validation

1st INCF Workshop on

Validation of Analysis Methods

June 18 - 19, 2012
Stockholm, Sweden

incf

1st INCF Workshop on Validation of Analysis Methods

Date and Location

18 – 19 June 2012, Stockholm, Sweden

Organizers

Gaute Einevoll, Felix Franke, Thomas Wachtler, Daniel Wójcik

Authors

Michael Denker
Gaute Einevoll
Felix Franke
Sonja Grün
Espen Hagen
Jason Kerr
Martin Nawrot
Torbjørn Bækø Ness
Thomas Wachtler
Daniel Wójcik

Participants:

Michael Denker, Forschungszentrum Jülich, Germany

Gaute Einevoll, Norwegian University of Life Sciences, Norway

Felix Franke, ETH Zürich, Switzerland

David Greensberg, Max Planck Institute for Biological Cybernetics, Germany

Sonja Grün, Forschungszentrum Jülich, Germany

Espen Hagen, Norwegian University of Life Sciences, Norway

Jason Kerr, Max Planck Institute for Biological Cybernetics, Germany

Amir Khosrowshahi, Norwegian University of Life Sciences, Norway

Pilipp Meier, Technische Universität Berlin, Germany

Martin Nawrot, Freie Universität Berlin, Germany

Torbjørn Bækø Ness, Norwegian University of Life Sciences, Norway

Adrien Peyrache, New York University Medical Center, USA

Robert Pröpper, Technische Universität Berlin, Germany

Leslie Smith, University of Stirling, UK

Andrey Sobolev, Ludwig-Maximilians-Universität München, Germany

Thomas Wachtler, Ludwig-Maximilians-Universität München, Germany

Daniel Wójcik, Nencki Institute of Experimental Biology, Poland

INCF Secretariat

Sean Hill, Executive Director

Raphael Ritz, Scientific Officer

Contents

Executive Summary

To ensure that research results can be trusted, it is essential to use methods that are validated. This holds in particular for sharing of data and in the context of the large-scale concerted brain projects that are presently emerging both in Europe and in the US. This Workshop brought together scientists concerned with the validation of methods for data analysis from different perspectives.

The Workshop was motivated by a collaboration involving members of the Norwegian, Polish and German Nodes on establishing a community site for the evaluation of spike sorting methods (spike.g-node.org). However, the need for validation of data analysis methods is not restricted to spike sorting, but pertains to all physiological and anatomical measurement methods used in neuroscience. Besides spike sorting of extracellular recordings, the workshop addressed the extraction of spikes from two-photon calcium imaging, methods for analysis of local-field potentials, and statistical analysis of spike trains. In addition, analysis workflow management and documentation, which are crucial for validation of complex multi-stage analyses and thus an essential element of reproducible science, were discussed. An important additional measurement technique, functional magnetic resonance imaging (fMRI), was not considered extensively in the workshop, but it was understood by the Workshop participants that the issues, problems and needs identified for the other fields, as well as the conclusions and recommendations, would equally apply to the field of fMRI.

Typically, the efforts to define validation procedures for analysis methods, including the collection of benchmark data, start in single labs. However, they should be made available for use in the wide community. As requirements, benchmarks must be:

- Broadly accepted by a wide range of laboratories,
- Available to these laboratories, and
- Easily evaluated.

The Workshop participants discussed what is necessary to bring a validation effort from a prototype-like state, typically achieved in the initiating lab, to a community resource. This process must involve the initiating scientists, but also the community, and ideally an organization that has built up the expertise to support this process. An example of this scheme is the spike sorting validation project which has been enabled by a close collaboration of the involved scientists with the German INCF Node. The lesson learned from this example is that such a task should be taken on at the scale of an INCF program with expertise and support built up at the secretariat.

Discussing the different examples, the Workshop made clear that methods validation is at different stages for the various measurement techniques. Thus, specific recommendations differed, but overall the clear picture emerged that there is an unequivocal need for benchmarking activities. The participants agreed on the following overall key recommendations for supporting the various method validation initiatives:

1. Assure development and maintenance of the web site for validation of algorithms for spike sorting of electrical recordings hosted by the G-Node (spike.g-node.org).
2. Use the same technical resources to develop, host and maintain a corresponding website for validation of spike-detection algorithms for calcium imaging data.
3. Develop extensible framework allowing for validation of methods of analysis of other types of data.
4. Involve the community in the development of benchmarks and other means to validate methods for analysis.
5. Initiate and support training activities to educate users in methods validation.
6. Gather experts and users to discuss workflow standardization, and start activities to support reproducibility in data analysis.

1. Introduction

Our ideas and theories for how the nervous system works are typically formulated in terms of their cellular building blocks, that is, in terms of neurons, glia and their interactions. To test our ideas and generate new ones, we thus need to interpret measurements in terms of activity in these entities. For example, the firing of neuronal action potentials has for almost a century been assessed by measuring electrical potentials in the brain with sharp electrodes. Information about the spiking of neurons in the vicinity of the electrode tip is then contained in the high-frequency part of the recorded potential. However, this signal, called the multiunit activity (MUA), typically represents a mix of contributions from a handful of neighboring neurons, and the typical data analysis involves spike sorting, that is, the assignment of the various spikes seen in the MUA to specific neurons.

Such spike recordings continue to be a main workhorse in neurophysiology, and for outsiders it may come as a surprise that spike sorting involves a large manual component, even in the leading labs. This is not only

labor intensive, it also makes the sorting inherently irreproducible and idiosyncratic, for example, preventing proper comparison of recordings done in the various labs. Automatic algorithms are thus called for, not only to make the sorting reproducible (a key feature of the scientific method), but also to allow for the analysis of the vast amounts of data that can be recorded with the new generation of silicon-based multicontact electrodes allowing for long-time, simultaneous recording of spikes from hundreds or thousands of neurons (Buzsaki, Nature Neuroscience 7:446 (2004)).

Numerous automatic spike sorting algorithms have been developed, but lack of resources in the labs has typically prevented the facilitation of the methods to use by the general neuroscience community. Further, the methods have typically not been properly validated, making it difficult to assess their accuracy in various recording situations. Proper validation will require establishment of community-accepted benchmark data sets, that is, data sets where the ground-truth spike trains are known. Moreover, neuroinformatics infrastructure is needed to facilitate testing of candidate methods against these benchmark data (Einevoll et al., Current Opinion in Neurobiology 22:11 (2012)).

The need for validation of data analysis methods is certainly not restricted to spike sorting, but pertains for all physiological and anatomical measurement methods used in neuroscience. The belief that such validation is a natural part of the development and facilitation of key neuroinformatics infrastructure, was the backdrop for the arrangement of the 1st INCF Workshop on Validation of Data Analysis Methods in Stockholm on the 18th and 19th of June in 2012. Two prime examples were focused on: the aforementioned spike sorting from extracellular recordings and the extraction of spikes from two-photon calcium imaging. The former problem is already being addressed in an on-going collaboration between several national INCF nodes (spike.g-node.org), and several of the organizers got the impetus to suggest the workshop from working on this project. In addition, the workshop addressed validation of methods for analysis of local-field potentials and spike trains.

After this workshop several projects have been started and initiatives made that further shows the timeliness and importance of the validation issue, most prominently:

- The Human Brain Project (www.humanbrainproject.eu),
- The BRAIN initiative (formerly known as BAM (Brain Activity Map project)), and
- Project Mindscope at the Allen Brain Institute.

All these projects highlight the need for validated methods for automatic spike sorting and analysis of two-photon imaging data (see, for example, Alivastos et al., Nanotools for neuroscience and Brain Activity Mapping, ACSNano 7:1850-1866 (2013), DOI: 10.1021/nn401288k).

We therefore think that the topic addressed at the 1st INCF Workshop on Validation of Data-Analysis Methods and in the present report offers a unique opportunity for INCF to address a neuroinformatics infrastructure challenge that will likely be of large interest for the neuroscience community and thus highlight the importance of INCF.

2. Methods validation

Scientific findings are usually the final outcome of a long chain of work steps each of which can, potentially, influence the ultimate result. These steps encompass the design and execution of experiments, choice of equipment, data recording, storage, management and analysis, selection of data sets, compilation and presentation of results. A key requirement of the scientific method is reproducibility that imposes strong demands on every step. To allow for reproducibility and the possibility to compare results from different laboratories each of the steps must be well documented and standardized wherever possible. Data analysis, becoming more and more complex as more complicated algorithms and processing stages are involved, should be a prime focus of this standardization.

Standardization of data analysis can be done in two ways: either by the common use of the same algorithms/procedures and their implementations by many laboratories or, when this is not desired or possible, e.g. because a manual component is involved, by benchmarking in what way the use of different procedures and their implementations influence analysis results. For that purpose, it is essential that a benchmark, a standardized procedure to assess the quality and characteristics of a processing step, be:

- Broadly accepted by a wide range of laboratories,
- Available to these laboratories and
- Can be easily evaluated.

Only then the benchmark will be used and this way allow for the comparison and evaluation of results from different laboratories.

2.1 Spike Sorting

2.1.1 Background

Spike sorting is the estimation of the number of neurons and their electrical activity from extracellular recordings. It is a processing step that in itself encompasses many non-trivial sub-processing steps, e.g. prefiltering of the raw recordings, detection of action potentials in the recording, preprocessing of detected action potentials, estimation of the number of neurons and assignment of individual action potentials to neurons to name but a few (Lewicki, 1998; Einevoll, 2012). The result of spike sorting is the individual activity of single neurons called spike trains, i.e. the time points at which a given neuron fired an action potential. Since it is unknown which spike sorting algorithm works well on which data set, currently, spike sorting in most laboratories involves a large manual component. This is, as already mentioned in the introduction, not only labor intensive and thus costly but makes results idiosyncratic and hard to compare.

Furthermore, manual spike sorting can introduce large variance and errors (Wood, 2004) and, additionally, it was established that spike sorting errors could directly influence scientific results obtained from the estimated spike trains (Won, 2003; Pazienti and Grün, 2006).

A main reason for the confusing situation in the spike sorting literature is the lack of a commonly accepted benchmark data set - one of the more most prominent ones is the one from (Quiroga, 2004) but even that was so far only used by a handful of publications (Quiroga, 2004; Herbst, 2008; Franke, 2011; Matthews, 2011; Bestel, 2012). And even those used the benchmark in different ways (e.g. in (Quiroga et al., 2004) spike detection and classification was evaluated in separated steps, in (Franke, 2011) the whole data was used as an input using but the available ground truth was used to estimate a classifier and in (Herbst, 2008) the data was preprocessed using a wavelet technique).

It is, therefore, a very important undertaking to foster the creation, acceptance and availability of spike sorting benchmark data sets and, also, the procedures how to use those benchmarks. Triggered by the unsatisfactory situation any spike sorting evaluation has currently to face, an international collaboration (including the Norwegian, German, Polish and Swiss INCF nodes as well as the Technische Universität Berlin) has been formed that aims at improving the availability of spike sorting benchmarks and also suggested this workshop. This initiative addresses the issue from two sides:

- The creation of biophysically plausible benchmark data sets (see section 2.1.3) and,
- The development of an interactive website that will make it possible to collect and publish benchmark data sets and automatically evaluate spike sorting results on those benchmarks (see next section).

2.1.2 Website based spike sorting evaluation

In 2009 the idea was formed that it would be beneficial to collect available spike sorting benchmarks on a website. However, as already mentioned, a large fraction of spike sorting is manual work carried out in many labs with the help of commercial software or custom build scripts. That means the person who in fact does spike sorting will most likely be not a researcher working on spike sorting algorithms. To allow a spike sorting user to evaluate his own, manual, spike sorting it is not enough to just supply benchmarks. The actual evaluation process itself needs to be standardized and provided in an easily accessible way. This is why the website was planned to have three dynamic and interactive parts: First, it should be possible for users to supply their own benchmark data files to allow the benchmark collection to grow over time. Second, it should be possible for spike sorting users to upload their own spike sorting of one of the benchmarks and retrieve an automatic and standardized evaluation on his/her own performance. Third, it should be possible for the users to compare their own evaluation results to those of other users.

The idea was realized in the form of a prototype in the lab of Prof. Klaus Obermayer by a PhD student and two undergraduate students within 6 months; however, it was found that the developed prototype was, while fully functional, too complicated, hard to understand and thus not easily accessible by someone not working on the project. Furthermore, hosting of the website and its maintenance as well as the maintenance of the servers were found to take considerable resources not easily provided by students. The lesson learned was that, while being able to develop prototype systems, a research group at a university is not well suited to bring such a system to a user friendly, easy to use and sustainable level.

In other research contexts the "lifting" of a prototype to a usable product might be carried out with the help or directly by companies. But for a noncommercial website that has a relatively small number of researchers as target group, this was not an option. In this situation, support from INCF became essential. The German INCF Node provided infrastructure, funding, and implementation support to make the site available and a usable resource to the community (Franke, 2012). The site now runs at spike.g-node.org.

2.1.3 Benchmark Data Generation

Benchmark test data for evaluation of spike sorting methods with known ground truth, that is, complete knowledge of the action potentials generated by neuronal sources in vicinity of some recording electrode, is currently not something that can be recorded for any considerable amount of neurons (for a review, see Einevoll et al., 2012, Curr Op Neurobiol 22:11-17).

By employing a forward modeling scheme for extracellular field potentials using realistic biophysically detailed neuron models, LFPy (http://software.incf.org/software/lfpy), we generate artificial extracellular recordings from neuronal populations mimicking cases where spike sorting is commonly used to extract neuronal activity, such as hippocampal tetrode recordings, cortical polytrode recordings and slice recordings with multi-electrode arrays (MEA). As all the underlying neuronal activity in the model populations is known, the ground truth for all spiking units can be extracted, and used to evaluate the performance of different spike sorting algorithms.

As neuron models with active membrane dynamics are used in conjugation with synaptic events with times from spiking network models, the spike-time correlations and firing rates can be varied to different levels of complexity. Frequency dependent noise covariance and power spectral density is extracted from real recordings and used to generate noise with similar properties with an adjustable overall noise level. The test data can thus be made arbitrary difficult to spike sort in terms of the degree of overlapping spikes, inclusion of bursting neuron models, and noise level.

2.1.4 Recommendations regarding validation of methods for spike sorting

1. **Supporting the development of benchmarking sites from prototype to public platform.** Although initiatives and first developments for building benchmarking platforms originate in research labs which include staff with the required skills and expertise in the field, such work does not fit well with what is expected of university researchers. Moreover, designing such sites to be easy to use requires investment of time and effort in design and implementation. There is therefore a need for technical staff to build and maintain such platforms in collaboration with the active researchers and maintain them. We believe this could be a natural activity for the INCF Secretariat as (i) it assures the necessary long-term perspective and thus gives the web-site credibility in the eyes of the user and (ii) it will build up competence in operating science data bases at the INCF Secretariat which can be utilized in similar projects in the future (see, for example, section 2.2 below).

2. **Bringing together experts to share experience and coordinate development of benchmarks.** There are a number of projects internationally that have developed projects similar to benchmarking portals. INCF could organize workshops and conferences bringing these scientists together to share their experience. This would advance the development, and avoid errors being repeated. To achieve accepted benchmark standards, it is necessary to bring together researchers from the relevant user communities, in order that they can agree on a set of benchmarks. This will require a relatively large effort, because there are different types of problems faced by different parts of the community (for example, those working in vivo, or those working with slices, or those working with cultures, or (alternatively) those working with movable electrodes, vs those working with electrodes which cannot be moved). INCF would be in a unique position to guide such efforts.

3. **Raise visibility.** Benchmarks need to be publicly visible and recognized by their possible user community. INCF could play an important role in publicizing them as well as encouraging and facilitating their use.

Of these three actions, we consider no. 1 to be critical for the continuation of the overall validation endeavor. Without allocation of sufficient resources to assure a successful further development and maintenance of the website hosted by G-Node (spike.g-node.org), this INCF initiative will likely wither away.

2.2 Calcium Imaging

2.2.1 Background

Over the past 10 years, calcium imaging has emerged as a key tool for interrogating activity from neuronal populations *in vivo* (Stosiek, 2003; Kerr, 2005) for review see (Kerr and Denk, 2008). Using two-photon excitation (Denk, 1990), it has become possible to resolve activity in individual neurons and subcellular structures, even at considerable depth within the brain of a living animal (Dombeck, 2010; Mittmann, Wallace, 2011). Recently developed imaging techniques have also extended these capabilities to awake (Dombeck, 2007; Greenberg, 2008)) and freely moving animals (Sawinski, 2009).

Two-photon calcium imaging does not measure spiking in neurons directly. Instead, it allows observation of a change in the brightness of calcium indicators, small molecules or proteins whose fluorescence intensity depends on the concentration of calcium. Using chemical protecting groups to deliver small molecules (Stosiek, 2003) or viral vectors (Wallace, 2008) to deliver proteins (Tian, 2009), it is possible to stain many or all neurons in a local population, after which individual neurons appear clearly as distinct structures in the two-photon image. Detection of neuronal spiking using calcium imaging relies on the fact that the action potential is accompanied by an influx of calcium

through voltage-gated channels (Tsien and Tsien, 1990), which in turn leads to an increase in fluorescence intensity. Consequently, detecting neural activity using *in vivo* two-photon calcium imaging requires analyzing each neuron's fluorescence time series to determine the spike train responsible for the observed fluorescence changes (for review see: Kerr and Denk, 2008).

Compared to multiunit electrical techniques for detecting neural activity *in vivo*, two-photon calcium imaging possesses both advantages and disadvantages. While assigning detected spikes to individual neurons is a major challenge and the focus of much attention for extracellular recordings, calcium imaging's inherent spatial resolution makes this trivial. However, due to the need to scan a laser focus from one neuron to the next and the noisiness of optical signals emanating from deep within light scattering tissue (Helmchen and Denk, 2005), for calcium imaging it is more difficult to determine whether a spike has happened at all, or to determine spike times precisely (Kerr, 2007). Calcium imaging's spatial resolution also allows maps of spontaneous and stimulus-evoked neural activity to be constructed with single-neuron precision and compared to known anatomical features, while such precision is not possible with extracellular recordings. The real advantage is that imaging can sample from neurons that are rarely active in such that the dependence of detecting a neuron is not based on activity, unlike with 'blind' electrical recordings where spiking is sorted to 'detect' active neurons. Another advantage is that imaging from neuronal populations allows the sampling of subpopulations of neurons, such as interneurons and allows the exact anatomical position to be recorded. Lastly, when combined with permanent genetically encoded activity indicators, imaging allows for recording activity from the same populations of neurons from days to months. However, extracellular recordings can easily reach any brain area whereas noninvasive optical imaging penetrates about 1 mm at best, albeit the depth limit is theoretically well beyond this current value. Overall, calcium imaging and extracellular recording play complementary roles in understanding neural activity in the intact brain. This ability to simultaneously resolve neurons and glass electrodes can be exploited to produce 'ground truth data' by simultaneously recording activity from the neuron electrically and optically (Kerr, 2005).

2.2.2 Inferring Spike Trains from Calcium Imaging

Similar to the requirement for "spike sorting" methods to analyze extracellular recordings, calcium imaging requires "spike finding" analysis procedures that transform raw fluorescence data into estimates of neural activity. This necessity leads naturally to several important tasks in the development and evaluation of data processing algorithms for calcium imaging, several of which are analogous to the challenges for spike sorting method development mentioned above, but with the clear advantage that 'ground truth' data can be generated.

The challenges include:

- Metrics for evaluating the accuracy of an AP detection approach should be established and agreed upon.
- Ground truth datasets, in which both the true spike train and the observed fluorescence are available, must be collected and used to test AP detection algorithms (see section 'Ground Truth Data' below for details on these datasets).
- Algorithms for AP detection from fluorescence time series should be accurate, clearly described for reproducibility, and available.
- Determination of which experimental factors can influence the accuracy of AP detection methods or change their optimal parameters. These might include species, cell type, animal age, brain region, depth within light scattering tissue, choice of calcium indicator, fluorescence excitation wavelength, microscope configuration, etc.

Additional goals worth pursuing include:

- Software tools for AP detection should be integrated into new or existing software interfaces to provide an easy to use data processing pipeline, along with previous data processing steps such as motion correction (Dombeck, 2007; Greenberg, 2009) and subsequent steps such as spike train analysis (see below). Additionally, tools for low-latency online analysis of incoming fluorescence signals should be developed and integrated into microscope control software.
- Data formats for ground truth datasets (combined optical / electrical recordings) should be standardized, along with the form that the output of an AP detection algorithm should take. Ultimately we should strive for the scenario where all AP detection algorithms will run on all ground truth datasets without being "tailored" to them.
- Online tools should be made available to facilitate reproducibility of AP detection results. Ideally, it would be straightforward to test a new candidate AP detection method against a battery of ground truth datasets in an online database, or to quickly compare the results of multiple algorithms on a novel ground truth data set from a newly recorded cell type or calcium indicator. Also, a researcher not pursuing methods development but rather

hoping to apply a previously validated calcium indicator, detection algorithm or some other tool to go from *in vivo* experiment to inferred spike trains should have access to a database of extensive testing results.

2.2.3 Ground-Truth Data

For two-photon calcium imaging, ground truth data is available through simultaneous optical and electrical recording of the same neuron (Kerr, 2005). Electrical recording is performed using a glass micropipette to form usual tight seal with a single neuron of interest, using simultaneous two-photon imaging to guide the pipette visually onto the target neuron. Under these circumstances, spiking in the target neuron can be unambiguously determined. Meanwhile, the entire local population of neurons can be stained with calcium indicator as usual. Thus electrical and optical signatures of spiking can be acquired simultaneously in the same neuron.

Using this ground truth data, various algorithms can be tested for their ability to predict electrically recorded spike trains from optical signals alone. Various accuracy measures such as correlation over various time windows, detection rate and false positive rate, mutual information, and others can be explored.

Both the relationship between spiking and calcium and the relationship between calcium and fluorescence can vary across neurons. For example, the relationship between spiking and calcium is governed by such factors as the calcium influx per AP, the rate of calcium extrusion from the neuron, drift in baseline calcium levels, and indicator binding kinetics. Factors influencing the relationship between calcium and spiking include the gain factor of fluorescence measurements, the signal to noise ratio, the concentration of calcium indicator, etc. In light of this potential variability across neurons, it is necessary to compile a library of ground truth recordings that includes many neurons from multiple cell types. Consequently, to be effective an algorithm for AP detection should accurately measure spiking even when the cell type is unknown. This may involve estimating some or all of the aforementioned parameters along with the spike train.

While ground truth cell-attached data ultimately provides the authoritative means of validating an AP detection method, simulations are also likely to play an important role in the development and validation of AP detection algorithm. One advantage of simulations is that ground truth is available not only for the spike train, but also for the full generative model that produces the data. This allows the accuracy of parameter estimates to be evaluated along with the accuracy of spike train reconstruction. It is also possible to tune various aspects of simulated data such as SNR, etc. to gain a full sense of an algorithm's capabilities and limitations beyond the available ground truth data.

2.2.4 Establishing Data Sharing Practices

Optical AP detection is a relatively new technique that is nonetheless rapidly gaining widespread adoption with the development of new activity indicators and imaging approaches. As the methods and practices acceptable in published research become established, it is important to foster a culture of rigor and carefully validated methods. An essential first step toward establishing rigorous application of validated algorithms will be simply to make these algorithms freely available to use. The use of these algorithms will also be a starting point for collaboration and conversation between the theorists developing AP detection algorithms and the experimentalists applying them, which may provide the basis for further collaborations that may involve collection of new ground truth data. A suite of available AP detection algorithms will also aid the development of new calcium indicators, especially if these algorithms can estimate and adapt to a variable calcium-fluorescence relationship.

2.2.5 Recommendation regarding validation of methods for spike-detection from calcium signals

To foster the use and validation of algorithms for AP detection from optical signals, it will be necessary to host an online repository of algorithms, along with ground truth data for testing. In addition to hosting data and algorithms, such a resource could provide technical assistance to researchers using the software and could assist them in formatting ground truth data for upload. For example, an opt-in email list could alerts subscribers when a bug-fix or other important update is made to software that has been downloaded. As in the case of resources for the spike sorting problem, this is a task particularly ill suited to active researchers. INCF may therefore be uniquely positioned to make this resource available. Thus, our recommendation is that a website similar to the present spike.g-node.org for spike sorting electrical recordings is developed and maintained by the INCF Secretariat. This could be done in parallel with, or after, the finalization of a first fully operable version of the spike sorting site.

2.3 CSD and LFP

2.3.1 Background

Local field potential (LFP) is the low-frequency part of the extracellular electric potential. It is believed to reflect mainly the postsynaptic activity (Mitzdorf, 1985;

Pettersen, 2012), although other phenomena may also contribute (Buzsaki, 2012). The relationship between the extracellular potential and the transmembrane currents that generate it is given by the Poisson equation with appropriate boundary conditions. The advantage of LFP in neurophysiology is that it is easy to record stably over long time scales (on the order of months), much longer than is realistic for single units in any setup. Its disadvantage is that due to the long-range nature of the electric field the LFP reflects the neural population activity, which in certain cases can transmit over millimeters (Linden, 2011; Hunt, 2011), complicating its analysis.

Typical strategies in analysis of LFP are either direct analysis of signals, e.g. using Fourier methods, different tests for interaction between signals (linear and nonlinear cross-correlation variants), etc. An alternative is to first reconstruct the current sources which is called Current Source Density analysis (CSD; Mitzdorf, 1985; Pettersen, 2006; Łęski, 2007; Potworowski, 2012). CSD analysis requires multi-electrode recordings, but it renders quantities which are better localized in space and simplify the analysis.

In both approaches, through direct analysis of recordings and after CSD reconstruction, it is common to justify the results of the analysis post-hoc by reference to known underlying physiology. However, due to the complexity of tissue structure and the number of active structures, especially in vivo, this is a challenging task.

2.3.2 Ground-truth data

To advance our understanding of the slow part of the extracellular potential we should follow the same strategy as in case of spike sorting or inferring spikes from calcium imaging data. What we need is a stable environment that allows validation of methods of LFP analysis on reliable model data. Since the LFP is a population signal its proper simulation requires large-scale network models of cells with realistic morphologies. Optimally, one should take into account the physical properties of the tissue as well as the effect of the electrodes. Depending on the problem studied different level of realism of the model data might be required. For example, to test methods of signal decomposition, one can initially assume a network in a uniform, homogeneous, and infinite medium, only later including tissue and electrode models, ambient noise, etc. One could also consider hybrid models, with spiking networks used to set up dynamics and single neuron templates with realistic morphologies used to generate extracellular potential. These different approaches are under current studies. Combining network simulation of neurons with realistic morphologies, including realistic tissue and electrode models, is among the most challenging simulation types, as it requires combining compartmental modeling of cells with finite-element models of the field in the tissue. Such obtained data, properly validated, would be of great interest to theoreticians willing to test their data analysis methods, and for experimentalists wishing to verify their experimental conjectures.

One should decide what kind of data should be made available. This is both a conceptual and technical issue. As an example, if we wish to provide data for testing different CSD reconstruction methods, the ground-truth data set includes positions and morphologies of all neurons, as well as the distribution of trans-membrane currents over the cells in time. Even for moderate networks such as Traubs' model (~3500 cells, about 200 thousands compartments; Traub, 2005) and short runs (~100ms) the datasets count in hundreds of gigabytes. This leads to the issues of storage and accessibility organization. To make the data practical, one should provide also precomputed extracellular potentials. These can be raw or filtered (spikes, MUA, CSD), split for any cell, a population, the whole network. Practicality of data organization and accessibility should be considered, as well as metadata – description of the model and the "experimental" protocol.

2.3.3 Recommendations

1. **Provide a framework for the validation of methods of LFP analysis using ground-truth data similar in spirit to the facilities for spike sorting project.** To obtain a useful environment we need a repository for ground-truth (model) data. One could expand the spike sorting infrastructure or build on other initiatives (e.g. CRCNS). The data provided should be linked to models used for their generation deposited in ModelDB, Neuromorpho, etc. and provided with the necessary simulation protocols. Given the breadth of LFP analysis one can first focus on well-defined areas, such as CSD estimation, testing phase-synchrony, etc.

2. **Education.** Efforts showing the necessity to use methods validated on ground-truth data are needed.

2.4 Spike-Train Statistics

2.4.1 Background

Modern electrophysiology recording techniques (e.g. Buzsáki, 2004; Peyrache, 2012; Riehle, 2013) enable to record from hundreds of individual neurons at a time. Such data offer the chance to observe the concerted action in the cortical network, but they also pose new challenges for analysis approaches of such data. As long

as such recording techniques are used for enhancing effectively the number of recorded neurons while considering the neurons as independent entities, each of the neuronal responses can be analyzed independently for the rate responses one by one. However, if the interest is to get insight into interactions within the network, the neuronal activities have to be analyzed for their correlated activities. This poses new challenges for the analysis of data. In contrast to estimates of rate responses that are quite tolerant against individual statistical properties of the neurons and their potential variability across trials, the estimation of correlations across neurons depends crucially on such properties. In order to distinguish a) spike correlations due to network interaction from b) chance correlations as a trivial by-product of (changes of) firing rates of the neurons, one has to perform a statistical test. For doing that empirical spike pattern counts are compared (in the simplest case) to the expected counts given the firing rates (Grün, 2002a), or an even more complicated null-hypothesis that e.g. also considers sub-correlations in order to test if higher-order correlations are present (Staude, 2010a,b; Shimazaki, 2012). However, even in the former case, the exact formulation of the null-hypothesis is crucial. If the assumptions included in the formulation of the null-hypothesis, e.g. that the neuronal spike trains follow a Poisson process, or/and are stationary in time and/or across-trials, are violated, there is the danger of false positive outcomes and/or wrong interpretation of the data (Grün, 2009). Thus, in order to avoid this it is recommended (i) to include the statistical properties in the null-hypothesis (Grün, 2002b; Grün, 2003; Grün, 2009; Pipa, 2013) and (ii) to validate the analysis method to make sure that these goals are achieved or that at least the method operates in a regime that does not cause harm (Grün, 2010).

The significance of spike correlation may be performed either in a parametric or in a non-parametric test. In this context, a parametric test implies to choose a distribution for the significance test that can be parameterized by the requested mean. The mean may be calculated as the product of the firing probabilities of the neurons (statistical independence; Grün, 2002a). Estimation of the firing rates of the neurons is easy to do, as long as e.g. the spike trains are stationary and Poisson. However, including more complex properties of the data, e.g. non-stationary firing rates or an interval statistics that deviates from Poisson, is a much more difficult task (Pipa, 2013 that requires reliable estimation of these properties.

However, the analytical derivation of the expected mean of the number of spike patterns or even more properties of the distribution turned out to be difficult. To avoid this one may instead choose a non-parametric approach by deriving the null-hypothesis and thus the distribution for the significance test by surrogate data. This is done in the following way. Surrogate data are artificially generated data, which have the same statistical properties as the to be tested experimental data, but the property that is tested for is not present in this data. Thus, in our context parallel surrogate data are generated that do not express spike correlations. These data are then analyzed in the same way as the experimental data for spike correlations, e.g. the number of spike synchrony patterns. The surrogate data are generated many times (e.g. 1000 times, to test for a significance level of 1%), and the expected distribution of spike pattern counts under is generated. In experimental data, the significance of the empirical count of spike synchrony patterns is derived by comparing the count to the derived distribution to measure the p-value (Louis, 2010a,b; Berger, 2010).

2.4.2 Approaches

The surrogate data can be generated in quite different ways, ranging from manipulating the original data (e.g. by shuffling spike times, or dithering the spike times, or trial shuffling, etc) to generating the spike trains according to specific stochastic models, all with the aim to destroy potentially correlated spike timing across the neurons (see for a summary Grün, 2009). However, this process requests for particular care, since a manipulation of spike trains is typically not only affecting / destroying spike correlation, but also affects other properties of the data. For example, spike time randomization applied to individual spike trains effectively destroys spike correlation across neurons, but it also flattens firing rate modulations and changes the inter-spike interval statistics. This in turn may lead to the conclusion that the original spike trains were correlated although they were not (i.e. generate false positives). However, the effects of the manipulation of the spike trains may also depend on the specific analysis method used or on data preprocessing (Louis, 2010a,b); therefore, the surrogate methods need to be validated in the context of the chosen analysis method.

Validating methods on experimental data is not an option, since the ground truth, here the underlying correlation structure, is not known; therefore, we suggested to use artificial simultaneous spike trains which are realized according to a well defined correlation model, and to test if the analysis method (including the surrogates) extract well the incorporated correlation structure, or if it generates false positive or false negative outcomes. For doing that one can generate data according to stochastic models (Kuhn, 2003; Pazienti and Grün, 2006; Pazienti, 2008; Staude, 2010a,b) or network simulations (Schrader, 2008; Gerstein, 2012) with known correlation structures, embedded assemblies and correlation strength. Other features typically found in experimental spike trains can and should also be incorporated to test realistic cases. Trivially, a well working analysis method should extract the incorporated spike correlation, no matter

what other features the spike trains have (Berger, 2010; Schrader, 2008; Staude, 2010b; Gerstein, 2012; Torre, 2013).

2.4.3 Recommendations

1. **Establish Collaborative websites**. Such website should serve for providing software, literature, test data of known ground truth, tools for generating test data, data analysis tools, exchange of experience (blogs).

2. **Education / Courses.** Essential requirement for the useful application of statistical analysis methods is appropriate training, including teaching of best practices. We learned over the last years in which we were organizing or involved in data analysis courses (NWG course: 'Analysis and models in neurophysiology, yearly, 4-5 days, since 2003), or the G-Node data analysis course (yearly, 1 week, since 2009), Berkeley CNCRS data analysis course (yearly, since 2010) that courses composed of lectures and strongly guided hands-on-work as done in the above mentioned courses is not enough to educate on data analysis. Therefore we recommend courses for data analysis where besides lectures people can bring their own data to analyze them under supervision. This requires longer courses, e.g. for 2-4 weeks with intense tutoring, e.g. as done for the 'Advanced course in computational neuroscience' in the field of modeling and simulations. The Wood's whole course 'Neuroinformatics' has a related concept, but tutoring was very sparse.

2.5 Providing data analysis services: Experiences from CARMEN

In the UK, the CARMEN project (see www.carmen.org.uk) has attempted to enable portal-based cross-laboratory working by enabling upload of electrophysiological datasets and services for processing this time-series data. One aspect of this has been services that enable spike detection and spike sorting. This project (which involves 11 UK Universities) started in October 2006, funded by the UK EPSRC, and continues, now funded by the UK BBSRC. The project's aims were lofty: "to deliver a virtual laboratory for neurophysiology, enabling sharing and collaborative exploitation of data, analysis code and expertise". But this has not been easy: The datasets were not all of one type, and even although the project used the Neuroshare libraries to enable conversion, creating services that are not tied to specific data types was difficult. As a result, a new data format, NDTF (see: www.carmen.org.uk/standards/CarmenDataSpecs.pdf) was created, with services made available for converting uploaded datasets in various proprietary formats into NDTF. NDTF is a container format: it not only contains time series data for various formats, but also records metadata. Indeed, one of the aims of CARMEN was to have the metadata about the experiment creating the electrophysiological datasets uploaded along with the data itself, and for NDTF to maintain this metadata, and to add to it information about what additional processing had occurred, so that there would be an "audit trail" enabling reproducibility.

The current situation of CARMEN is that the services enabling a variety of spike sorting techniques to be tested on a particular piece of data exist: there are high pass filters, spike detectors and at least one variety of spike sorter. Unfortunately, there have been problems implementing workflows in the system, making running multiple tests of different spike sorters time-consuming. (Note that these particular problems should, we hope be ironed out soon.) The technical problems have not been the only ones: it has proven surprisingly difficult to get a large volume of good electrophysiological data to be shared publicly by neurophysiologists. CARMEN permits data to be shared with any subset of users, and most of the users of CARMEN use it to share data between geographically distributed research teams. Electrophysiologists seem unwilling to provide the volume of metadata that would really be required to make their datasets effectively reusable, even although considerable efforts have been made to make the upload of such data straightforward. The automated documentation produced by the NDTF system should be useful for reproducibility, but without large volumes of good electrophysiological data, (and perhaps because of the rather difficult current user interface) there has not been much work on trying out different techniques on multiple datasets.

It is clear that trying to use university research resources to develop and maintain these types of services is difficult: University staff are generally assessed on their research publications, and not on the effectiveness of the tools they produce. The people we need to make these systems work are programmers, who have an interest and an understanding of the issues involved, but not really researchers developing state-of-the-art novel techniques. Yes, we want to include novel techniques as services, but we also need to simply make the system work effectively, with a good user interface, and to be able to run the types of workflows that make utilization of the existing services and datasets effective. It is now six years since the project started: quite a considerable amount of use has been made of CARMEN by specific projects, but much less use has been made of the large-scale data-sharing capabilities. However, now that the NDTF is working, and now that there are a good set of services that read and write NDTF data, we are hopeful that, once workflows are finally available, larger scale use will be made of the system.

3. Workflow Documentation/ Reproducibility

3.1 Background

In recent years the complexity of electrophysiological experiments has steadily increased. This complexity arises firstly from the interest in simultaneously analyzing the activity recorded from large numbers of channels in order to investigate the role of concerted neural activity in brain function, as opposed to classifying the responses of individual neurons in isolation. These efforts have led to advances in data analysis methods (Brown, 2004) that exploit the parallel properties of such data sets (Stevenson, 2011). A second source of complexity is in the sophistication of stimulus protocols. To take the visual system as an example, typical visual stimulation has progressed from simple moving bars or drifting gratings to natural movies, Gabor noise, and apparent motion stimuli. Another example, in the somatosensory system, is that new technology now allows the entire rodent whisker array to be stimulated in essentially arbitrary patterns, yielding a high-dimensional stimulus space (Jacob, 2010). However, an often neglected aspect of these technological advances is that both massively parallel data streams and highly complex stimuli place new demands on handling their complexity during all stages of the project (Buzsáki, 2004): from the initial recording, throughout the analysis process, to the final publication.

In particular, the documentation of the recording and analysis process becomes highly involved. For each step it is necessary not only to track in detail what and how something was done (both of which could be automated to some degree by software), but also to record or later reference the reasons for performing certain steps and for making certain choices in the design of these steps. The latter is typically a tedious process for the researcher in itself, but as the complexity of the analysis increases, becomes error-prone, if not virtually impossible, without the aid of standards and tools that can structure and simplify the analysis process. As a result, without such documentation, thorough reproducibility cannot be guaranteed.

The workflow describes this sequence: how experiments and data analysis are performed, how they are dependent on one another, and where they lead. Logically, re-doing a data analysis workflow on the same data sets should produce the same results. Thus, in principle, the problem of assuring reproducibility of data recording and analysis is tightly related to the establishment of well-defined workflows. However, there may be sources of variability - and in the case of electrophysiology, this variability can be quite considerable if one takes into account, e.g., repeating a complete behavioral experiment with the same workflow, the uncertainties of some processing steps such as spike sorting, analysis parallelization or the fragility of higher-order analysis methods with respect to parameter variations (cf., Fig. 1). Moreover, this workflow should not be seen in a purely computer science centered view, as a certain sequence of analysis steps to be performed by the computer. The workflow encompasses a much larger range of facets, such as animal protocols (e.g. anesthesia used, perfusion details, etc.), recording and quantifying animal training, organizing data and sharing it with team members. In fact, given the need for collaborative work on complicated data sets, it may even include social aspects, detailing how collaborators divide work and share information effectively and without getting in each other's way. In general, keeping detailed track of all these parts of the work and their dependencies is critical for preventing a situation where scientific results can no longer be reproduced and thus lose their scientific value.

In order to describe and share workflows or establish them to assist a collaboration between laboratories, it becomes necessary to define best practices and guidelines that can be easily implemented. One manner to describe a basic workflow is state-based – considering state machines with transitions, branching, merging: logically these are finite state automata. However, what we are concerned with here are activity-based workflows: sequences of steps, with choice and iteration. That is, there is a sequence of actions (steps), each labeled (perhaps with a name, and a set of parameters that apply to that step): there can also be choices, that is the selection of the next step depends on some local values (perhaps on the result of previous steps), and iteration (that is, one can loop around a sequence of steps, possibly altering the parameters used at each step in a controlled way). Given the complexity of recording and analysis witnessed today, the deviations from a purely linear workflow are quite considerable in practice. Conceptually, workflows in computing have been around for a *very* long time, going all the way back to the job control languages of 1970's mainframes.

Sharing workflows across groups of researchers requires first and foremost some form of standardisation of format and terminology. Frameworks for dealing with these are coming into existence: W3C is proposing an ML based technique. Such tools imply data models, and these need to be agreed across collaborators and communities. Conceptually, they should enable, for example, that one lab performs an electrophysiological investigation: and another lab can easily re-use it, and even modify it. A successful example of the implementation of such a standard data model is the NineML/NeuroML (software.incf.org/software/nineml/home/; www.neuroml.org/) for specification systems of modelling techniques: it provides an interoperability standard that enables a generic reproducibility. To be successful, data models and formats describing the workflow, data and metadata need to be supported by all relevant components in the tool chain. Indeed, the likelihood of their general adoption by the community also depends on manufacturers supporting it, perhaps in the same style as

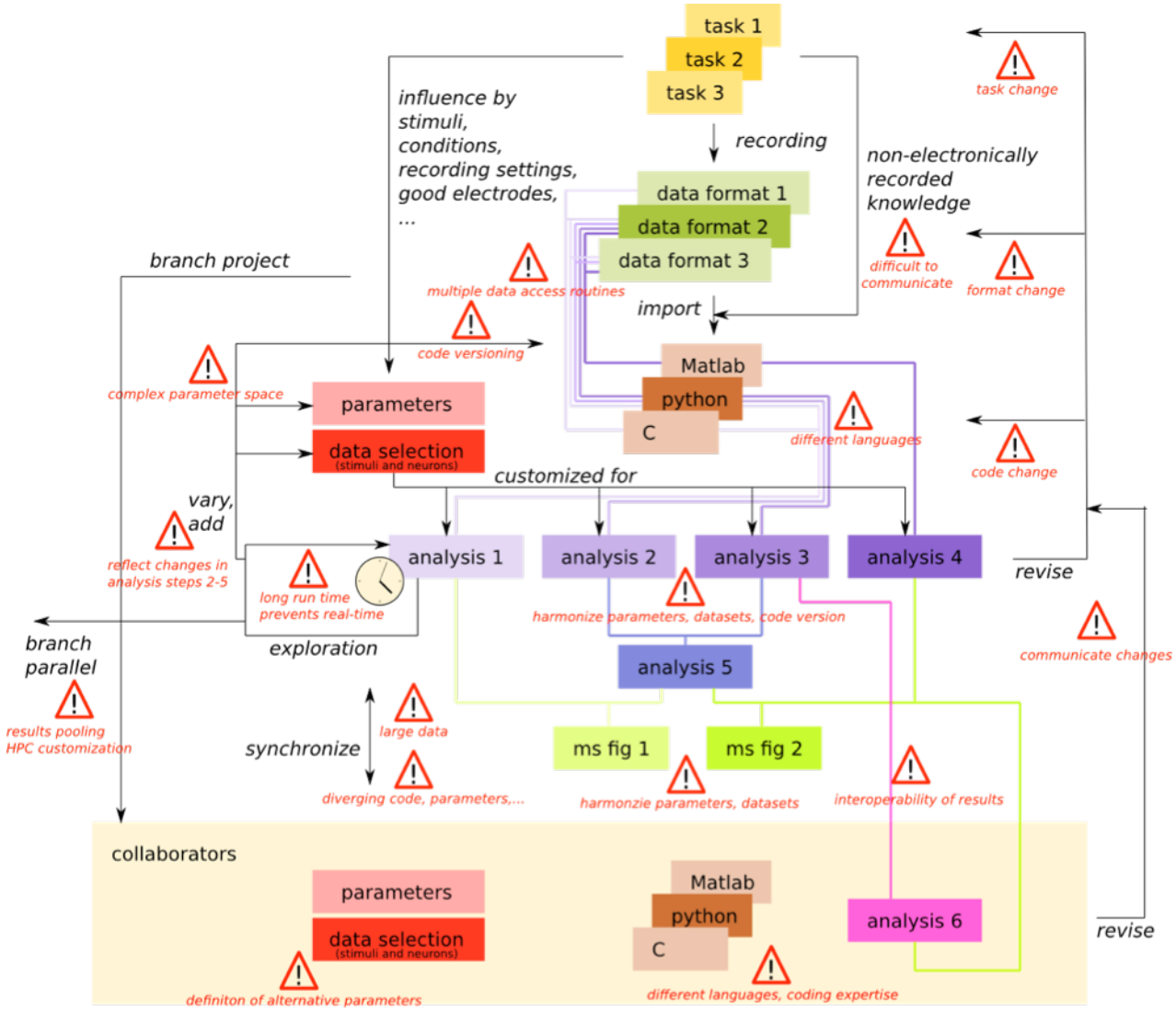

Figure 1. *A graphical representation of a workflow in electrophysiology from experiment (top) to publication (bottom) illustrates its complicated nature. Exclamation marks indicate critical aspects of the workflow where support by tools and guidelines is in demand (figure: Michael Denker, Jülich).*

the manufacturers of electrophysiology equipment support the Neuroshare libraries. Moreover, if it is to be an open source, community driven project care must be taken that the developers are actively supported.

Care must be taken to prevent over-engineering. Any standards on workflow will only receive acceptance if they are useful to the researchers and easy to implement. A simple case is to consider a workflow as a record of a simple sequence of steps in spoken language. In fact, this is a process similar to what should be done in labs nowadays by filling laboratory notebooks and journals, and represents in many cases the *status quo*. While this method may seem reasonably straightforward for researchers, there are two major caveats:

- Current book keeping methods may not even necessarily be sufficient to enable reproducibility locally within the laboratory. Automated agglomeration of meta-data during the project can help reduce errors, and increase the wealth of information recorded. Both factors act positively towards better reproducibility of experiment and analysis.
- To facilitate general reproducibility, in the sense that results could be reproduced in a different laboratory, more generic standards need to be created which allow the unambiguous interpretation of metadata and annotations by third parties.

The designers of the goal of proposed workflow guidelines that attempt to improve on these two points should always keep in mind that the additional overhead created by workflow sophistication should also reflect immediate usefulness to the researchers in their daily routine.

Yet another matter of concern in designing workflow guidelines for electrophysiology lies in the diverse levels

of descriptions required to capture the complete project. At a low level, the description is simply about how each step is carried out. At a higher level, there should be some sort of motivation - something of the "why" the workflow is as it is. The latter part becomes especially important for a number of reasons unique to electrophysiological investigations, e.g., the complexity of animal training in behavioral paradigms, or the conceptual challenges faced with when analyzing massively parallel datasets. Getting tool manufacturers to integrate their products into proposed workflow guidelines requires a clear understanding of this multi-level description of the workflow.

As the analysis of neuronal data becomes increasingly more difficult, both due to advanced, parallel and intensive computational methods and due to the data complexity, it will become vital to share analysis code between researchers. In neuronal modeling, similar problems were encountered when trying to compare results obtained by independent modeling efforts. Experience in this field has demonstrated that infrastructure that enables the sharing of methodology (in this case via efforts such as *ModelDB*; senselab.med.yale.edu/modeldb) vastly improves the ability to reproduce and cross-validate results, and is seen as of great benefit to the researcher.

3.2 Approaches and tools in need

1. Systems to help manage workflow and provenance information, all the way from experimenter to analysis. Examples for such systems (partly from other scientific communities) are Sumatra (neuralensemble.org/trac/sumatra), Kepler (kepler-project.org/), Taverna (www.taverna.org.uk/), Vistrails (www.vistrails.org/), Galaxy (galaxy.psu.edu/), LabLog (lablog.sourceforge.net/). Such systems need to include the ability to deal with large data sets, to be flexible and non-obtrusive, easy to set up, and to utilize a common description level.

2. Services to aid researchers in validating their methods, using calibrated models and data with known properties. A good example of how this may look, are the efforts on validating spike sorting techniques outlined earlier in this report.

3. Better interoperability of the multitude of tools available to the researchers, enabling in particular a smoother interaction between collaborators. Transparency of the workflow is essential for all team members to better grasp data and analysis steps in their completeness, leading to more data robust analysis. In the field of electrophysiology, already the rivalry between the two most widely used programming languages (Matlab and python) is one such challenge.

3.3 Recommendations

1. **Collect experience from the community.** Contact researchers who have tried using the different existing systems to manage workflows and get their reports. What tools exist and how useful are they?

2. **Organize workshops** to discuss workflow issues with electrophysiologists, data analysts, tool developers, and people in **other communities** who are further ahead in workflow implementation and usage. As a concrete suggestion, a first workshop could be help in combination with a planned workshop on these topics in the context of the EU funded BrainScaleS initiative at the Institute of Neuroscience and Medicine (INM-6), at Forschungszentrum Jülich, Germany in 2013 **Promote interoperability of analysis tools**. INCF should initiate and coordinate activities to develop standards for formats and interfaces to facilitate integration of the multitude of tools and data formats.

3. **Support tool development**. INCF should identify which tools need to be developed to help researchers organize and document data analysis workflows. The possibility to collaborate with commercial e-Lab book developers should be considered, or the option of combining such efforts with a *Google summer of code* project.

4. **Education.** Awareness of workflow issues and their importance should be taught in courses. Perhaps this can be achieved using specific practical tools, so that students immediately see benefits of using such approaches (Example: version control in python using git or subversion). These efforts would teach the younger generation best practices early on, help researchers to overcome initial difficulties in using such tools, and finally carry the theoretical concepts of workflow management to the individual labs in a practical form. Courses funded by INCF should be encouraged to include such training components.

5. **Establish collaborative websites**. These websites could host an ecosystem of tools that is required for the easy implementation of workflows (example: the python ecosystem of tools being developed for the modeling community). However, this requires interoperability of all the elements of the workflow.

4. Conclusions and key overall recommendations

The Workshop has shown that validation of data analysis methods is an issue in many fields. Accordingly, this report contains numerous recommendations. We therefore reiterate what we consider to be the main conclusions and key recommendations. The recommendations strongly indicate the need of a coordinated effort to establish benchmarks and sites for validation for different types of data. A leading role of INCF would avoid double work and guarantee standardized processes. **We recommend that INCF establishes dedicated activities to support validation initiatives by providing expertise, resources, and coordination, including:**

- Support the development of validation websites from prototype to public platform and maintain them. This requires expertise and resources to build sites that provide the required services and are easy to use. Interaction with the active researchers is essential
- Support the development of benchmarks for method validation. Bring together experts in the respective fields to define and establish benchmarks
- Promote method validation by raising awareness in the community, encouraging use of benchmarking sites, and supporting training activities

The topics covered at this workshop present different use cases providing specific goals for these activities.

The main recommendations are:

1. Assure development and maintenance of the web site for validation of algorithms for spike sorting of electrical recordings (spike.g-node.org), preferably by committing technical staff at the INCF Secretariat to further develop and maintain the site hosted by the German INCF Node (see section 2.1).
2. Use the same technical resources to develop, host and maintain a corresponding website for validation of spike-detection algorithms for calcium imaging data (see section 2.2).
3. Develop extensible framework allowing for validation of methods of analysis of other types of data or in other context using ground-truth data (e.g. for LFP - CSD analysis, phase-synchrony, etc.; similar facilities for spike-train, EEG, MRI analysis, etc; see sections 2.3-2.4).
4. Involve the community in the development of benchmarks and other means to validate methods for analysis in these fields (see sections 2.1-2.4)
5. Initiate and support training activities to educate users how to validate the methods on ground-truth data before using them (see sections 2.1-2.4).
6. Gather experts and users, for example at INCF workshops, to discuss workflow standardization, and start activities to support reproducibility in data analysis (see section 3).

References

- Berger, D., Borgelt, C., Louis, S., Morrison ,A., and Grün, S. (2010). Efficient Identification of Assembly Neurons within Massively Parallel Spike Trains., Computational Intelligence and Neuroscience Vol. 2010, Article ID 439648, 18 pages. DOI:10.1155/2010/439648
- Brown, E.N., Kass, R.E., and Mitra, P.P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. Nat Neurosci, **7:**456-461.
- Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. Nat Neurosci, **7:**446-451.
- Buzsáki, G., Anastassiou, C. A, and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, *13*(6), 407–420.
- Denk, W., Strickler, J. H., and Webb, W. W. (1990). Two-photon laser scanning fluorescence microscopy. *Science* **248**: 73-76.
- Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L., and Tank, D. W. (2007). Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56**: 43-57.
- Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L., and Tank, D. W. (2010). Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat Neurosci* **13**: 1433-1440.
- Einevoll, G. T., Franke, F., Hagen, E., Pouzat, C., and Harris, K. D. (2011). Towards reliable spike-train recordings from thousands of neurons with multielectrodes. Current Opinion in Neurobiology *27*, 1–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22023727.
- Franke, F. (2011). Real-Time Analysis of Extracellular Multielectrode Recordings (Berlin: PhD Thesis, Technische Universität Berlin) Available at: http://opus.kobv.de/tuberlin/volltexte/2012/3387/.
- Franke F, Meier P, Sobolev A, Hagen E, Hierlemann A, Obermayer K, Einevoll GT, Wachtler T (2012) Automatic spike sorting evaluation: A website based community approach. 5th INCF Congress of Neuroinformatics, Sep 10-12, Munich. http://www.neuroinformatics2012.org/abstracts/automatic-spike sorting-evaluation- a-website-based-community-approach
- Gerstein, G.L., Williams, E.R., Diesmann, M., Grün, S., Trengove, C. (2012). Detecting synfire chains in parallel spike data. J Neurosci Methods 206 (1): 54-64 doi: 10.1016/j.jneumeth.2012.02.003
- Greenberg, D. S., Houweling, A. R., and Kerr, J. N. (2008). Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat Neurosci* **11**: 749-751.
- Grün, S., Diesmann, M., and Aertsen, A. (2002). `Unitary Events' in multiple single-neuron activity. I. Detection and significance. Neural Computation 14(1): 43-80.
- Grün, S., Diesmann, M., and Aertsen, A. (2002). `Unitary Events' in multiple single-neuron activity. II. Non-Stationary data. Neural Computation 14(1): 81-119.
- Grün, S., Riehle, A., and Diesmann, M. (2003). Effect of cross-trial nonstationarity on joint-spike events. Biological Cybernetics 88(5): 335-351.
- Grün, S. (2009). Data-driven significance estimation of precise spike correlation. J Neurophysiology, 101: 1126-1140 (invited review)
- Grün, S., Diesmann, M., Aertsen, A. (2010). Unitary Event Analysis In: Analysis of parallel spike trains. eds.
- Grün and Rotter. Springer Series in Computational Neuroscience, DOI:10.1007/978-1-4419-5675-0_10
- Helmchen, F. and Denk, W. (2005). Deep tissue two-photon microscopy. *Nat Methods* **2**: 932-940.
- Herbst, J. a, Gammeter, S., Ferrero, D., and Hahnloser, R. H. R. (2008). Spike sorting with hidden Markov models. Journal of neuroscience methods *174*, 126–134. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18619490 [Accessed July 12, 2011].
- Hunt, M. J., Falinska, M., Łęski, S., Wójcik, D. K., and Kasicki, S. (2010). Differential effects produced by ketamine on oscillatory activity recorded in the rat hippocampus, dorsal striatum and nucleus accumbens. *Journal of psychopharmacology. 25(6)*, 808-21.
- Jacob, V., Estebanez, L., Le Cam, J., Tiercelin, J.Y., Parra, P., Parésys, G., Shulz, D. (2010). The Matrix: a new tool for probing the whisker-to-barrel system with natural stimuli. J Neurosci Methods, **189:**65-74.

- Kerr, J. N., Greenberg, D., and Helmchen, F. (2005). Imaging input and output of neocortical networks in vivo. *Proc Natl Acad Sci U S A* **102**: 14063-14068.
- Kerr, J. N., de Kock, C. P., Greenberg, D. S., Bruno, R. M., Sakmann, B., and Helmchen, F. (2007). Spatial organization of neuronal population responses in layer 2/3 of rat barrel cortex. *J Neurosci* **27**: 13316-13328.
- Kerr, J. N. and Denk, W. (2008). Imaging in vivo: watching the brain in action. *Nat Rev Neurosci* **9**: 195-205.
- Kuhn, A., Aertsen, A., and Rotter, S. (2003) Higher-order statistics of input ensembles and the response of simple model neurons. Neural Comput. 15(1):67-101
- Lewicki, M. S. (1998). A review of methods for spike sorting: the detection and classification of neural action potentials. Network (Bristol, England) *9*, R53–78. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10221571.
- Lindén, H., Tetzlaff, T., Potjans, T. C., Pettersen, K. H., Grün, S., Diesmann, M., and Einevoll, G. T. (2011). Modeling the Spatial Reach of the LFP. *Neuron*, *72*(5), 859–872.
- Louis, S., Gerstein, G.L., Grün, S., Diesmann, M. (2010a). Surrogate spike train generation through dithering in operational time. Front. Comput. Neurosci. 4: 127, doi: 10.3389/fncom.2010.00127
- Louis, S., Borgelt, C., Grün, S. (2010b). Generation and selection of surrogate methods for correlation analysis. In: Analysis of parallel spike trains. eds. Grün & Rotter. Springer Series in Computational Neuroscience
- Łęski, S., Wójcik, D. K., Tereszczuk, J., Świejkowski, D. A., Kublik, E., and Wróbel, A. (2007). Inverse current-source density method in 3D: reconstruction fidelity, boundary effects, and influence of distant sources. *Neuroinformatics*, *5*(4), 207–22.
- Matthews B, Clements M (2011). Joint modeling of observed inter-arrival times and waveform data with multiple hidden states for neural spike sorting. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, p. 637–640.
- Mittmann, W., Wallace, D. J., Czubayko, U., Herb, J. T., Schaefer, A. T., Looger, L. L., et al. (2011). Two-photon calcium imaging of evoked activity from L5 somatosensory neurons in vivo. *Nat Neurosci* **14**: 1089-1093.
- Mitzdorf, U. (1985). Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. *Physiological reviews*, *65*(1), 37.
- Pazienti, A., and Grün, S. (2006). Robustness of the significance of spike synchrony with respect to sorting errors. Journal of computational neuroscience *21*, 329–342. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16927209 [Accessed August 10, 2012].
- Pazienti, A., Maldonado, P., Diesmann, M., Grün, S. (2008). Effectiveness of systematic spike dithering depends on the precision of cortical synchronization Brain Research, 1225: 39-46.
- Pettersen, K. H., Devor, A., Ulbert, I., Dale, A. M., and Einevoll, G. T. (2006). Current-source density estimation based on inversion of electrostatic forward solution: effects of finite extent of neuronal activity and conductivity discontinuities. *Journal of neuroscience methods*, *154*(1-2), 116–33.
- Pettersen, K. H., Linden, H., Dale, A. M., and Einevoll, G. T. (2012). Extracellular spikes & multi-electrode recordings. *Handbook of Neural Activity Measurement*.
- Peyrache, A., Dehghani, N., Eskandar, E.N., Madsen, J.R., Anderson, W.S., Donoghue, J.A., Hochberg, L.R., Halgren, E., Cash, S.S., and Destexhe, A. (2012). Spatiotemporal dynamics of neocortical excitation and inhibition during human sleep. Proc Natl Acad Sci U S A. 2012 Jan 31;109(5):1731-6. doi: 10.1073/pnas.1109895109.
- Pipa, G., Grün, S., and van Vreeswijk, C. (2013). Impact of spike-train autostructure on probability distribution of joint-spike events. Neural Comput. 25(5): 1123-1163 doi: 10.1162/NECO_a_00432
- Potworowski, J., Jakuczun, W., Łęski, S., and Wójcik, D. K. (2012). Kernel current source density method. *Neural computation*, *24*(2), 541–75.
- Quiroga, R. Q., Nadasdy, Z., and Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural computation *16*, 1661–1687. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15228749.
- Riehle, A., Wirtssohn, S., Grün, S., and Brochier, T. (2013). Mapping the spatio-temporal structure of motor cortical LFP and spiking activities during reach-to-grasp movements. Front. Neural Circuits 7: 48 doi: 10.3389/fncir.2013.00048

- Sawinski, J., Wallace, D. J., Greenberg, D. S., Grossmann, S., Denk, W., and Kerr, J. N. (2009). Visually evoked activity in cortical cells imaged in freely moving animals. *Proc Natl Acad Sci U S A* **106**: 19557-19562.
- Schrader, S., Grün, S., Diesmann, M., and Gerstein, G. (2008). Detecting synfire chain activity using massively parallel spike train recording. J Neurophysiol. 100(4):2165-2176
- Shimazaki, H., Amari, S., Brown, E.N., and Grün, S. (2012). State-Space Analysis of Time-Varying Higher-Order Spike Correlation for Multiple Neural Spike Train Data. PLoS Comput Biol 8(3): e1002385 doi: 10.1371/journal.pcbi.1002385
- Staude, B., Rotter, S., and Grün, S. (2010a). CuBIC: cumulant based inference of higher-order correlations in massively parallel spike trains. J Comput Neurosci, 29 (1-2): 327--350, DOI: 10.1007/s10827-009-0195-x
- Staude, B., Grün, S., and Rotter, S. (2010b). Higher-order correlations in non-stationary parallel spike trains: statistical modeling and inference. Front. Comput. Neurosci. 4:16. doi:10.3389/fncom.2010.00016
- Stevenson, I.H. and Kording, K.P.(2011). How advances in neural recording affect data analysis. Nat Neurosci, 14:139-142.
- Stosiek, C., Garaschuk, O., Holthoff, K., and Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. *Proc Natl Acad Sci U S A* **100**: 7319-7324.
- Tian, L., Hires, S. A., Mao, T., Huber, D., Chiappe, M. E., Chalasani, S. H., et al. (2009). Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nat Methods* **6**: 875-881.
- Torre, E., Picado-Muiño, D., Denker, M., Borgelt, C., and Grün, S. (2013). Statistical evaluation of synchronous spike patterns extracted by Frequent Item Set Mining. Front Comput Neurosci doi: 10.3389/fncom.2013.00132
- Tsien, R. W. and Tsien, R. Y. (1990). Calcium channels, stores, and oscillations. *Annu Rev Cell Biol* **6**: 715-760.
- Wallace, D. J., Meyer zum Alten Borgloh, S., Astori, S., Yang, Y., Bausen, M., Kugler, S., et al. (2008). Single-spike detection in vitro and in vivo with a genetic Ca2+ sensor. *Nat Methods* **5**: 797-804.
- Won, D. S., Chong, D. Y., and Wolf, P. D. (2003). Effects of spike sorting error on information content in multi-neuron recordings. In First International IEEE EMBS Conference on Neural Engineering, 2003. Conference Proceedings. (IEEE), pp. 618–621. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1196904 [Accessed September 5, 2012].
- Wood, F., Black, M. J., Vargas-Irwin, C., Fellows, M., and Donoghue, J. P. (2004). On the variability of manual spike sorting. Biomedical Engineering, IEEE Transactions on *51*, 912–918. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1300782 [Accessed August 14, 2011].

Program

International Neuroinformatics Coordinating Facility

INCF Secretariat
Seminar Room
Nobels väg 15A

Validation of Data - Analysis Methods
Agenda

Stockholm, Sweden
June 18th – June 19th
2012

Monday June 18th 2012

9:00 - 9:15 Gaute Einevoll and Thomas Wachtler
Opening Including Background and Goals

9:15 - 10:45 **Spike-sorting Validation Project**
- 9:15 - 9:45 Felix Franke
- 9:45 - 10:15 Espen Hagen
- 10:15 - 10:45 Andrey Sobolev

10:45 - 11:15 Coffee break

11:15 - 12:15 **Spike-train Extraction From Two-photon Calcium Data**
- 11:15- 11:45 Jason Kerr
- 11:45 - 12:15 David Greenberg

12:15 - 13:30 Lunch at *Königs*

13:30 - 14:30 **Testing of Methods for Analysis of LFP**
- 13:30 - 14:00 Daniel Wojcik
- 14:00 - 14:30 Gaute Einevoll

14:30 - 15:30 **Workflow and Collaboration**
- 14:30 - 15:00 Michael Denker
- 15:00 - 15:30 Leslie Smith

15:30 - 16:00 Coffee break

16:00 - 17:30 **Testing of Statistical Analysis Methods**
- 16:00 - 16:30 Sonja Grün
- 16:30 - 17:00 Martin Nawrot
- 17:00 - 17:30 Adrien Peyrache

17:30 - 18:00 **Planning for next day**

19:00 Dinner at *Bankomat*

Tuesday June 19th 2012

9:00 - 10:30 Work according to plan from previous day

10:30 - 11:00 Coffee break

11:00 - 12:30 Work according to plan from previous day

12:30 - 13:30 Lunch at *Königs*

13:30 - 15:00 Final discussion, summary of recommendations and future plans

15:00 - 15:30 Coffee break

15:30 - 16:30 Those still available work on the report

incf | Workshop Series

validation

INCF Secretariat
Karolinska Institutet
Nobels väg 15 A,
Stockholm, 171 77 Sweden

Tel: +46 8 524 870 93
Fax: +46 8 524 870 94
E-mail: info@incf.org
Web: www.incf.org