefore, in addition to efficient data annotation, users can append, change or
186 supplement existing annotations that can then be propagated through downstream analysis and

visualizations. Dynamically adaptable data management is extremely useful since biological questions and experiments are often in constant flux as new data are processed and analyzed.

Downstream Analysis

A *Project Dataset*, or sub-dataset, can be loaded into a flowchart where users can build analysis pipelines by connecting analysis nodes (Fig 1h-j). We provide nodes to perform many common signal processing routines, data handling/organization, dimensionality reduction, and clustering analysis. Mesmerize's default collection of nodes allows users to perform many common analysis procedures such as comparison of stimulus/behavioral periods (Fig 1h), peak detection (Fig 1i), and clustering analysis (Fig 1j). All analyses performed in the flowchart are logged with a description of the nodes and their parameters, thereby facilitating future reproducibility of the analyses. For more customized analysis, we provide documentation and an API for efficiently writing new analysis nodes or using the analysis data structures in external notebooks or scripts (http://docs.mesmerizelab.org/en/master/developer_guide/nodes.html). The flowchart builds upon a pyqtgraph[28] widget. The stock assortment of nodes implement various signal processing, dimensionality reduction, and clustering analysis using scipy[29], sklearn[30] and tslearn[31] libraries. We use common and mature libraries to simplify customization by more advanced users or developers.

Visualization

The ultimate result of almost any analysis procedure and scientific study is the creation of visualizations that convey an experiment's results. The vast majority of visualizations in most research are static. This makes it difficult or impossible to instantly link datapoints from a plot with the original imaging data and analysis procedures[8,9,16], which greatly hampers reproducibility[16]. Recent developments help address these issues; tools such as Jupyter[32] notebooks delivered via MyBinder[33] allow the data and analysis procedures to be shared. However, these methods are not readily accessible to non-programmers and do not aid in the creation of FAIR and functionally linked datasets. Mesmerize allows users to create interactive visualizations through a GUI and share them in their

212 interactive state (Fig 1k). Many interactive plots are attached to a *Datapoint Tracer* (Fig 1l) which
213 highlights the spatial localization of the selected datapoint and displays all its associated annotations
214 and the analysis history log which can be visualized using an analysis graph (Fig 1m), a graphical
215 visualization that intuitively communicates the analysis steps. A rich variety of built-in plots are
216 provided, such as heatmaps, spacemaps, scatterplots, beeswarm, and more. As with other components
217 of the Mesmerize platform, we provide developer instructions for the creation of new plots that can
218 integrate with the *Datapoint Tracer*
219 (http://docs.mesmerizelab.org/en/master/developer_guide/plots.html). Thus far, no other calcium
220 imaging analysis suite offers such a rich variety of interactive visualizations for downstream analysis
221 (Table 1). Lastly, we are currently creating a set of standardized web-based visualizations that mirror
222 the current options available for matplotlib[34] and pyqtgraph[28] based plots in Mesmerize. This will
223 further improve the shareability of data since a user will be able to interactively explore visualizations
224 from a Mesmerize dataset without installing anything on their end.

225 ***Shareable Datasets***

226 In summary, Mesmerize is the first platform to address common difficulties with
227 reproducibility, data reusability, and organization in calcium imaging data analysis by comprehensively
228 encapsulating image analysis, data annotation, analysis, and interactive visualizations. Mesmerize
229 allows analysis procedures and annotations to be transparent at the level of the individual datapoints in
230 a plot. This is achieved by tagging Universally Unique Identifiers (UUID) to the data at various layers
231 of analysis, a key principle for the creation of a FAIR dataset. Mesmerize's unique capacity for the
232 robust maintenance of rich and complex annotations encourages users to exhaustively describe their
233 datasets. A Mesmerize project is entirely self-contained within a single directory tree, making it easy to
234 share entire datasets, analysis workflows, and interactive visualizations with the scientific community.
235 Another scientist can open a Mesmerize project and immediately explore visualizations, analysis
236 procedures, and view the raw data associated with the datapoints on a published figure. This ease of

237 opening a Mesmerize project and exploring datasets in conjunction with interactive visualizations will
238 help scientists in making their data easily accessible and reusable.

239 Lastly, in order to reach a broad range of users, Mesmerize is cross-platform and works on
240 Linux, Mac OSX and Windows. Mesmerize is free, open source, uses the GNU General Public License
241 v3.0 and is hosted on GitHub. In facilitate fast and easy installation on all major platform, we provide
242 an importable Virtual Machine with Mesmerize pre-installed so that users can get up and running
243 within minutes. Mesmerize is also on PyPI, which allows it to be installed via pip – the prevailing
244 package manager for Python. We have a dedicated YouTube channel with more than 150 minutes of
245 video tutorials, we host an active GitHub community to provide troubleshooting help, software
246 maintenance, and a gitter room for open discussions. Mesmerize is regularly updated and there have
247 been five releases in the past year (excluding bug-fix releases). This paper describes Mesmerize v0.7.1.
248 See section "Documentation, source code and assistance" for details.

249 **Usage Examples**

250 **Calcium imaging in the mouse visual cortex in response to visual sinusoidal grating stimuli**

251 Before we illustrate the more complex and novel analysis that can be performed with
252 Mesmerize, we demonstrate its use for basic neurobiological analysis using a well-known phenomenon
253 and a simple dataset. We used a mouse visual cortex dataset (dataset name: CRCNS pvc-7) contributed
254 by the Allen Brain Institute, which consists of *in-vivo* 2-photon imaging data from layer 4 cells in the
255 mouse visual cortex[35] (Fig 2a). The recording was performed while the mouse was presented with
256 visual stimuli consisting of sinusoidal bands at various orientations, spatial frequencies, and temporal
257 frequencies. The stimulus mapping module in Mesmerize allows users to map temporal annotations,
258 such as the characteristics of the visual stimuli in this experiment (Fig 2b). However, it can be used to
259 map any temporal variable, such as behaviors and other forms of stimuli, with any number of
260 characteristics. These temporal mappings can be entered manually through the GUI, or the scripting
261 interface can be used to import a temporal mapping from a spreadsheet file. As we will show, these

temporal mappings can be incorporated into downstream analysis – an essential feature for streamlined analysis in systems neuroscience. The CaImAn NoRMCorre[22] module and CNMF[23] were used for motion correction and signal extraction respectively (Fig 2c). A flowchart, illustrated in Fig 2d, can then be used to determine how cells are tuned to various characteristics of the visual stimuli. An interactive heatmap can be used to visualize the result (Fig 2e). The heatmap can be labelled and sorted according to any categorical variable in the dataset, such as the orientation, spatial frequency, and temporal frequency that each cell is tuned to. As mentioned previously, clicking a datapoint in the heatmap will update the *Datapoint Tracer,* which then 1) highlights the spatial localization of the ROI that the datapoint originates from, 2) displays all other data associated to the datapoint (Fig 2e, bottom center), and 3) lists the analysis log (Fig 2e, top center) which can be exported as an analysis graph (Supplementary Figure 1). Another visualization that is appropriate for these data are Spacemaps. These allow users to spatially visualize categorical analysis results or annotations within the imaging field. For example, we show orientation tuning (Fig 2f), spatial frequency tuning (Fig 2g) and temporal frequency tuning (Fig 2h) of the cells in the "CRCNS pvc-7" dataset. The analysis of this basic dataset illustrates how Mesmerize can encapsulate entire analysis workflows.

Analysis of a volumetric zebrafish calcium imaging dataset coupled to somatosensory stimulation

Mesmerize is also capable of handling 3D volumetric imaging datasets with the same annotation and analysis capabilities that are provided for 2D datasets. In order to demonstrate some of these features we analyzed an *in-vivo* 2-photon imaging dataset where zebfrafish larvae expressing a nuclear localized GCaMP are presented with various forms of heat stimuli[36] (Fig 3a). Users are provided with multiple options for ROI extraction from 3D data. Mesmerize can interface with the Caiman 3D CNMF[23] implementation, or each plane can be processed individually using Caiman 2D CNMF. Furthermore, Mesmerize can utilize the NuSeT[26] network to provide a deep-learning based segmentation tool for ROI extraction. These NuSeT-segmented ROIs that can then be used to initialize

286 CNMF. This example demonstrates how Mesmerize's modular platform greatly simplifies the process
287 of combining multiple cutting-edge tools, allowing them to be more easily adopted by a broader range
288 of users. For this 3D dataset, CNMF with greedy initialization performed poorly (Fig 3b), which is
289 likely due to lower signal-to-noise ratios that are more common with 2-photon volumetric imaging[37].
290 However, the performance of CNMF is greatly improved when it is initialized with binary masked
291 produced by NuSeT (Fig 3b). After ROI extraction, the stimulus information was temporally mapped
292 and a few imaging samples were used to create a Mesmerize project and perform downstream analysis.
293 Interactive stimulus tuning plots can be obtained for every cell (Fig 3c-d), and these can be used to sort
294 cells according to the stimulus they are tuned for (Fig 3e) and visualized using a spacemap (Fig 3f).
295 Lastly, we used Mesmerize to train a Linear Discriminant Analysis (LDA) model and classified three
296 distinct brain states that are observed during heat-on, heat-on-delayed and pre-stimulus (none) periods
297 (Fig 3g). Put together, these demonstrate Mesmerize's capabilities in handling 3D calcium imaging
298 data and identifying distinct brain states using standard machine learning approaches, such as LDA
299 decomposition. This example demonstrates how Mesmerize's suite of analysis tools and annotation
300 capabilities makes it a game-changer for cutting-edge systems neuroscience researchers in the present
301 and into the future as volumetric imaging becomes more widespread.

302 **Functional fingerprinting of neuronal and non-neuronal cell types in *C. intestinalis***

303 Having demonstrated how Mesmerize can be used to tackle several popular experimental
304 paradigms in neuroscience, where neuronal dynamics are analyzed in the context of stimuli or
305 behavior, we next addressed more contemporary/non-standard forms of analysis, with the aim of
306 making novel biological findings. We thus turned our attention to spontaneous calcium activity datasets
307 from both neuronal and non-neuronal cells in the absence of well-defined stimuli, in cells where typical
308 neuronal spike trains have not been observed previously by leveraging the emerging model organism
309 for systems neuroscience, the protochordate *Ciona intestinalis*. Neurobiological studies in *C.*

310 *intestinalis* have just gained momentum, with a handful of ethological studies[38–40] and a few studies of
311 calcium dynamics[41]. However no pan-neuronal calcium imaging analysis has been performed and such
312 a study would be a great resource for the Ciona and greater chordate community.

313 We chose *C. intestinalis* as model system to address the unique and fundamental question of
314 spontaneous neuronal activity in neuronal and non-neuronal cells for multiple reasons. First, the recent
315 completion of the larval connectome[42–44] in conjunction with the generation of comprehensive single-
316 cell transcriptomes[45,46] establishes the nervous system of *C. intestinalis* as likely the most thoroughly
317 mapped chordate nervous system to date. Second, despite the established connectome, there has not
318 been a comprehensive functional study to investigate neuronal activity across its diverse neuronal
319 populations. Third, its small nervous system, flat head, and the ability to label genetically defined
320 populations of cells using various promoters that drive GCaMP6s expression allow us to approximate
321 the identity of neuronal cells in reference to the connectome[42,43]. Finally, to showcase comprehensive
322 comparative calcium dynamics analysis within the same organism for applications beyond
323 neuroscience, we additionally performed calcium imaging in two non-neuronal cell types in *C.*
324 *intestinalis*, the epidermis and a population of migratory mesenchymal cells termed trunk lateral cells[47]
325 (TLCs).The analysis methods developed in this work can be employed by cell and developmental
326 biologists to study calcium-dependent mechanisms that underlie a broad range of cell biological and
327 morphogenetic processes.

328 Since our goal here was to quantitatively define calcium activities in cells and domains where
329 typical neuronal spike trains have not been observed previously, we implemented techniques which
330 have not been used prior to our study to analyze calcium dynamics. These methods can also be applied
331 to understand calcium dynamics in other systems. Frequency-domain analysis has previously been used
332 to compare calcium dynamics between experimental groups[48,49]and during cortical development[50],
333 however it has not been used for global clustering analysis to deduce more complex relationships

334 between cell types or experimental conditions. To fill this gap, we introduce the application of Earth
335 Mover's Distances[51,52] (EMD) between frequency domain representations of calcium traces data as a
336 distance metric for hierarchical clustering. The EMD is commonly used for pattern recognition and
337 image retrieval systems through histogram comparison[52]. Intuitively, the EMD can be thought of the
338 amount of work that must be done to transform one distribution into another. Therefore, in contrast to
339 the Euclidean distance, the Earth Mover's Distance accounts for the order of elements along two
340 feature vectors that are being compared. This makes it a useful metric for performing clustering
341 analysis using Discrete Fourier transforms (DFTs) of calcium traces since similar weights in
342 neighboring, but not identical, frequency domains are measured as a small EMD whereas the same
343 weights in far-apart frequency domains result in a large EMD between the feature vectors. To illustrate
344 this, consider the traces from two cells that appear to have similar dynamics (Fig 4a), and their
345 corresponding Fourier transforms (Fig 4b). If the order of elements along the DFT, shown as feature
346 vectors u & v (Fig 4b), are randomly shuffled, the EMD between the shuffled vectors is different
347 whereas the Euclidean distance is identical (Fig 4c).

348 Next, we show how we used the EMD to cluster calcium dynamics of neuronal and non-
349 neuronal cells from *C. intestinalis*. To conceptually demonstrate the application of Earth Mover's
350 Distances, consider ten example traces (Fig 4d). It is important to note that these traces were not
351 acquired over the same time period and we were not interested in finding neurons/cells that fire
352 together (i.e. neural assemblies). Instead, we were interested in quantitatively categorizing neurons
353 based on their overall dynamics. The EMD-based distance matrix shows better grouping than the
354 distance matrices calculated using Euclidean distances (Fig 4e-f). To quantitatively demonstrate that
355 the EMD performs better than Euclidean distances we performed hierarchical clustering and calculated
356 the agglomerative coefficient (denoted by $\boldsymbol{\alpha}$) - a score between 0 and 1 where values approaching 1
357 indicate better clustering structure. With the ten example traces, the hierarchical clustering obtained by
358 using the EMD metric results in an agglomerative coefficient $\alpha \approx 0.841$ (Fig 4g), whereas the clustering

359 obtained from Euclidean distances results in a coefficient $\alpha \approx 0.574$ (Fig 4h). When applied to a larger
360 dataset the clustering structure found through EMD is even stronger with an agglomerative coefficient
361 $\alpha \approx 0.983$ (Fig 4i), compared to $\alpha \approx 0.663$ for Euclidean distances (Fig 4j). Agglomerative coefficients
362 tend to increase with the size of a dataset, therefore smaller datasets (Fig 4e-f) are more useful for
363 evaluating performance between different metrics. Euclidean distances in the time-domain can be
364 useful for grouping cells that fire together, however this is irrelevant since the traces were not acquired
365 over the same time period.

366 To compare our methods with techniques that have previously been used in clustering analysis
367 of spontaneous neuronal activity, such as comparisons between various stages of the circadian cycle[53],
368 we benchmarked Silhouette and Davies-Bouldin scores using both hierarchical and k-means clustering.
369 Earth Mover's Distance based hierarchical clustering far outperforms standard hierarchical clustering
370 using Euclidean distances, and k-means using both the time and frequency domain (Fig 4k-l). Since the
371 data are not temporally aligned, k-means clustering would be unsuitable for our task and mostly results
372 in aligned traces as expected (Supplementary figure 2). From these dendrograms and agglomerative
373 coefficients, we demonstrate that the EMD metric between frequency-domain representations of
374 calcium traces results in better separation of disparate dynamics and an aggregation of similar
375 dynamics. Since this method is suitable for data that are not temporally aligned, it opens the potential
376 for novel analysis of spontaneous activity during circadian cycles[53], development[50], and during
377 pathological states using psychiatric disease-relevant models and paradigms[49,54].

378 To illustrate how the EMD is a simple and effective method for characterization of calcium
379 dynamics across a diverse range of cell types, we performed hierarchical clustering on traces obtained
380 by imaging various neuronal and non-neuronal populations of cells in the *C. intestinalis* head.
381 Clustering of both neuronal and non-neuronal cells resulted in a dendrogram which was cut to form 4
382 clusters, separating these cells into 4 distinct populations based on their activity profile (Fig 5a).
383 Example traces from each of the four clusters show that Cluster 1 consists of cells with very low levels

384 of activity (Fig 5b). Cells within Cluster 2 show slightly more activity, and Cluster 3 is enriched with
385 cells showing moderately more activity and shorter peaks. Cluster 4 is highly enriched with cells that
386 show very high levels of activity. The cluster centroids help further describe the characteristics of the
387 four clusters. Cluster 1 shows very high spectral energy in the lowest frequency domains, and relatively
388 no spectral energy in higher frequency domains (Fig 5c). The amount of spectral energy in the lowest
389 frequency domains increases progressively from Cluster 1 to Cluster 4, whereas the opposite is true for
390 spectral energy in higher frequency domains. Cluster 4 shows the most spectral energy in higher
391 frequency domains. Biologically, each of these 4 clusters are enriched with distinct populations of cells
392 (Fig 5d). Cluster 1 is almost exclusively composed of CESA and HNK-1 cells exhibiting wide and
393 large peaks, with large spectral energy in lower frequency domains. In contrast, neuronal cells are
394 predominantly found in Clusters 3 and 4, with a few peripheral sensory neurons also found in Cluster
395 2. Peripheral sensory neurons, such as Pap, aATEN, pATEN and RTEN, are highly enriched in Cluster
396 2 and 3. Cluster 4, with cell showing very high activity, mostly consists of various types of
397 photoreceptor cells and interneurons.

398 This analysis demonstrates that the combination of DFT with EMD allow us to identify
399 different activity states in non-neuronal cell types and to classify different neuronal cell types in
400 different groups based on their activity dynamics. We show that this clustering separates genetically
401 defined populations of peripheral & sensory neurons, from populations located within the brain vesicle
402 which form the Central Nervous System. Most interestingly, four cell types involved in peripheral
403 sensory networks namely the Palp Sensory Neurons (PSNs), the rostal trunk epidermal neurons
404 (RTEN), and the apical trunk epidermal neurons (aATEN & pATEN) exhibit similar modes of activity
405 and are enriched in Cluster 2 and 3. Previous anatomical studies[44,55,56] postulated that PSNs provide
406 feedforward excitation to the RTENs, while all four cell types appear to exhibit a glutamatergic
407 molecular signature[55,57]. The similarity in their activity 'signatures' that we observe in our imaging
408 analysis provides functional support for this hypothesis. Cells that are mostly primary interneurons

409 within the brain vesicle all exhibit high levels of activity and cluster together (Fig 5d). These cell types
410 include interneurons that are postsynaptic to the RTENs such as the peripheral interneurons (PNIN),
411 interneurons closely associated with photoreceptors such as the photoreceptor tract interneuron (trIN)
412 and the photoreceptor relay neurons (prRN), antenna relay neurons (antRN) which receive input from
413 the gravity sensing cells and finally the Eminens (Em) peripheral relay neurons which are thought to be
414 one of the main centers of integration in the larval nervous system[42]. The high activity that these
415 different types of interneurons exhibit could reflect the possibility that they receive more complex
416 inputs due to their intermediate positions in different sensory networks.

417 The distinct clustering of cell types shown here is likely indicative of cellular function and
418 molecular composition. For example, the slower calcium dynamics observed in Cluster 1 likely reflect
419 the contribution of calcium signaling in homeostatic cellular processes[58] such as epidermal barrier
420 formation and maintenance, and processes mediating motility and cell-shape changes in mesenchymal
421 cells. Neuronal cells are inherently noisy compared to other excitable cell types[59], such as epithelial
422 cells, even in the absence of any discernable stimuli. However noise, or spontaneous activity, is often
423 important for many neurobiological processes such as development[50], encoding[60] and stochastic
424 resonance[61–64] - a signal-boosting strategy employed by sensory circuits and other neurophysiological
425 systems where noise from neurons exhibiting spontaneous activity is injected to increase the sensitivity
426 of sensory circuits. Spontaneous activity in developing circuits have been studied semi-quantitatively,
427 including frequency analysis[50]. These fields could greatly benefit from a method to quantitatively
428 compare and cluster large numbers of diverse cell types to create *cell-type signatures* at various stages
429 of development, which could complement the ever growing transcriptomic data that are more
430 commonly used to generate cell-type signatures[65]. Put together, this work reveals how spontaneous
431 activity is sufficient to broadly derive cell-specific functional fingerprints in *C. intestinalis* larvae. This
432 simple but broadly applicable technique can be used in other model systems to define discrete
433 functional domains for specific populations or sub-types of neurons and provides a novel way to

434 quantitatively characterize the overall dynamics of calcium, or other molecules and ions.

435

436 **Motif extraction from shape-based analysis of calcium imaging data**

437 To extract additional valuable information from our calcium imaging datasets, here we
438 demonstrate another downstream analysis method, k-Shape clustering[31,66], on our *C. intestinalis* dataset
439 using Mesmerize. Many experiments in neuroscience and cell biology require a quantitative method to
440 define discrete archetypical shapes from calcium traces, as well as traces that may represent changes in
441 the levels of other molecules such as those obtained from neurotransmitter or voltage indicators, etc.
442 Thus, the methods described here will be broadly applicable to trace-containing datasets and not
443 limited to calcium datasets. In the early days shape archetypes were defined subjectively[67–70], and
444 currently the most common method is to describe peak-features such as amplitude, width, slope, etc[71].
445 However, certain biological systems such as the developing nervous system or adult nervous system in
446 the context of pathological conditions (e.g. seizures) display complex and irregular types of calcium
447 activity, which makes the use of such metrics less suitable. Here we apply k-Shape clustering, a
448 contemporary time-series analysis technique to tackle this problem. This method allows us to
449 comprehensively compare peaks directly so that we can reduce calcium traces to sequences of discrete
450 motifs. K-Shape clustering uses a normalized cross-correlation function to derive a shape-based
451 distance metric that can be used to extract a finite set of discrete archetypical peaks from calcium traces
452 (Fig 6a). These clusters can be visualized using PCA of peak features to illustrate how the k-Shape
453 clustering maps to more traditional peak-features based measures. K-Shape derived archetypes can then
454 be used to reduce calcium traces to sequences of discrete letters, and statistical models, such as Markov
455 Chains (Fig 6d-g), can be applied to describe calcium dynamics between different types of cells or
456 experimental groups. For example, the Markov Chains created using k-Shape-sequences derived from
457 HNK-1 traces (Fig 6d-e) are very simple, characteristic of the simple calcium dynamics that these cell
458 exhibit. On the other hand, Markov Chains that represent photoreceptor cells (Fig 6f-g) are much more

459 complex. In summary, we show that k-Shape clustering could provide a contemporary approach to
460 answering questions in various systems, such as examining stimulus-response profiles, behavioral
461 periods, etc. This approach can likely be further tailored to extract motifs from imaging calcium,
462 neurotransmitters, voltage or other Genetically Encoded Indicators (GEIs) using different organisms, to
463 investigate conserved and species-specific mechanisms.
464

465 **Conclusion**

466 We demonstrate here that Mesmerize is a platform that can be used to perform novel, complex,
467 and reproducible calcium imaging data from a diverse range of cell types and organisms.
468 Mesmerize addresses a contemporary need in the field of functional imaging namely, the requirement
469 for a platform with cutting edge analytical tools capable of tackling 2D and 3D datasets that is
470 accessible to biologists with a broad range of competence in terms of computational skills and
471 biological interests. We show that Mesmerize can analyze a wide range of datasets from multiple
472 organisms with morphologically diverse brains and cell types, which were acquired using different
473 imaging techniques (e.g., 2-photon imaging, epifluorescence) in the absence or presence of
474 spatiotemporally defined external stimuli.

475 While the creation of a user-friendly platform was of paramount importance, this should not
476 come at the expense of novelty, expandability, traceability and broad applicability. Mesmerize provides
477 new analyses techniques such as EMD based hierarchical clustering and k-Shape clustering in
478 combination with Markov Chains, equipping users with new tools to extract functional fingerprints and
479 to delineate the basic building blocks and organization of calcium activity from diverse cell types. Our
480 platform can be readily integrated with popular imaging processing tools such as Suite2p and can
481 utilize newly published cutting-edge tools such as the deep learning tool NuSeT, which as we
482 demonstrate can markedly improve the performance of the well-established and popular signal
483 extraction method CNMF(E). Importantly, Mesmerize's capacity to produce FAIR datasets by the

484 encapsulation of raw data, analysis procedures and interactive plots *en masse* provides a blueprint for
485 other projects and future software platforms. In future directions, Mesmerize could provide
486 neuroscientists with a user-friendly interface to back-end tools such as DataJoint[6] and NWB[7]. This will
487 help create a community where traceable visualizations and reproducible analysis become more
488 common in the biological sciences.

489 Mesmerize provides the opportunity to combine functional fingerprinting (calcium signal or
490 other using GEIs) with genetic fingerprinting (e.g. regulatory elements) in genetically tractable
491 organisms with the potential to simplify systems-level analyses that utilize complex combinations of
492 categorical variables that include multiple genotypes, drugs, and other experimental groups. Our
493 functional imaging analysis of genetically defined neuronal and non-neuronal cell types in *C.*
494 *intestinalis* showed that different neuronal cell types can be grouped together based on their calcium
495 fingerprint. In addition, it also revealed for the first time some of the basic building blocks that build
496 the observed calcium activity (k-Shape derived archetypes) and how these building blocks can be
497 organized (Markov Chains) in order to generate distinct calcium dynamics. The *C. intestinalis* datasets
498 (both neuronal and non-neuronal) generated in this work will enrich an ever-growing ecosystem of
499 openly available genomic[45,46], morphological and genetic[72–74] resources for an emerging model system
500 for neuroscience and beyond.

501

502 **Methods**

503 **Obtaining *C. intestinalis***

504 Adult *Ciona intestinalis* were obtained and maintained as described previously[38]. Briefly, the
505 adults were collected from Døsjevika, Bildøy Marina AS near Bergen, 5353, Norway and housed in
506 filtered seawater at 10°C in constant illumination.

507 **Electroporation of zygotes**

508 Electroporation was performed as described by L. Christiaen *etc. al*.[75]; adult *Ciona intestinalis*
509 were dissected to obtain eggs & sperm to perform fertilization *in-vitro*. Zygotes were then
510 dechorionated using sodium-thioglycolate solution and placed on a rocker for ~6 minutes until zygotes
511 were fully dechorionated. Zygotes were electroporated in a mannitol solution with 70-100µg of DNA
512 depending on the typical expression levels of a given construct. Embryos were cultured in ASW
513 (artificial sea water, Red Sea Salt) at 14°C until they were swimming larvae to be used for imaging.
514 The pH of the ASW was 8.4 at 14 °C. The salinity of the ASW was 3.3–3.4%.

515 **Imaging**

516 Stage 26 larvae were embedded in 1.5% low melting point agarose (Fisher BioReagents,
517 BP1360-100) between two coverslips to minimize scattering and bathed in artificial sea water.
518 Illumination was provided by a mercury lamp with a BP470/20, FT493, BP505-530 filterset. A
519 Hamamatsu Orca FlashV4 CMOS camera acquired images at 10Hz with exposure times of 100ms
520 using a custom application[76] using a python library for interfacing with Hamamatsu cameras[77]. Imaging
521 was performed at 16°C using a Zeiss Examiner A1 with a water immersion objective ZEISS W B-
522 ACHROPLAN 40x.

523 **Signal Extraction**

524 Images were motion corrected using NoRMCorre[22] and signal extraction was performed using
525 CNMFE[25] with parameters optimized per video. Extracted signals that were merely movement or noise
526 were excluded. All parameters for motion correction and CNMFE can be seen in the available dataset.
527 Cells were identified with the assistance of the connectome[42,43] to the best of our capability with 1-
528 photon data (Supplementary Figure 3). Only regions that covered cell bodies were tagged, axons were
529 not tagged with cell identity labels.

530 **Hierarchical Clustering**

531 Analysis was performed using the Mesmerize flowchart. All traces extracted from CNMFE
532 were normalized between 0 – 1. The Discrete Fourier Transform (DFT) of the normalized data was
533 calculated using `scipy.fftpack.rfft` from the SciPy (v1.3) Python library[29]. The logarithm of the
534 absolute value of the DFT data arrays were taken, and the first 1000 frequency domains (corresponding
535 to frequencies between 0 – 1.67 Hz) were used for clustering. This cutoff was determined by looking at
536 the sum of squared differences (SOSD) between the raw curves and interpolated Inverse Fourier
537 Transforms (IFTs) of the DFTs with a step-wise increase in the frequency cutoff (Supplementary
538 Figure 4). The SOSD changes negligibly beyond 1.67 Hz, and inclusion of higher frequencies would
539 likely introduce noise. At 1001 frequency domains, corresponding to 1.676 Hz, the cumulative sum of
540 the mean SOSD corresponds to 94.5% of the total cumulative sum from all frequency domains (i.e. all
541 domains up to Nyquist frequency). Earth Mover's Distance (EMD) was used as the distance metric
542 through the OpenCV[78] (v3.4) EMD function and complete linkage was used for constructing the tree.
543 The dendrogram was cut to obtain 4 clusters according to the maxima of the silhouette scores (Fig 4k).
544 The Davies-Bouldin score was also relatively low for 4 clusters (Fig 4l). Silhouette scores were
545 calculated using sklearn[30] v0.23 and a custom written function was used to adapt the Davies-Bouldin
546 score for Earth Mover's Distances. Euclidean Davies Bouldin scores were calculated using sklearn[30]
547 v0.23.

548 **k-Shape Clustering**

549 This method uses a normalized cross-correlation function to derive a shape-based distance
550 metric[66]. The tslearn[31] implementation is used in Mesmerize. Tslearn v0.4 was used. Peak-curves were
551 used as the input data for k-Shape clustering and the parameters can see seen in Supplementary Figure
552 5. A gridsearch was performed to optimize the hyperparameters and obtain a set of clusters with
553 minimum inertia (sum of within cluster distances) with no empty clusters. The search range for the
554 number of clusters to form was 2-14. For each iteration of the gridsearch, peak-curves were ordered
555 based on half-peak-width and partitioned into *n_cluster* partitions and a random centroid seed was
556 picked from each partition.

557 **Markov Chains**

558 Cluster membership of peaks, as determined through k-Shape clustering, was used to express
559 calcium traces as discretized sequences. These sequences were used to create Markov Chain models
560 using the pomegranate[79] Python library.

561 **Determining stimulus tuning of cell within the CRCNS pvc-7 and zebrafish datasets**

562 All stimulus periods were extracted and the average response was calculated for each stimulus,
563 such as an orientation, spatial frequency, or temporal frequency for the pvc-7 data set; or heat-on, heat-
564 off, and none (inter-trial period). The stimulus tuning of the cell was then determined as the stimulus
565 which produced the highest mean response in that cell. For more details, this is calculated by the
566 `get_tuning_curves()` function within `mesmerize.plotting.widgets.stimulus_tuning.widget`. The
567 analysis graph for the analysis of the pvc-7 dataset can be seen in Supplementary Figure 1, and the
568 analysis graph for the analysis of the zebrafish dataset can be seen in Supplementary Figure 6.

569 **Linear Discriminant Analysis**

570 The "Neural Decompose" node was used in the Mesmerize flowchart to perform supervised
571 LDA. Each timepoint of the recording is used as a feature vector containing the intensity values for
572 each cell at that timepoint. The model was trained using the stimulus periods (heat-on, heat-on-delayed,
573 and none) for classification.

574 **Promoters**

575

576 To drive the expression of GCaMP6s population in different cell types in *Ciona intestinalis* larvae we
577 used the following promoters:

Gene Unique ID	Gene Model ID	Name	Abbr.	Length
Cirobu.g00010959	KH.L128.92	Proprotein/Prohormone convertase 2	pc2	2.86kb
Cirobu.g00008038	KH.C7.211	CesA	cesa	2.2kb
Cirobu.g00014653	KH.S544.3	DMRT1	dmrt1	1.29kb
Cirobu.g00004616	KH.C2.42	Brn3b/POU4	brn3b	3.78kb
Cirobu.g00006491	KH.C4.403	HNK1[33]	hnk1	3.0kb
Cirobu.g00010171	KH.C9.608	PDE9	pde9	4.43kb
Cirobu.g00012642	KH.L42.6	CNG Channel 4	cng_ch4	1.48kb
Cirobu.g00003963	KH.C14.52	EEF1A1	eef1a	1.96kb

578 Sequences for several of these promoters were obtained from DBTGR[74].

579

Primer name	Primer sequence
PC2 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gCAGCAGTCAAAGGGTTTCTTGAAACAC
PC2 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gGCTGCTTTAAGAATTCTTCGTTTTTTCAC
CesA GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gCCCGGTGCTTTGAAAATTGACAAG
CesA GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gGAACTCGTATATCTTGATGGTTTGG
DMRT1 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gTCAGAACGAGGCGCTACATGATC
DMRT1 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gCACTGTTCTAAGCAAGGTATCAAGG
Brn3b/Pou4 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gCGACTGTAACAAGTTCTAAACAGAGC
Brn3b/Pou4 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gATATCGTATCAAAAAATATACAATAAGTCTG
HNK1 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gCAGCACGGGTTGAGTCAATGAAAC
HNK1 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gACGCACCAGGAAGTTAAATAAAACC

PDE9 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gATTCATGGCTGATATACCCGGTTG
PDE9 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gCTATGCTGTTGTAGAATCTGTATATAG
CNG4 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gCTCCGTTTCGTGGAAAACTCATTTTTC
CNG4 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gACTGGACTCTAGACACAGACAGC
EEF1A1 GW-FW	g g g g a c a a c t t t g t a t a g a a a a g t t gGTGACGGGAAAACGATAGTCG
EEF1A1 GW-RV	g g g g a c t g c t t t t t t g t a c a a a c t t gTTTGGAAGGTTGGGGTTAACC

The amplified PCR products were gel purified and inserted into P4-P1R vector using BP Clonase II. Positive clones identified by restriction digest were sequenced. Subsequently we performed a 4-way Gateway Recombination using one of the promoters in the 1st position, GCaMP6s in the 2nd positon and unc-54 3'UTR in the 3rd position. These were recombined into a pDEST II. Expression constructs were electroporated at a range of concentrations (80-120µg).

C. elegans strain generation and imaging

To generate construct drg1 [prab-3::GCaMP6m::NLS::unc-54 3'UTR] we performed a 4-way Gateway recombination reaction using LR Clonase II (Invitrogen). We recombined pDEST II with the following entry clones: 1st position a 1.2kb promoter of rab-3 (a kind gift from Dr. Inja Radman, Chin lab, MRC LMB); 2nd position GCaMP6m fused to SV40NLS at the N-terminus and EGL-13 NLS sequence at the C-terminus and 3rd position unc-54 3'UTR. The resulting construct was injected into N2 animals at 100µg/µl to generate strain SCB1. C. elegans young adults were immobilized on 1% agarose pads (in M9) using DERMABOND (2-Octyl Cyanoacrylate) glue.

Dataset availability

The datasets are available as a Mesmerize project and can be downloaded from figshare:

C. intestinalis: https://doi.org/10.6084/m9.figshare.10289162

C. elegans: https://doi.org/10.6084/m9.figshare.10287113

CRCNS pvc-7 as a Mesmerize dataset: https://doi.org/10.6084/m9.figshare.10293041

Zebrafish dataset as a Mesmerize dataset: https://doi.org/10.6084/m9.figshare.14748915

Notebooks that produce some of the figures and the Markov Chains are available on GitHub and can be used on binder.

https://github.com/kushalkolar/mesmerize_manuscript_notebooks

https://mybinder.org/v2/gh/kushalkolar/mesmerize_manuscript_notebooks/master

Author Contributions

K.K. wrote Mesmerize and analyzed all experiments. D.D. aided and contributed to the development of Mesmerize and provided critical input. Imaging experiments were performed by K.K. and M.C. GCaMP6s constructs were cloned by M.C., J.C.Z. and J.H. assisted with significant user testing of the Mesmerize platform and aided in development. J.C.Z. created the Mesmerize logo. The manuscript was written by K.K. and M.C.

Documentation, source code and assistance:

Mesmerize documentation: http://docs.mesmerizelab.org/

GitHub repository: https://github.com/kushalkolar/MESmerize

Gitter community for discussion: https://gitter.im/mesmerize_discussion/community

Video tutorials: https://www.youtube.com/playlist?list=PLgofWiw2s4REPxH8bx8wZo_6ca435OKqg

Additional video tutorials:

https://www.youtube.com/playlist?list=PLgofWiw2s4RF_RkGRUfflcj5k5KUTG3o_

Acknowledgements

We would like to thank Mesmerize users and the community for their engagement in the gitter channel and GitHub for constant feedback and bug reports. We would like to thank Pietro Vertechi and Julius Parulek for technical advice and members of the Chatzigeorgiou lab for user feedback during Mesmerize's early development. We thank Mie Wong and Dario Sarra for comments on the manuscript. Work in MC's laboratory was funded by Sars Centre core budget.

Competing interests

The authors declare no competing interests.

References

1. Giovannucci, A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *Elife* (2019) doi:10.7554/eLife.38173.
2. Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv* (2016) doi:10.1101/061507.
3. Kaifosh, P., Zaremba, J. D., Danielson, N. B. & Losonczy, A. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Front. Neuroinform.* (2014) doi:10.3389/fninf.2014.00080.
4. Cantu, D. A. *et al.* EZcalcium: Open-Source Toolbox for Analysis of Calcium Imaging Data. *Front. Neural Circuits* **14**, 25 (2020).
5. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* vol. 9 671–675 (2012).

6. Yatsenko, D. *et al.* DataJoint: managing big scientific data using MATLAB or Python. *bioRxiv* 031658 (2015) doi:10.1101/031658.
7. Teeters, J. L. *et al.* Neurodata Without Borders: Creating a Common Data Format for Neurophysiology. *Neuron* **88**, 629–634 (2015).
8. Chessel, A. An Overview of data science uses in bioimage informatics. *Methods* vol. 115 110–118 (2017).
9. Jennings-Antipov, L. D. & Gardner, T. S. Digital publishing isn't enough: the case for 'blueprints' in scientific communication. *Emerg. Top. Life Sci.* (2018) doi:10.1042/etls20180165.
10. Stall, S. *et al.* Make scientific data FAIR. *Nature* (2019) doi:10.1038/d41586-019-01720-7.
11. Allan, C. *et al.* OMERO: Flexible, model-driven data management for experimental biology. *Nature Methods* vol. 9 245–253 (2012).
12. Rubens, U. *et al.* BIAFLOWS: A Collaborative Framework to Reproducibly Deploy and Benchmark Bioimage Analysis Workflows. *Patterns* **1**, 100040 (2020).
13. Marée, R. *et al.* Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics* **32**, 1395–1401 (2016).
14. Bauch, A. *et al.* OpenBIS: A flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics* **12**, 468 (2011).
15. Fillbrunn, A. *et al.* KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology* vol. 261 149–156 (2017).
16. Perkel, J. M. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature* **554**, 133–134 (2018).
17. Introducing eLife's first computationally reproducible article | Labs | eLife. https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article.
18. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* (2016) doi:10.1038/sdata.2016.18.
19. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* (2010).
20. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* (2011) doi:10.1109/MCSE.2011.37.
21. Harris, C. R. *et al.* Array programming with NumPy. *Nature* vol. 585 357–362 (2020).
22. Pnevmatikakis, E. A. & Giovannucci, A. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* (2017) doi:10.1016/j.jneumeth.2017.07.031.
23. Pnevmatikakis, E. A. *et al.* Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron* **89**, 285 (2016).
24. Pnevmatikakis, E. A. *et al.* A structured matrix factorization framework for large scale calcium imaging data analysis. 1–16 (2014).
25. Zhou, P. *et al.* Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife* 1–37 (2018) doi:10.7554/eLife.28728.
26. Yang, L. *et al.* NuSeT: A deep learning tool for reliably separating and analyzing crowded cells. *PLOS Comput. Biol.* **16**, e1008193 (2020).
27. Kolar, K. GitHub - kushalkolar/nuset-lib: NuSeT packaged as a library with an easy to use API. https://github.com/kushalkolar/nuset-lib.
28. Campagnola, L. pyqtgraph. www.pyqtgraph.org.
29. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
30. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

31. Tavenard, R. *et al.* Tslearn, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **21**, 1–6 (2020).
32. Kluyver, T. *et al. Jupyter Notebooks—a publishing format for reproducible computational workflows*. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016). doi:10.3233/978-1-61499-649-1-87.
33. Jupyter, P. *et al.* Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. in *Proceedings of the 17th Python in Science Conference* (2018). doi:10.25080/majora-4af1f417-011.
34. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* (2007) doi:10.1109/MCSE.2007.55.
35. Garner, A. In vivo calcium imaging of layer 4 cells in the mouse using sinusoidal grating stimuli. (2014) doi:10.6080/K0C8276G.
36. Haesemeyer, M., Robson, D. N., Li, J. M., Schier, A. F. & Engert, F. A Brain-wide Circuit Model of Heat-Evoked Swimming Behavior in Larval Zebrafish. *Neuron* **98**, 817-831.e6 (2018).
37. Keller, P. J. & Ahrens, M. B. Visualizing whole-brain activity and development at the single-cell level using light-sheet microscopy. *Neuron* vol. 85 462–483 (2015).
38. Rudolf, J., Dondorp, D., Canon, L., Tieo, S. & Chatzigeorgiou, M. Automated behavioural analysis reveals the basic behavioural repertoire of the urochordate Ciona intestinalis. *Sci. Rep.* **9**, 1–17 (2019).
39. Kourakis, M. J. *et al.* Parallel visual circuitry in a basal chordate. *Elife* **8**, e44753 (2019).
40. Salas, P., Vinaithirthan, V., Newman-Smith, E., Kourakis, M. J. & Smith, W. C. Photoreceptor specialization and the visuomotor repertoire of the primitive chordate Ciona. *J. Exp. Biol.* **221**, (2018).
41. Okawa, N. *et al.* Cellular identity and Ca2+ signaling activity of the non-reproductive GnRH system in the Ciona intestinalis type A (Ciona robusta) larva. *Sci. Rep.* **10**, 18590 (2020).
42. Ryan, K., Lu, Z. & Meinertzhagen, I. A. The CNS connectome of a tadpole larva of Ciona intestinalis (L.) highlights sidedness in the brain of a chordate sibling. *Elife* **5**, 1–34 (2016).
43. Ryan, K. & Meinertzhagen, I. A. Neuronal identity: the neuron types of a simple chordate sibling, the tadpole larva of Ciona intestinalis. *Curr. Opin. Neurobiol.* **56**, 47–60 (2019).
44. Ryan, K., Lu, Z. & Meinertzhagen, I. A. The peripheral nervous system of the ascidian tadpole larva: Types of neurons and their synaptic networks. *J. Comp. Neurol.* (2018) doi:10.1002/cne.24353.
45. Sharma, S., Wang, W. & Stolfi, A. Single-cell transcriptome profiling of the Ciona larval brain. *Dev. Biol.* (2019) doi:10.1016/j.ydbio.2018.09.023.
46. Cao, C. *et al.* Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* (2019) doi:10.1038/s41586-019-1385-y.
47. Jeffery, W. R. *et al.* Trunk lateral cells are neural crest-like cells in the ascidian Ciona intestinalis: Insights into the ancestry and evolution of the neural crest. *Dev. Biol.* (2008) doi:10.1016/j.ydbio.2008.08.022.
48. Tibau, E., Valencia, M. & Soriano, J. Identification of neuronal network properties from the spectral analysis of calcium imaging signals in neuronal cultures. *Front. Neural Circuits* **7**, 1–16 (2013).
49. Rosch, R. E., Hunter, P. R., Baldeweg, T., Friston, K. J. & Meyer, M. P. Calcium imaging and dynamic causal modelling reveal brain-wide changes in effective connectivity and synaptic dynamics during epileptic seizures. *PLoS Comput. Biol.* **14**, 1–23 (2018).
50. Luhmann, H. J. *et al.* Spontaneous neuronal activity in developing neocortical networks: From single cells to large-scale interactions. *Front. Neural Circuits* **10**, 1–14 (2016).
51. Monge, G. Mémoire sur la théorie des déblais et de remblais. in *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*

(1781).
52. Rubner, Y., Tomasi, C. & Guibas, L. J. Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* (2000) doi:10.1023/A:1026543900054.
53. Cox, J., Pinto, L. & Dan, Y. Calcium imaging of sleep-wake related neuronal activity in the dorsal pons. *Nat. Commun.* **7**, 1–7 (2016).
54. Seshadri, S., Hoeppner, D. J. & Tajinda, K. Calcium imaging in drug discovery for psychiatric disorders. *Front. Psychiatry* **11**, 1–8 (2020).
55. Horie, T., Kusakabe, T. & Tsuda, M. Glutamatergic networks in the Ciona intestinalis larva. *J. Comp. Neurol.* **508**, 249–263 (2008).
56. Takamura, K., Minamida, N. & Okabe, S. Neural Map of the Larval Central Nervous System in the Ascidian Ciona intestinalis. *Zoolog. Sci.* **27**, 191–203 (2010).
57. Horie, T. *et al.* Ependymal cells of chordate larvae are stem-like cells that form the adult nervous system. *Nature* **469**, 525–528 (2011).
58. Clapham, D. E. Calcium Signaling. *Cell* **131**, 1047–1058 (2007).
59. Stein, R. B., Gossen, E. R. & Jones, K. E. Neuronal variability: noise or part of the signal? *Nat. Rev. Neurosci.* **6**, 389–397 (2005).
60. Tkačik, G., Prentice, J. S., Balasubramanian, V. & Schneidman, E. Optimal population coding by noisy spiking neurons. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14419–14424 (2010).
61. Gammaitoni, L., Hänggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–287 (1998).
62. Longtin, A. Stochastic resonance in neuron models. *J. Stat. Phys.* **70**, 309–327 (1993).
63. Longtin, A. Autonomous stochastic resonance in bursting neurons. *Phys. Rev. E* **55**, 868–876 (1997).
64. Gluckman, B. J. *et al.* Stochastic Resonance in a Neuronal Network from Mammalian Brain. *Phys. Rev. Lett.* **77**, 4098–4101 (1996).
65. Miller, J. A. *et al.* Common cell type nomenclature for the mammalian brain. *Elife* **9**, 1–23 (2020).
66. Paparrizos, J. & Gravano, L. k-Shape: Efficient and Accurate Clustering of Time Series. *ACM SIGMOD Rec.* (2016) doi:10.1145/2949741.2949758.
67. Wiltgen, S. M., Dickinson, G. D., Swaminathan, D. & Parker, I. Termination of calcium puffs and coupled closings of inositol trisphosphate receptor channels. *Cell Calcium* **56**, 157–168 (2014).
68. Tovey, S. C. *et al.* Calcium puffs are genetic InsP3-activated elementary calcium signals and are downregulated by prolonged hormonal stimulation to inhibit cellular calcium responses. *J. Cell Sci.* **114**, 3979–3989 (2001).
69. Swillens, S., Dupont, G., Combettes, L. & Champeil, P. From calcium blips to calcium puffs: Theoretical analysis of the requirements for interchannel communication. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 13750–13755 (1999).
70. Bootman, M. D., Berridge, M. J. & Lipp, P. Cooking with calcium: The recipes for composing global signals from elementary events. *Cell* **91**, 367–373 (1997).
71. Mackay, L., Mikolajewicz, N., Komarova, S. V. & Khadra, A. Systematic characterization of dynamic parameters of intracellular calcium signals. *Front. Physiol.* **7**, (2016).
72. Sasakura, Y., Suzuki, M. M., Hozumi, A., Inaba, K. & Satoh, N. Maternal factor-mediated epigenetic gene silencing in the ascidian Ciona intestinalis. *Mol. Genet. Genomics* (2010) doi:10.1007/s00438-009-0500-4.
73. Sasakura, Y. *et al.* Transposon-mediated insertional mutagenesis revealed the functions of animal cellulose synthase in the ascidian Ciona intestinalis. *Proc. Natl. Acad. Sci. U. S. A.* (2005) doi:10.1073/pnas.0503640102.
74. Sierro, N. DBTGR: a database of tunicate promoters and their regulatory elements. *Nucleic*

Acids Res. (2006) doi:10.1093/nar/gkj064.

75. Christiaen, L., Wagner, E., Shi, W. & Levine, M. Isolation of sea squirt (Ciona) gametes, fertilization, dechorionation, and development. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5344 (2009).

76. Kolar, K. & Chatzigeorgiou, M. Simple GUI for acquiring images from a Hamamatsu Orca Flash 4.0 CMOS camera. (2019) doi:10.5281/ZENODO.3370464.

77. Babcock, H. *et al.* ZhuangLab/storm-control: v2019.06.28 release. (2019) doi:10.5281/ZENODO.3264857.

78. Bradski, G. The OpenCV Library. *Dr Dobbs J. Softw. Tools* (2000) doi:10.1111/0023-8333.50.s1.10.

79. Schreiber, J. pomegranate: Fast and Flexible Probabilistic Modeling in Python. *J. Mach. Learn. Res.* **18**, 1–6 (2018).

Package	Type	Suited for Ca imaging	3D calcium imaging	Motion correction	ROI Extraction	Project Management	ROI Annotation	Temporal annotation	Sample Annotation	Graphical Interfaces	Scripting interfaces	Downstream analysis	FAIR Dataset Creation	Visualization	Interactive Visualization
Mesmerize	Platform	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize	Mesmerize
Caiman	Pipeline	Available	Available	Available	Available	Not Available	Not Available	Not Available	Not Available	Limited	Available	Not Available	Not Available	Not Available	Not Available
Suite2p	Pipeline	Available	Available	Available	Available	Not Available	Not Available	Not Available	Not Available	Limited	Available	Limited	Not Available	Limited	Not Available
EZCalcium	Pipeline	Available	Not Available	Available	Available	Not Available	Not Available	Not Available	Not Available	Limited	Not Available	Not Available	Not Available	Not Available	Not Available
SIMA	Pipeline	Available	Not Available	Available	Available	Not Available	Not Available	Not Available	Not Available	Limited	Available	Not Available	Not Available	Not Available	Not Available
S. A Romano	Pipeline	Available	Available	Available	Available	Not Available	Not Available	Not Available	Not Available	Limited	Limited	Limited	Not Available	Limited	Not Available
SamuROI	GUI Tool	Available	Limited	Available	Available	Not Available	Not Available	Not Available	Not Available	Limited	Limited	Not Available	Not Available	Not Available	Not Available
DataJoint	Workflow Management	Available	Available	Not Available	Not Available	Available	Not Available	Available	Available	Limited	Available	Available	Available	Not Available	Not Available
OMERO	Platform	Not Available	Not Available	Not Available	Available	Available	Limited	Not Available	Available	Available	Available	Available	Available	Not Available	Not Available
Biaflows	Platform	Not Available	Not Available	Not Available	Available	Available	Not Available	Not Available	Limited	Available	Available	Limited	Available	Not Available	Not Available
Cytomine	Platform	Not Available	Not Available	Not Available	Available	Available	Available	Not Available	Available	Available	Available	Available	Available	Not Available	Not Available
openBIS	Platform	Not Available	Not Available	Not Available	Available	Available	Not Available	Not Available	Not Available	Not Available	Not Available	Not Available	Available	Not Available	Not Available
KNIME	Platform	Not Available	Not Available	Not Available	Available	Available	Available	Not Available	Available	Available	Available	Available	Available	Not Available	Not Available

Mesmerize:
Available:
Limited:
Not Available:

Table 1 | **Overview of various image analysis tools**
An overview of various tools for calcium imaging analysis and dataset organization. The availability of various features for calcium imaging analysis, data annotation, data management, analysis, and, visuzalition are shown.

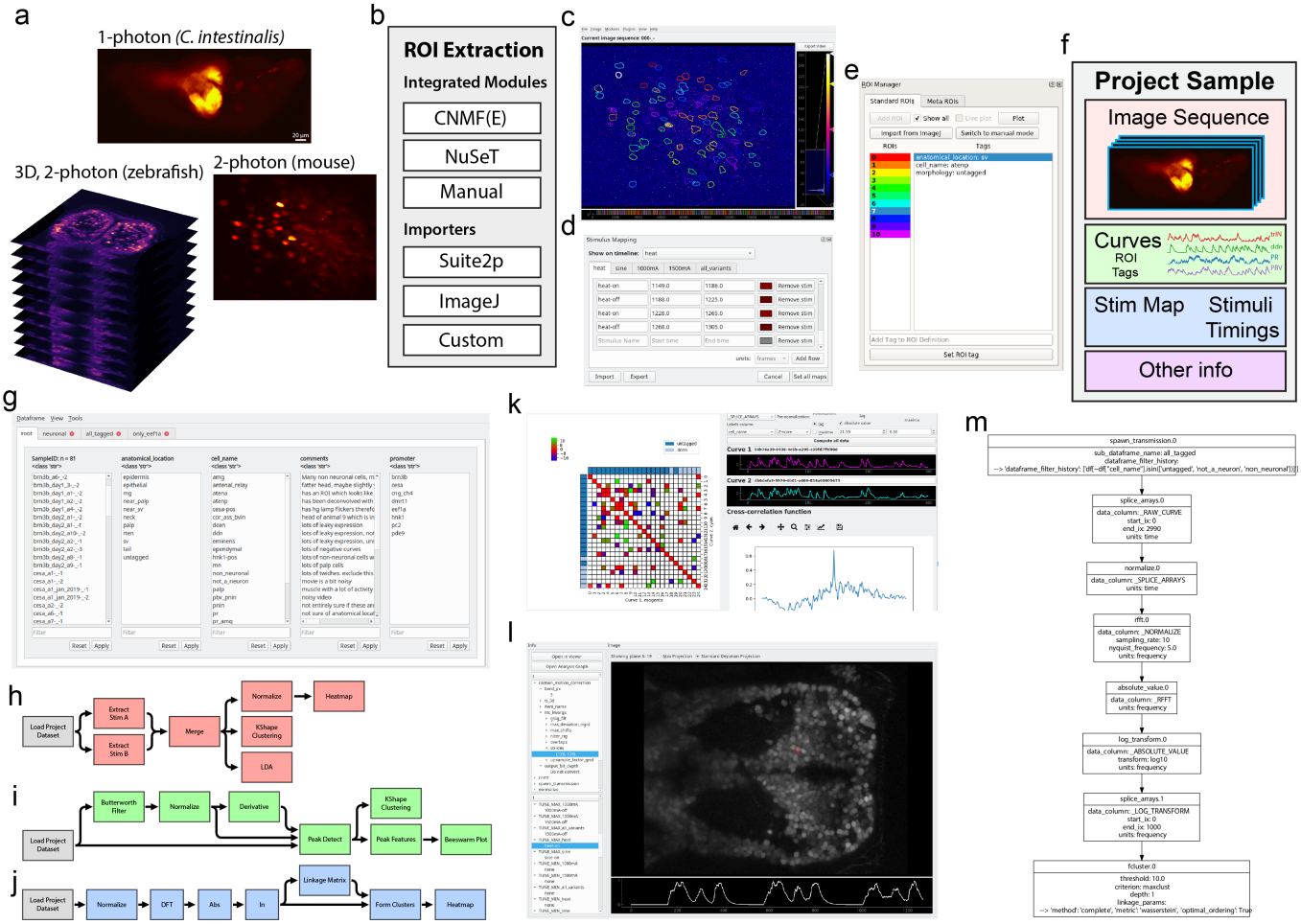
a
1-photon (C. intestinalis)
3D, 2-photon (zebrafish)
2-photon (mouse)
b
ROI Extraction
Integrated Modules
CNMF(E)
NuSeT
Manual
Importers
Suite2p
ImageJ
Custom
c
d
e
f
Project Sample
Image Sequence
Curves
ROI
Tags
Stim Map
Stimuli Timings
Other info
g
h
Load Project Dataset
Extract Stim A
Extract Stim B
Merge
Normalize
KShape Clustering
LDA
Heatmap
i
Load Project Dataset
Butterworth Filter
Normalize
Derivative
Peak Detect
KShape Clustering
Peak Features
Beeswarm Plot
j
Load Project Dataset
Normalize
DFT
Abs
ln
Linkage Matrix
Form Clusters
Heatmap
k
l
m

Fig 1 | **Mesmerize platform overview**
(a) Raw imaging data that can originate from a variety of sources, examples shown from 1-photon calcium imaging of *Ciona intestinalis*, 2-photon imagine of the mouse visual cortex neurons and volumetric 2-photon imaging of zebrafish. (b) Mesmerize's highly modular design allows ROI extraction to be performed through a variety of methods such as CaImAn CNMF(E), NuSeT deep learning, or manually. ROIs can also be imported from Suite2p, ImageJ, or a custom module can be written using the API to import ROIs from other sources. (c) The Mesmerize Viewer lets users explore their imaging data and integrates with various viewer modules such as: (d) Stimulus Mapping module which allows users to map temporal information, such as stimulus or behavioral periods; (e) ROI Manager which can work with ROIs originating from a variety of sources, as shown in [b], and allows users to tag an unlimited variety of categorial information such as anatomical location, cell type, morphology, etc. for each ROI. (f) All data pertaining to an imaging session, i.e. the image sequence, calcium curves, ROIs, tags (annotations), stimulus mappings, and all other categorical information are packaged into a "Project Sample" and saved to the Project Dataset. (g) The samples within a Project Dataset can be interactively managed using the Project Browser. (h, i, j) Project Datasets, or sub-datasets, can be loaded into a flowchart which allows users to interactively perform downstream analysis. Simplified examples of how flowcharts can be used to (e) explored stimulus or behavioral responses, (i) analyze peak features (width, amplitude, slope etc.) or perform k-Shape clustering and (j) perform hierarchical clustering. (k) Downstream analysis in flowcharts are integrated with various forms of highly interactive plots such as cross-correlation analysis. Many interactive plots are associated with a (l) Datapoint Tracer where users can click on individual datapoints to view the spatial location of the ROI that it originates from, along with all other data associated with that datapoint. (m) The Datapoint Tracer shows in [l] also lets users view the analysis history log for every datapoint in the form of an Analysis Graph. This graph will also allow users to view any pre-processing or ROI extraction parameters that were used (not shown here for simplicity).

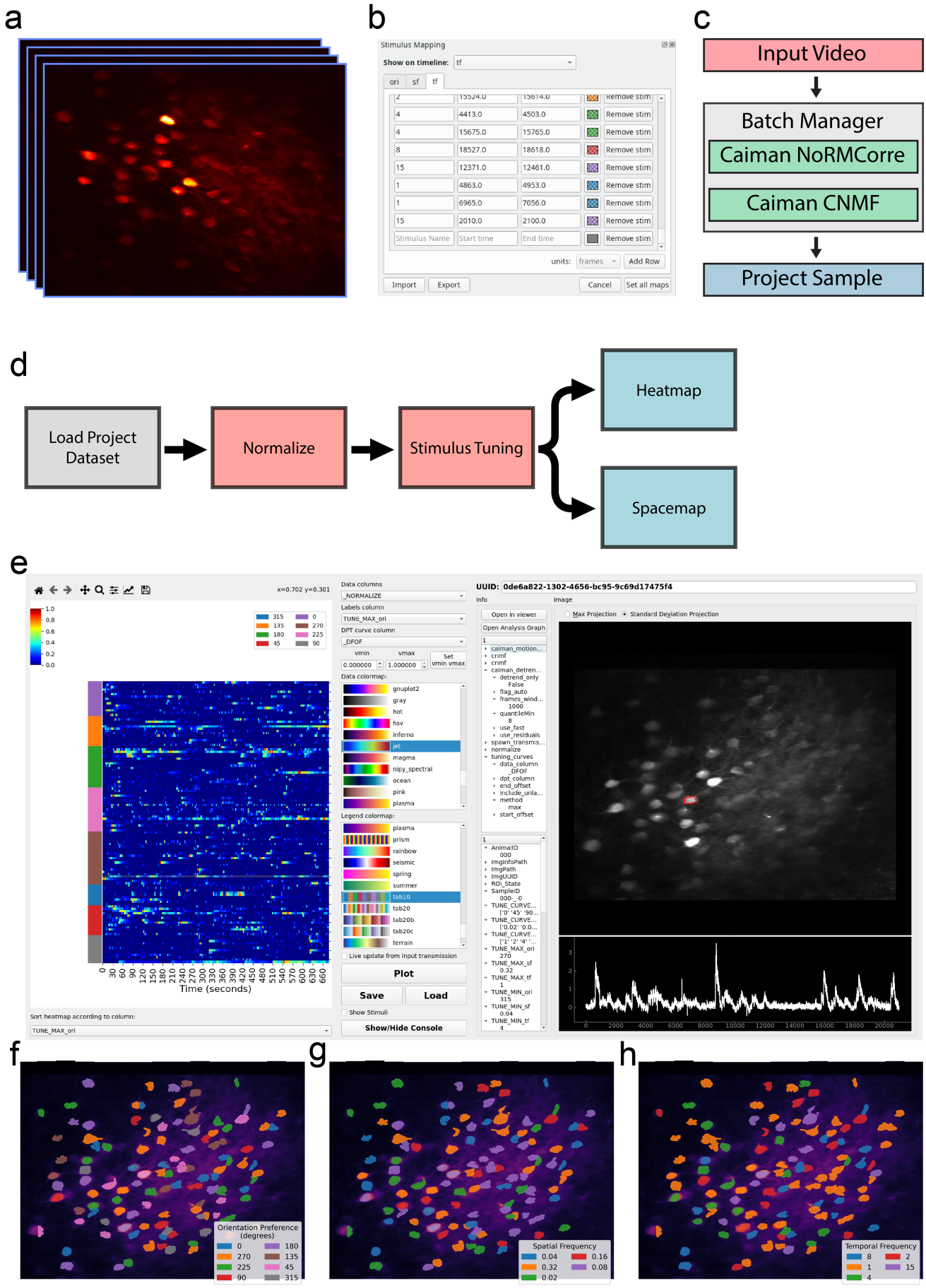
a
b
Stimulus Mapping
Show on timeline: tf
ori sf tf
Remove stim
Stimulus Name
Start time
End time
units: frames
Add Row
Import
Export
Cancel
Set all maps
c
Input Video
Batch Manager
Caiman NoRMCorre
Caiman CNMF
Project Sample
d
Load Project Dataset
Normalize
Stimulus Tuning
Heatmap
Spacemap
e
UUID: 0de6a822-1302-4656-bc95-9c69d17475f4
Data columns
_NORMALIZE
Labels column
TUNE_MAX_ori
DPT curve column
_DFOF
Data colormap:
Legend colormap:
Plot
Save
Load
Show Stimuli
Show/Hide Console
Sort heatmap according to column:
TUNE_MAX_ori
Time (seconds)
Open in viewer
Open Analysis Graph
Max Projection
Standard Deviation Projection
f
Orientation Preference (degrees)
g
Spatial Frequency
h
Temporal Frequency

Fig 2 | **Stimulus tuning of cells from the CRCNS PVC-7 dataset**.
(a) Video of cells within the visual cortex of a mouse being presented with visual stimuli consisting of sinusoidal gratings. These stimuli can be mapped onto the imaging data using the (b) Stimulus Mapping module of the Mesemrize Viewer. (c) The video was processed using the Mesmerize Batch Manager, which allows users to conveniently manage computationally intestensive tasks such as CaImAn NoRMCorre motion correction and CNMF(E). The CNMF results are imported in the Mesmerize Viewer and are packaged into a Project Sample with the imaging data and stimulus maps. (d) Flowchart which illustrates basic stimulus tuning analysis that can be performed in Mesmrize flowcharts. (e) Heatmap widget showing the results of the stimulus tuning anlysis flowchart in [d]. The heatmap shows min-max normalized calcium traces. The y-axis color labels show the orientation tuning of the cells. These plots are interactive, allowing the user to plot various forms of numerical data, such as raw traces, normalized traces, $\Delta F/F_0$, z-scored traces etc., the relationships between numerical data and various form of categorial data such as stimulus tuning, ROI tags, etc. The spatial location of the ROI and calcium trace, along with any other tagged data, can be seen on the right-hand-side panels of the widget (Datapoint Tracer). The stimulus tuning of individual cells can also be visualized using "Space maps" to visualize the (f) orientation tuning of cells, (g) spatial frequency tuning, and (h) temporal frequency tuning. Space maps can be used to visualize ROIs with respect to any categorial variables.

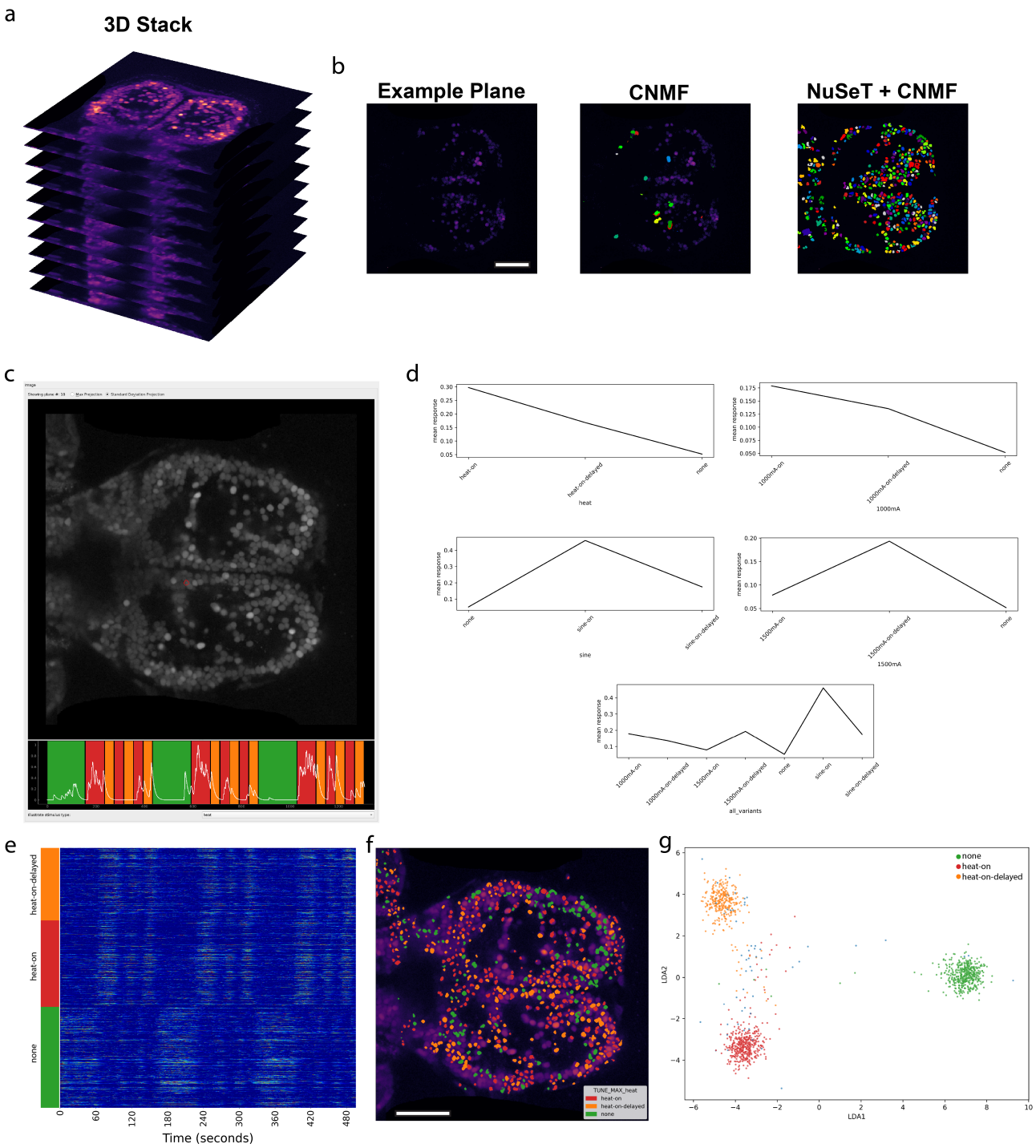
a
3D Stack
b
Example Plane
CNMF
NuSeT + CNMF
c
d
mean response
heat-on
heat-on-delayed
none
heat
1000mA-on
1000mA-on-delayed
1000mA
sine-on
sine-on-delayed
sine
1500mA-on
1500mA-on-delayed
1500mA
all_variants
e
heat-on-delayed
heat-on
none
Time (seconds)
f
TUNE_MAX_heat
heat-on
heat-on-delayed
none
g
none
heat-on
heat-on-delayed
LDA1
LDA2

Fig 3 | **Mesmerize handles 3D calcium imaging data**
(a) Mesmerize can work with volumetric calcium imaging data. (b) Frame from one plane of the volumetric dataset. CNMF with greedy initialization is unable to detect many cells in this relatively noisy dataset, however CNMF seeded with NuSeT segmentation picks up many more cells. (c) Mesmerize datapoint tracer showing a cell highlighted in red, and the corresponding calcium trace. The tuning curves of this cell are shown in (d), which shows that this cell is tuned to heat-on stimulus. (e) Min-max normalized calcium traces sorted by their stimulus tuning profiles, heat-on, heat-on-delayed, and none. (f) Space map showing the stimulus tuning characteristic of each cell. (g) LDA projection showing distinct brain states for heat-on, heat-on-delayed and none between each stimulus trial. Scalebars: 100 microns.
heat-on: cells which respond to the heat stimulus
heat-on-delayed: cells which show a delayed response to the heat stimulus
none: cells which are more active between the stimulus trials and less active during heat-on and heat-on-delayed stimulus periods.

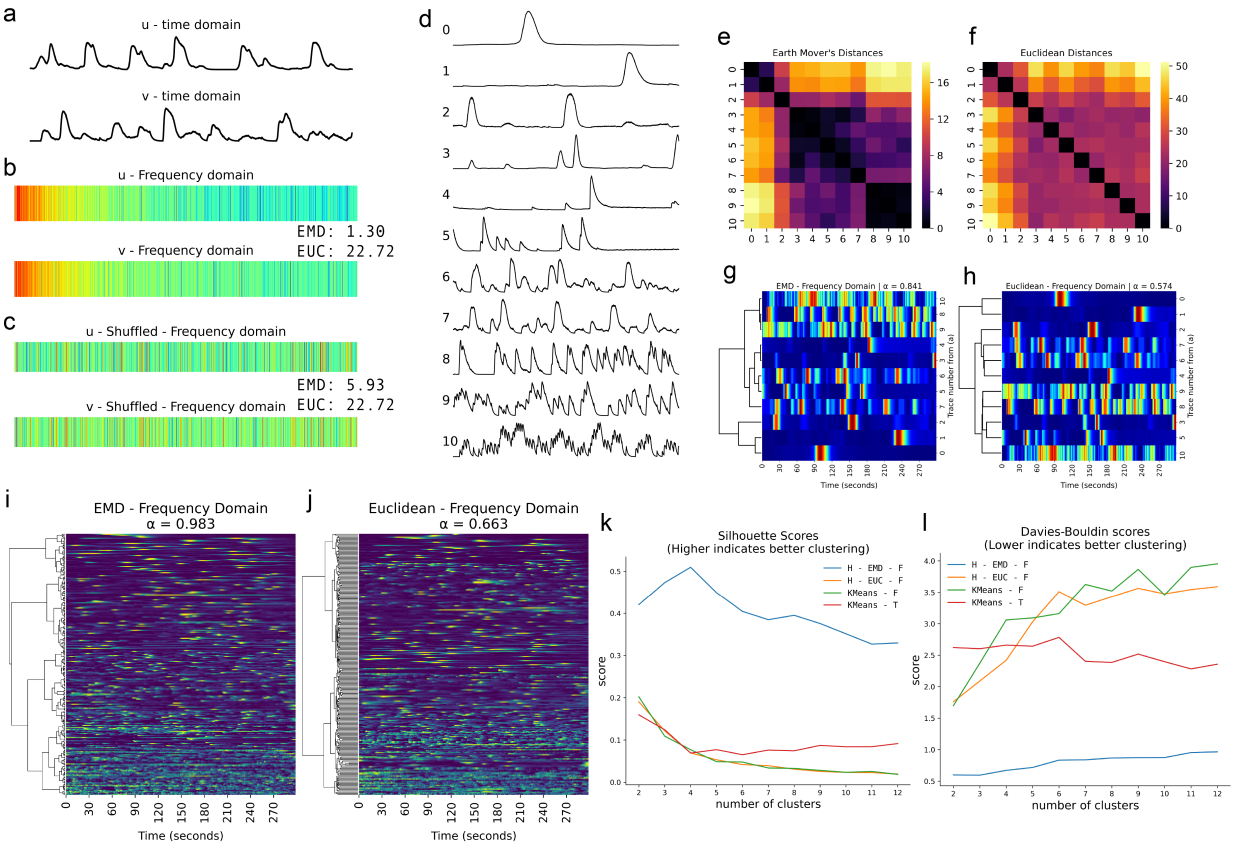
a
u - time domain
v - time domain
b
u - Frequency domain
v - Frequency domain
EMD: 1.30
EUC: 22.72
c
u - Shuffled - Frequency domain
v - Shuffled - Frequency domain
EMD: 5.93
EUC: 22.72
d
e
Earth Mover's Distances
f
Euclidean Distances
g
EMD - Frequency Domain | α = 0.841
h
Euclidean - Frequency Domain | α = 0.574
Trace number from (a)
Time (seconds)
i
EMD - Frequency Domain
α = 0.983
j
Euclidean - Frequency Domain
α = 0.663
Time (seconds)
k
Silhouette Scores
(Higher indicates better clustering)
H - EMD - F
H - EUC - F
KMeans - F
KMeans - T
score
number of clusters
l
Davies-Bouldin scores
(Lower indicates better clustering)
H - EMD - F
H - EUC - F
KMeans - F
KMeans - T
score
number of clusters

Fig 4 | **The Earth Mover's Distance is a robust metric for broadly characterizing calcium activity.** (a) Two example calcium traces, *u* & *v,* in the time domain. (b) Discrete fourier transforms of *u* & *v* are used as feature vectors. The Earth Mover's Distance (EMD) between *u* & *v* is 1.30, the Euclidean (EUC) distance between *u* & *v* is 22.72. (c) A random shuffle is applied to feature vectors *u* & *v.* The Earth Mover's Distance (EMD = 5.93) is altered by the random shuffle, however the Euclidean distance (EUC = 22.72) is identical. This demonstrates how the order of elements along a feature vector is captured by the EMD, which is necessary for effectively comparing discrete Fourier Transforms. (d) Eleven example calcium traces from *C. intestinalis*. (e) Distance matrix showing Earth Mover's Distances between discrete Fourier Transforms of the eleven calcium traces from [d]. (f) Distance matrix showing Euclidean Distances between discrete Fourier Transforms of the eleven calcium traces from [d]. (g) Dendrogram constructed from [e], with a high agglomerative coefficient ($\alpha \approx 0.841$, best = 1, worst = 0) indicating good hierarhical clustering. (h) Dendrogram constructed from [f], with a low agglomerative coefficient ($\alpha \approx 0.574$) indicating poor hierarhical clustering. (i-j) Dendrograms showing hierarhical relationships between over 200 calcium traces. (i) Dendrogram calculated using EMD, showing a very high agglomerative coefficient ($\alpha \approx 0.983$) that indicates good clustering performance. Cells near the top of the tree show slow and sparse calcium dynamics, cells closer to the bottom of the tree show much more active and complex calcium dynamics. (j) Dendrogram calculated using Euclidean distances, showing a moderate agglomerative coefficient ($\alpha \approx 0.663$). (k) Silhouette scores comparing clustering performance of various methods. Higher scores indicate better clustering performance. Hierarchical clustering using the EMD between discrete Fourier Transforms outperforms other methods. (k) Davies-Bouldin scores comparing clustering performance of various methods. Lower Davies-Bouldin scores indicate better clustering performance. This score also demonstrates that hierarchical clustering using the EMD between discrete Fourier Transforms outperforms other methods.
Legend:
H – EMD – F : Hierarchical clustering using the EMD between discrete Fourier transforms.
H – EUC – F : Hierarchical clustering using the Euclidean distances between discrete Fourier transforms.
KMeans – F : k-means clustering using discrete Fourier transforms as feature vectors.
KMeans – T : k-means clustering using calcium traces in the time domain as feature vectors.

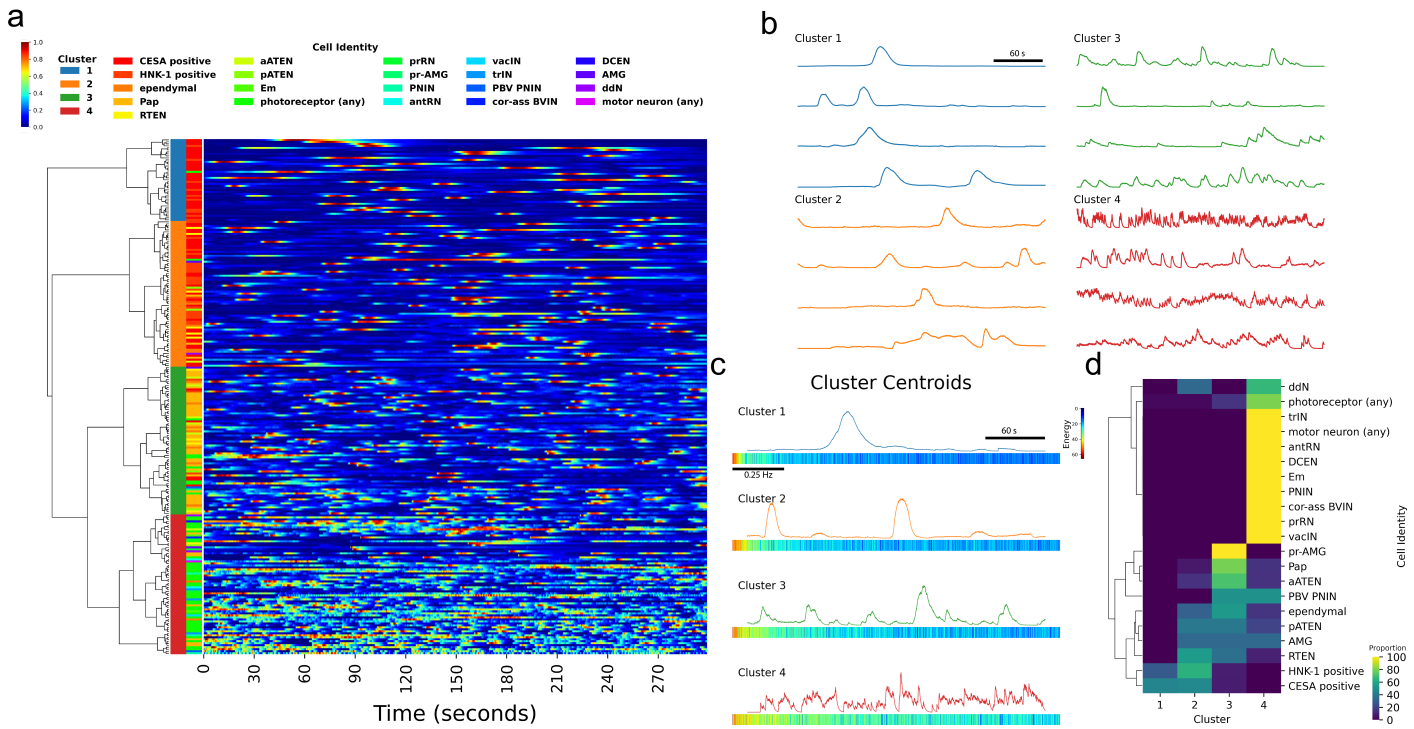
a
Cell Identity
Cluster
1
2
3
4
CESA positive
HNK-1 positive
ependymal
Pap
RTEN
aATEN
pATEN
Em
photoreceptor (any)
prRN
pr-AMG
PNIN
antRN
vacIN
trIN
PBV PNIN
cor-ass BVIN
DCEN
AMG
ddN
motor neuron (any)
0
30
60
90
120
150
180
210
240
270
Time (seconds)
b
Cluster 1
Cluster 2
Cluster 3
Cluster 4
60 s
c
Cluster Centroids
Cluster 1
Cluster 2
Cluster 3
Cluster 4
60 s
0.25 Hz
Energy
d
ddN
photoreceptor (any)
trIN
motor neuron (any)
antRN
DCEN
Em
PNIN
cor-ass BVIN
prRN
vacIN
pr-AMG
Pap
aATEN
PBV PNIN
ependymal
pATEN
AMG
RTEN
HNK-1 positive
CESA positive
Cell Identity
1
2
3
4
Cluster
Proportion
100
80
60
40
20
0

Fig 5 | **Spontaneous calcium dynamics in *C. intestinalis* reveals cell type signatures.**
(a) Hierarchical clustering of calcium dynamics observed in neuronal and non-neuronal cells within the head of ***C. intestinalis***. Dendrogram show hierarchical relationships. Left colorbar between the dendrogram and heatmap indicates cluster membership. Right colorbar legend indicates cell identity. Heatmap shows normalized traces. (b) Example traces from each cluster. (c) Cluster centroids in both the time domain (top) and frequency domain (bottom). (d) Proportion of cells that appear in each of the four clusters. For each cell type, proportions sum up to 100% across all 4 clusters.

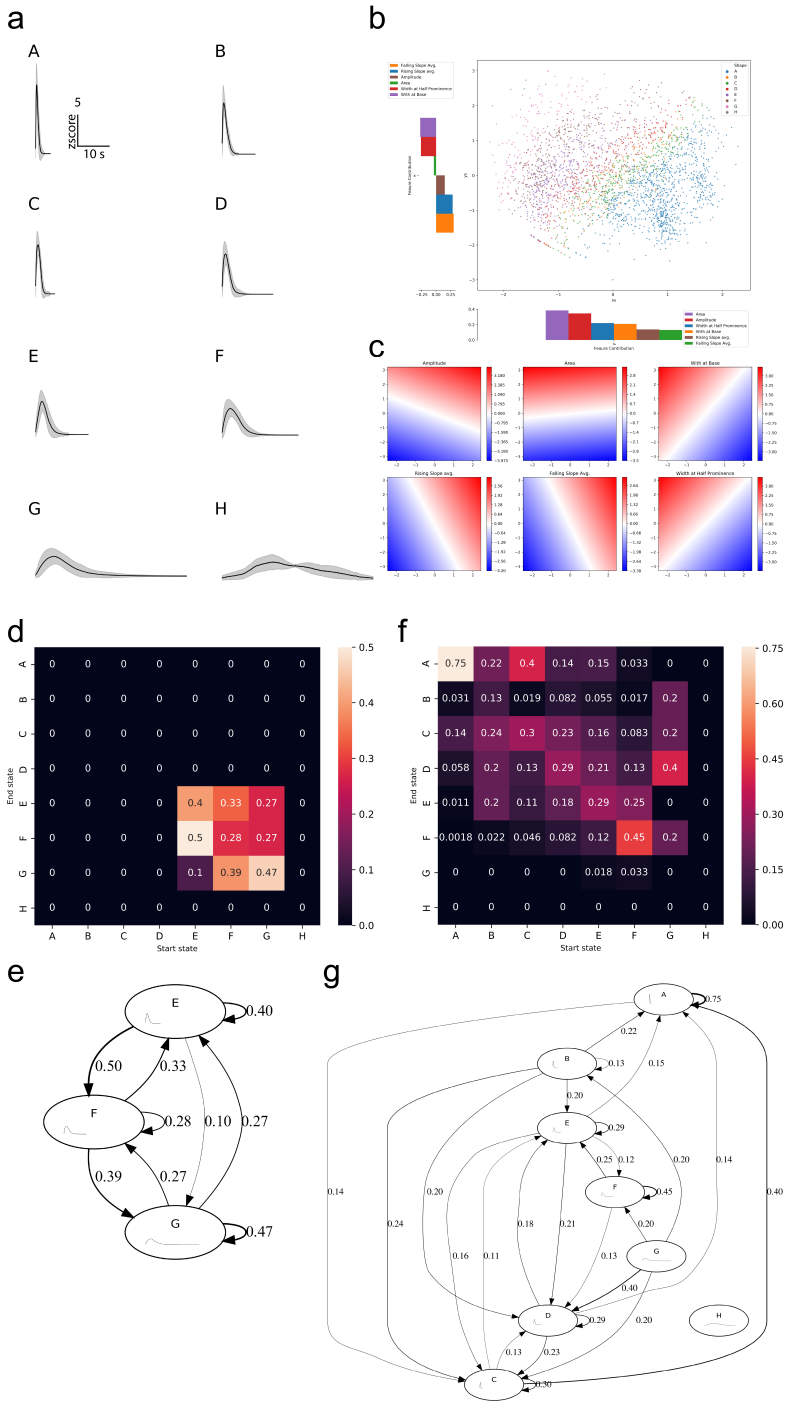
a
A
B
C
D
E
F
G
H
5
zscore
10 s
b
c
Amplitude
Area
Width at Base
Rising Slope avg.
Falling Slope Avg.
Width at Half Prominence
d
End state
Start state
A B C D E F G H
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0.4 0.33 0.27 0
0 0 0 0 0.5 0.28 0.27 0
0 0 0 0 0.1 0.39 0.47 0
0 0 0 0 0 0 0 0
0.5 0.4 0.3 0.2 0.1 0.0
f
End state
Start state
A B C D E F G H
0.75 0.22 0.4 0.14 0.15 0.033 0 0
0.031 0.13 0.019 0.082 0.055 0.017 0.2 0
0.14 0.24 0.3 0.23 0.16 0.083 0.2 0
0.058 0.2 0.13 0.29 0.21 0.13 0.4 0
0.011 0.2 0.11 0.18 0.29 0.25 0 0
-0.0018 0.022 0.046 0.082 0.12 0.45 0.2 0
0 0 0 0 0.018 0.033 0 0
0 0 0 0 0 0 0 0
0.75 0.60 0.45 0.30 0.15 0.00
e
E
F
G
0.40
0.50
0.33
0.28
0.10
0.27
0.39
0.27
0.47
g
A
B
C
D
E
F
G
H
0.75
0.22
0.13
0.15
0.20
0.29
0.25
0.12
0.20
0.14
0.45
0.14
0.20
0.40
0.24
0.18
0.21
0.20
0.16
0.11
0.13
0.40
0.29
0.20
0.13
0.23
0.30

Fig 6 | **k-Shape clustering and Markov Chains**
(a) Cluster means from k-Shape clustering of peaks from neuronal and non-neuronal cells in the head of *C. intestinalis*. Clusters are assigned alphabetical labels according to their half peak width. Error bands show within-cluster standard deviation. (b) PCA of peak-features showing how k-Shape clustering maps onto the PCA space. (c) Inverse transform for each of the input features showing the characteristics of the PCA space. (d) State Transition Matrix of a Markov Chain created from discretized sequences of HNK-1 cell calcium traces and the corresponding (e) state transition graph. (f) State Transition Matrix of a Markov Chain created from discretized sequences of photoreceptor cell calcium traces and the corresponding (g) state transition graph. Color scales in (d) and (f) are transition probability. Numbers on the transition graphs in (e) and (g) also show transition probability. Transition probabilities less than 0.1 were excluded to reduce visual clutter.