Reconstruction of complete connectivity matrix for connectomics by sampling neural connectivity with fluorescent synaptic markers

Yuriy Mishchenko

Columbia University, Department of Statistics, 1255 Amsterdam Ave
New York, New York, 10027

Abstract

Physical organization of the nervous system is a topic of perpetual interest in neuroscience. Despite significant achievements here in the past, many details of the nervous system organization and its role in animals' behavior remain obscure, while the problem of complete connectivity reconstructions has recently re-emerged as one of the major directions in neuroscience research (i.e. connectomics). We describe a novel paradigm for connectomics reconstructions that can yield connectivity maps with high resolution, high speed of imaging and data analysis, and significant robustness to errors. In essence, we propose that physical connectivity in a neural circuit can be sampled using anatomical fluorescent synaptic markers localized to different parts of the neural circuit with a technique for randomized genetic targeting, and that high-resolution connectivity maps can be extracted from such datasets. We describe how such an approach can be implemented and how neural connectivity matrix can be reconstructed statistically using the methods of Compressive Sensing. Use of Compressive Sensing is the key to allow accurate neural connectivity reconstructions with orders-of-magnitude smaller volumes of experimental data. We test described approach on simulations of neural connectivity reconstruction experiments in *C. elegans*, where real neural wiring diagram is available from past electron microscopy studies. We show that such wiring diagram can be in principle re-obtained using described approach in 1-7 days of imaging and data analysis. Alternative approaches would require currently at least 1-2 years to produce a single comparable reconstruction. We discuss possible applications of described approach in larger organisms such as *Drosophila*.

1. Introduction

In recent years the problem of complete reconstructions of neural connectivity has re-emerged as one of the major goals of modern neuroscience (Sporns et al., 2005; Briggman and Denk, 2006; Smith, 2007; Lichtman et al., 2008; Helmstaedter et al., 2009). Several "connectomics" projects had been thus proposed relying on classical electron microscopy (Briggman and Denk, 2006; Smith, 2007), injections of tracer viruses (Bohland et al., 2009), sparse expression of fluorescent cytoplasmic markers (Svoboda, Personal Communication), diffusion tensor imaging (Hagmann et al., 2007; Hagmann et al., 2008), among others. Despite significant progress made along these directions this problem remains far from being solved and significant limitations continue to exist in the resolution or practically feasible volume of possible reconstructions. E.g., diffusion tensor imaging (DTI) allows one to obtain connectivity maps for the entire human brain, but only very large-scale projections, such as between entire cortical areas, can be mapped (Hagmann et al., 2007; Hagmann et al., 2008). Electron microscopy (EM), on the other hand, can produce reconstructions at the level of individual synapses, but is very difficult to scale to large volumes (White et al., 1986; Briggman and Denk, 2006). Due to the ability of EM to map neural circuits at the level of individual synapses, EM remains the favorite approach for the new connectomics

reconstructions (White et al., 1986; Briggman and Denk, 2006; Smith, 2007). Yet, EM reconstructions are known to be extremely labor intensive (White et al., 1986; Briggman and Denk, 2006), while tracing of axons and dendrites over large distances through densely cluttered EM images has proven difficult to automate (Jurrus et al., 2006; Jain et al., 2007; Macke et al., 2008; Mishchenko, 2009). Also, significant sensitivity of EM reconstructions to point errors during imaging and analysis had been recently revealed (Mishchenko, 2009; Mishchenko et al., 2010b).

In this paper, we describe a different paradigm for reconstructions of neural connectivity that can combine high level of detail of produced connectivity maps with high speed of fluorescent light microscopy imaging and significant robustness to errors. In essence, we propose that neural connectivity can be sampled using anatomical fluorescent synaptic markers introduced into different parts of a neural circuit genetically or otherwise and that high-resolution connectivity maps can be subsequently extracted from such datasets using statistical techniques. We describe a specific setup for such an experiment that utilizes *two-component fluorescent synaptic markers* (Feinberg et al., 2008; Mishchenko, 2010) and a *recombinase system for stochastic gene expression* (Livet et al., 2007; Lichtman et al., 2008; Luo et al., 2008) to randomly express the synaptic marker in a neural circuit and produce a sample of certain simple connectivity measurements. We show that synaptic connectivity matrix can be accurately reconstructed from such data using the techniques of *Compressive Sensing* (Donoho, 2006; Candes and Wakin, 2008). Use of Compressive Sensing is the key to allow accurate neural connectivity reconstructions with orders-of-magnitude smaller volumes of the experimental data. We test the feasibility of this approach by simulating a hypothetical neural connectivity reconstruction experiment in *C. elegans*, a popular neuroscience model where a real neural wiring diagram is available from EM (White et al., 1986). Using such an actual wiring diagram, we show that the complete connectivity matrix in *C. elegans* can be in principle re-obtained with the described approach in only 1-7 days of imaging and data analysis, whereas the best alternative available EM approach currently would require at least 1-2 years for a single comparable reconstruction. We also discuss different strategies that can allow applying the described approach with different connectivity probes and genetic targeting techniques in larger organisms such as *Drosophila*. Our results open new possibilities for quantitative data-rich empirical studies of neural circuits' organization and functions in the brain.

2. Materials and Methods

2.1. Probing specific synaptic connectivity with fluorescent synaptic markers

Recently certain two-component fluorescent synaptic markers had been developed that can be used to label and individually observe synapses between specific neurons. These include the synaptic marker GRASP (Feinberg et al., 2008) and the two-component co-localization synaptic marker (Mishchenko, 2010).

The synaptic marker GRASP relies on the process of recombination of the two fragments of the split-GFP molecule to label synaptic contacts between specified neurons. Split-GFP molecule consists of a pair of proteins that by themselves are not fluorescent, but can chemically recombine into a functional fluorescent protein (GFP) when in close proximity from each other (Sarkar and Magliery, 2008). In GRASP, such split-GFP fragments are genetically tailored to the proteins normally present on the external surfaces of pre- and post-synaptic regions of neurons

and are separately expressed in selected neurons using gene-fusion (Fig. 1A). When two such neurons form a synapse, split-GFP fragments recombine over its synaptic cleft and produce a functional GFP, thus, fluorescently tagging that synapse.

The principle of (Mishchenko, 2010) is essentially the same with that of GRASP. This synaptic marker consists of a pair of different wavelength fluorochromes targeted to the proteins normally present on pre- and post-synaptic surfaces of neurons (Fig. 1B). Unlike in GRASP, such fluorochromes are not required to be chemically complementary; instead, detection of synapses is performed by observing spatial co-localization of fluorescence from these fluorochromes. At existing synapses such pre- and post-synaptic fluorochromes are co-localized naturally due to dimensions of the synaptic clefts. Such co-localization may also occur randomly, e.g., due to random oppositions of pre- and post-synaptic surfaces of nearby synapses. However, in (Mishchenko, 2010) it was shown using electron microscopy data that such random oppositions should be extremely rare for optical instruments with sufficient resolving power, including array tomography (Micheva and Smith, 2007) and structured illumination microscopy (Gustafsson, 2000, 2005), in which case synapses can be detected with such an approach typically with an accuracy of better than 95%.

The connectivity probe considered in this work is essentially a replica of the two-component construct described above with one difference. Because the above synaptic markers originally only label synapses of affected neurons, corresponding cell bodies remain unaffected. For our purposes, however, it will also be necessary to label cell bodies of the neurons expressing such markers. For that, to the above construct we introduce a "helper" element consisting of a fluorescent nuclei marker that can be used to identify neurons expressing the synaptic marker post-factum, Fig. 2A-B.

2.2. Probing synaptic connectivity stochastically with fluorescent synaptic markers and a recombinase system

Recombinase systems such as the Cre/Lox system or the Flp/Frt system can be used to express the synaptic marker in a neural circuit stochastically and, thus, to allow sampling different connections in that circuit in a high-throughput manner. Recombinase systems are well known genetic tools for manipulating gene expression and, in particular, for expressing genes stochastically in a population of cells. A particularly ingenious tour-de-force demonstration of this latter aspect had been recently produced with the Brainbow mouse using the Cre/Lox system (Livet et al., 2007; Lichtman et al., 2008; Lu et al., 2009a; Lu et al., 2009b). In the Brainbow mouse (or rather one of its implementations) a cassette of several genes encoding different emission spectra fluorochromes (e.g., GFP, YFP, etc.) is flanked in the genome with inversely oriented loxP-sites. The Cre action then consists in flipping orientation of such sequences (Fig. 1C). In this process, loxP-enclosed sequences stochastically assume either direct or reverse orientation in different cells. Since the cells where orientation of the sequence is reversed cannot transcribe the respective genes successfully, Brainbow succeeds in randomly deactivating different fluorochromes in different cells. Different cells then attain distinctive cytoplasmic colors depending on the combination of the fluorochromes that are "active" there, and can be distinguished merely by color. This makes analysis and tracing of projections of such neurons significantly simpler and at the same time more reliable (Livet et al., 2007). A number of larger

axons and synapses in several neural circuits had been thus reconstructed using this approach (Lu et al., 2009a; Lu et al., 2009b).

In our settings, we use the recombinase system to randomly deactivate expression of the synaptic marker in different cells of a neural circuit. In one possible implementation, the sequences for the post- and pre-synaptic components of such marker (*sA* and *sB*, respectively) are placed into reverse-tandem flanked jointly with loxP-sites, Fig. 2A. Two different outcomes of Cre-recombination then can occur with equal probability. In outcome (i) sequence *sA* is transcribed leading such neurons to express the post-synaptic marker fragment. The respective nuclei tag *nA* is also transcribed, thus, labeling the nuclei of such neurons with a distinctive color. In outcome (ii) sequence s*B* is transcribed leading such neurons to express the pre-synaptic fragment and the nuclei tag *nB*. Neurons always express exclusively either *sA* or *sB* and never *sA* and *sB* together. This may be advantageous if the expression of both marker fragments simultaneously in the same neuron is undesirable (Bargmann, Personal communication). Fluorescent puncta are formed by all synapses made by the neurons that express complementarily the pre- and post-synaptic marker fragments, Fig. 2C.

In another implementation, Fig. 2B, the sequences for the post- and pre-synaptic marker fragments are flanked with loxP individually. Four outcomes of Cre-recombination are then possible with equal probability. In outcome (i) both *sA* and *sB* are transcribed and both pre- and post-synaptic sites of the neurons are tagged and visualized. In outcome (ii) only *sA* sequence is transcribed and, thus, only post-synaptic sites of such neurons are visualized. In outcome (iii) only *sB* sequence is transcribed and only pre-synaptic sites of such neurons are visualized. And in outcome (iv) neither *sA* nor *sB* are transcribed and synapses of such neurons are not observed. If homogeneous action of a single recombinase system across all neurons cannot be achieved, mixtures of multiple recombinase systems can be used in the manner described above.

Other implementations are obviously possible; Fig. 2 provides only one possible "real life" example of such an implementation together with all appropriate recombination outcomes.

2.3. Sampling neural connectivity with stochastically expressed fluorescent synaptic markers

We suggest combining a two-component fluorescent synaptic marker and a recombinase system for stochastic gene expression to stochastically sample synaptic connectivity in a neural circuit as illustrated schematically in Fig. 3. In Fig. 3A three subsets of neurons from a hypothetical neural circuit are shown expressing at random the post- and pre-synaptic marker fragments and the respective nuclei tags. These three samples can correspond to three different specimens from the genetic line containing implementation of such construct, i.e. its *three different phenotypes*. Whenever two neurons that express complementarily the pre- and post-synaptic marker fragments form a synapse, a fluorescent punctum is also formed allowing one to observe such synapses with fluorescent microscopy. Obviously, different synapses are visualized in different specimens.

Because single synaptic marker labels all synapses in the same way (e.g., GRASP labels tags all its synapses with the same fluorochrome - GFP), in general, it will be impossible to determine which labeled synapses belong to which neurons. Instead, a cumulative measure should be obtained such as *the total count* of all puncta or *the combined fluorescence size* of all puncta. We will denote such measurement with symbol *n*. Identity of the neurons expressing the synaptic

marker in each experiment should be also recorded using associated nuclei tags. We will denote such expression patterns for the post- and pre-synaptic marker fragments with symbols *a* and *b*. That is, *a* is just a list of all neurons that express given "post-synaptic" nuclei tag in a given specimen and *b* is that for the neurons expressing "pre-synaptic" nuclei tag. The set of observations that is thus procured is that of the triples, {(*n*, *a*, *b*)}, for different specimens. Such a sample is a characteristic of the above stochastic phenotype (Fig. 3B), and can be used to exactly reconstruct the underlying neural connectivity matrix, as we discuss below.

2.4. Reconstruction of the neural connectivity matrix with stochastically expressed fluorescent synaptic markers

The key observation of this work is to recognize that the measurements introduced in Section 2.3. can be interpreted as certain sums over the neural connectivity matrix (Fig. 3C). Specifically, if C_{ij} is an element of such neural connectivity matrix describing, e.g., the number or the size of all synapses between different pairs of post-synaptic and pre-synaptic neurons *i* and *j*, respectively, the measurement *n* in a specimen *k* can be mathematically represented as the following sum,

$$n(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i^{(k)} b_j^{(k)} C_{ij} \; . \tag{1}$$

Here, *N* is the total number of neurons in the circuit, $a_i^{(k)}$ and $b_j^{(k)}$ are the indicator functions characterizing the expression patterns of the post- and pre-synaptic marker fragments in animal *k*, respectively. More specifically (see Fig. 3A-C for illustration),

$$
a_i^{(k)} = \begin{cases} 1, \text{neuron } i \text{ expressed the post - synaptic marker in specimen } k \\ 0, \text{neuron } i \text{ did not express the post - synaptic marker in specimen } k \end{cases}
$$
$$
b_j^{(k)} = \begin{cases} 1, \text{neuron } j \text{ expressed the pre - synaptic marker in specimen } k \\ 0, \text{neuron } j \text{ did not express the pre - synaptic marker in specimen } k \end{cases}
\tag{2}
$$

For recombinase systems such as the Cre/Lox or Flp/Frt, $a_i^{(k)}$ and $b_j^{(k)}$ will be essentially random vectors with approximately 50% of ones and 50% of zeros (Livet et al., 2007; Lichtman et al., 2008; Luo et al., 2008).

Using Eq. (1), we can expect that it may be possible to estimate the neural connectivity matrix C_{ij} statistically from the collection of measurements {(*n*(*k*), $\{a_i^{(k)}\}$, $\{b_j^{(k)}\}$), *k*=1,…,*K*}. Specifically, we can search for C_{ij} that can best describe the available set of measurements (1) within the class of sparse matrices. We know that connectivity in the neural systems is typically very sparse, and sparse priors previously had been shown to allow significant reduction in the amount of data necessary for a reliable reconstruction of linearly encoded signals. Although such a problem may appear to be hopelessly ill-defined (e.g., all expression patterns are broad and nonspecific, sparse prior is nonspecific as well, etc.), contrary to this impression it is possible to recover the connectivity matrix under these conditions quite accurately. In fact, it can be mathematically shown that the connectivity matrix can be extracted *exactly* from relatively small number of measurements (1) using the methods of Compressive Sensing (Candes and Romberg, 2005;

Donoho, 2006). In subsequent sections we describe the mathematical procedures that can be used to identify the connectivity matrix from such data.

2.5. Reconstruction of the neural connectivity matrix using Compressive Sensing

Compressive Sensing (CS) (Candes and Romberg, 2005; Candes et al., 2006; Donoho, 2006; Candes and Wakin, 2008) is a recently emerged field of signal processing that deals with decoding of linearly encoded sparse signals. Linear encoding here means that a signal, *f(t)*, is encoded by a collection of linear measurements $y(k) = \phi_k * f = \sum \varphi_k(t) f(t)$ (sum is over all *t*=1,…,*T*). Sampling *f(t)* at a set of predefined points $\{t_i\}$ or observing the Fourier components of *f(t)* would be examples of linear encoding. The goal of CS is to accurately reconstruct *f(t)* from an incomplete set of such linear measurements, while knowing that the original signal is sparse (i.e. that *f(t)*=0 for most *t*, but not knowing where *f(t)* is zero). For example, we know that the connectivity matrix in the neural systems should be sparse, but we may not know exactly which neurons are connected. CS is provably nearly optimal approach for addressing this problem.

Certain remarkable mathematical properties of CS had been recently rigorously established and should be mentioned (Candes and Romberg, 2005; Candes et al., 2006; Candes and Wakin, 2008). First, it was shown that it is possible to reconstruct sparse signal *exactly* from a small number of its measurements, $K \propto \log(T) T_{sp}$, where $T_{sp} << T$ is the number of nonzero elements in *f(t)*. Second, appropriate reconstruction procedure is tractable: with probability approaching to 1 exponentially in *T* the sought signal is the smallest l_1-norm solution of the set of linear equations $\{y(k) = \sum \varphi_k(t) f(t), \ k = 1,..., K\}$. Third, such exact reconstruction can be obtained essentially independently of the specific form of $\varphi_k(t)$, i.e., exact reconstructions can be achieved using nearly arbitrary set of probing waveforms $\varphi_k(t)$.

Recall now that our measurements of the count or combined size of labeled synapses, *n*, can be interpreted as following sums over the neural connectivity matrix,

$$n(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i^{(k)} b_j^{(k)} C_{ij} \ ,$$

where $a_i^{(k)}$ and $b_j^{(k)}$ are the indicator functions for the respective expression patterns of the post- and pre-synaptic marker fragments in animal *k*, Sections 2.3-2.4. We now immediately recognize in these settings the linear encoding problem described above. According to CS (Candes et al., 2006), then, the connectivity matrix can be accurately recovered from a sample of such measurements by solving a linearly constrained l_1-optimization problem,

$$\min \|C\|_{l_1} = \min \sum_{i=1}^{N} \sum_{j=1}^{N} |C_{ij}| , \text{ subject to} \tag{3a}$$

$$n(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i^{(k)} b_j^{(k)} C_{ij}, k = 1,..., K . \tag{3b}$$

Problem (3) is a standard linear-program (LP) and powerful methods exist for solving it efficiently, e.g., such as the interior point methods (Wright, 1997; Vanderbei, 2001; Boyd and Vandenberghe, 2004). For large connectivity matrices, however, the issue of computational

scalability should be addressed. In particular, Matlab's LP solver fails to solve problem (3) when N exceeds $N \approx 50$ neurons due to computer memory limitations. Remember that there are N^2 variables to be optimized in problem (3).

In (Candes and Romberg, 2005) an alternative method for solving Eqs. (3) for very large N^2 was proposed. Specifically, if l_1-norm of the connectivity matrix, $S = \sum_{i=1}^{N} \sum_{j=1}^{N} |C_{ij}|$, is known a-priori, solution of problem (3) can be found as the intersection of two convex sets – the l_1-cube $\|C\|_{l_1} = S$ and the hyper-plane (3b). In our case, S can be estimated directly from the measurements $\{(n, a, b)\}$, i.e. $S \approx \langle n(k) \rangle / \langle |a^{(k)}| \|b^{(k)}\| \rangle$, where the average is over all samples k and $|a^{(k)}| = \frac{1}{N} \sum_{i=1}^{N} a_i^{(k)}$. According to the central theorem of CS, such intersection is unique and corresponds to the exact solution of problem (3) when K is sufficiently large (Candes and Romberg, 2005).

To actually find said intersection, the algorithm of alternate projections was used (Bregman, 1965). In this algorithm one starts with a random guess for the connectivity matrix, $C^{(0)}$, and then repeats the two steps of consequently projecting the current guess, $C^{(l)}$, onto the hyper-plane (3b) and then onto the l_1-cube $\|C\|_{l_1} = S$,

$$C^{(l+1/2)} := C^{(l)} + P^T (PP^T)^{-1}(n - P \cdot C^{(l)}). \tag{4a}$$

$$C_{ij}^{(l+1)} := \max(0, C_{ij}^{(l+1/2)} - \gamma^{(l)}), \tag{4b}$$

In Eq. (4a) we used vector notation for clarity, i.e., NxN connectivity matrix C_{ij} was represented by a N^2x1 vector over joint indices (ij) and P denoted the KxN^2 matrix $P(k;ij) = a_i^{(k)} b_j^{(k)}$. In such notation Eq. (3b) clearly corresponds exactly to the dot-product of P and C, $n = P \cdot C = \sum_{(ij)} P(k;ij)C_{ij}$. Eq. (4a) describes the step of projecting $C^{(l)}$ onto the hyper-plane (3b): it can be trivially checked that $P \cdot C^{(l+1/2)} = n$. Eq. (4b) describes the step of projecting $C^{(l+1/2)}$ onto the l_1-cube $\|C\|_{l_1} = S$. $\gamma^{(l)}$ is chosen at each iteration to achieve $\|C^{(l+1)}\|_{l_1} = S$ and controls retraction of $C^{(l+1/2)}$ onto the nearest face of that cube. Steps (4a) and (4b) are repeated until convergence, which typically can be achieved rapidly because the process proceeds between two hyper-planes - the hyperplane of the constraints and one of the faces of the l_1-cube.

2.6. A hypothetical neural connectivity reconstruction experiment using stochastically expressed synaptic marker GRASP in *C. elegans*

We evaluate performance of the described approach using simulations of a hypothetical neural connectivity reconstruction experiment in *C. elegans*. *C. elegans* is a popular neuroscience model organism and the only organism where real wiring diagram is known from prior electron microscopy work (White et al., 1986). In particular, this motivated our attention specifically to that system, although our method obviously is not restricted to applications in *C. elegans*.

We simulated connectivity reconstruction experiments involving from 500 to 10,000 measurements performed with the described system using the synaptic marker GRASP (stochastic GRASP). For each observation we generated pre- and post-synaptic GRASP expression patterns according to Section 2.2. For the construct in Fig. 2A, each neuron was assumed to express pre- and post-synaptic GRASP fragments uniformly and mutually exclusively with probability 50%. For the construct in Fig. 2B each neuron was assumed to express pre- and post-synaptic GRASP fragments uniformly and non-exclusively with probability 50%. Using thus generated $a_i^{(k)}$ and $b_j^{(k)}$, we simulated measurements $n(k)$ according to Eq.(1).

We also considered a number of "noise" factors that can affect actual experiments. One of such factors considered was *biological variability*, i.e., that actual connectivity matrix could vary from one stochastic GRASP animal to another. In that case for each animal k a slightly different connectivity matrix $C_{ij}^{(k)}$ was prepared using following formula,

$$C_{ij}^{(k)} = \overline{C}_{ij}(1 - a_b) + v[a_b \overline{C}_{ij}]. \tag{5}$$

Here $\overline{C}_{ij}$ was the test wiring data from (White et al., 1986), i.e. for each pair of neurons i and j $\overline{C}_{ij}$ described the number of synaptic contacts between these neurons from the EM data in (White et al., 1986). $v[x]$ was a Poisson-distributed random variable with mean x and modeled the variation in C_{ij} among animals. Parameter a_b controlled the degree of such variation - $a_b=0$ corresponded to no variability and $a_b=1$ corresponded to the case where synapses were formed completely at random with certain number of synapses between given neurons on average. The objective of the reconstruction in this case was to recover the *average* connectivity matrix $\overline{C}_{ij}$.

Second, we considered noise that could be added into the observations during the measurement process itself, e.g., due to imperfections in the experimental setup. For that we altered the measurements $n(k)$ by adding white noise, $n(k) \rightarrow n(k)(1 + a_o v)$. Here, v was a Normally-distributed random variable with zero mean and unit variance, and a_o controlled the relative strength of such "measurement" noise.

Finally, we considered errors that could be made during measurements of the expression patterns $a_i^{(k)}$ and $b_j^{(k)}$. Since such errors typically would lead reconstruction algorithm to use wrong projecting matrix P, such errors clearly could result in deviations of the reconstruction from the truth. To consider this factor in our model, we corrupted the "true" expression patterns by randomly shuffling identities of a small number of neurons. In that sense, we assumed that the total number of neurons was known a-priori (as is in fact in *C. elegans*) and the only errors that could be committed were due to confusion of nearby neurons.

2.7. Comparison with alternative approaches for connectomics reconstructions

We performed a comparison of our approach with potential applications of several alternative methods for connectomics reconstructions – serial electron microscopy (EM) and pair-wise cell recordings (PCR). We inspected such applications vs. the amount of the reconstruction effort involved, E, and the impact of different degrees of noise and contributed errors, f. In order to

perform simulations we used real neural wiring data from *C. elegans*, similar in essence to Section 2.6.

In EM, reconstruction effort E refers to the fraction of the neural tissue volume imaged in respect to the total volume of the neural circuit, and in PCR that refers to the fraction of all neural pairs recorded from. The part of the neural connectivity matrix that is recovered for given E with EM is that corresponding to the neurons whose cell bodies are contained inside the reconstructed volume, of which the fraction is obviously E. Note that, although a larger number of synapses may be observed in the reconstructed volume, only a fraction of these can be identified corresponding to the neurons whose cell bodies can be found inside the reconstructed volume. Therefore, to simulate the result of EM reconstructions for different E we considered the connectivity matrix in which only connections between EN randomly chosen neurons were retained, while all other connections were assumed to be unknown. To simulate the result of PCR reconstructions for different E, respectively, we considered the connectivity matrix in which EN^2 randomly chosen neural connections were retained.

In order to inspect the impact of biological variability on the result of EM and PCR reconstructions, we simulated connectivity matrices for individual specimens using the scaled Poisson model, given by Eq. (5). Since EM or PCR would produce connectivity reconstructions essentially within individual such specimens, to simulate the impact of biological variability we compared such individual scaled Poisson matrices with the average $\overline{C}_{ij}$. In general, we found that the impact of biological variability on such reconstructions was insignificant, i.e., the members of scaled Poisson family could generally be expected to be similar with their average $\overline{C}_{ij}$.

Errors in EM reconstructions generally are comprised of the mistakes made in the traces of thin axons due to local misinterpretations of images or photographing defects (Mishchenko, 2009; Mishchenko et al., 2010b). For example, an axon can be confused with a similar nearby axon in an image and, thus, continued as such of a wrong neuron. Such errors can be characterized by their rate, f, i.e., the probability for one such error to occur in the reconstruction of an axon over a given interval, e.g., such as between two its consecutive synapses. To simulate the impact of such errors we proceeded as follows. For each neuron j we traversed all its post-synaptic connections in C_{ij} in (White et al., 1986) in random order, one by one. During each step the error event was generated with probability f. If error event was generated, then the effective identity of the pre-synaptic neuron was changed to another random neuron, j^{*}, and subsequent connections for the reconstruction of the original neuron j were selected from $C_{ij^{*}}$. This was meant to model confusion of similar nearby axons during tracing. This process was repeated until all connections from C_{ij} were exhausted, so that all rows of the corrupted reconstructed connectivity matrix were thus filled.

Errors in PCR reconstructions typically would correspond to failures to observe a connection when stimulating one neuron and recording from another. This can occur, e.g., due to failure of the target neuron to get stimulated or failure to observe a weak post-synaptic sub-threshold response, etc. Such errors are mathematically described as point errors in the connectivity matrix and can be characterized by their probability per one neural pair, f. To simulate the impact of such errors, therefore, we deleted at random a fraction f of the connections from the reconstructed connectivity matrix and compared such corrupted result with the original C_{ij}.

Additive noise in the measurement of the connection strength can also be present in PCR due to fluctuations in the membrane potential levels. This factor was not considered in this work.

3. Results

3.1. Feasibility of the wiring diagram reconstructions in *C. elegans* using stochastic GRASP

To test our approach we simulated a hypothetical neural connectivity reconstruction experiment in *C. elegans. C. elegans* is a popular neuroscience model and the only animal where complete wiring diagram is known from serial electron microscopy (White et al., 1986). This, in particular, motivated our attention specifically to this system, although our approach obviously is not restricted to *C. elegans*.

The neural wiring diagram in *C. elegans* in (White et al., 1986) was produced as follows. Entire body of one *C. elegans* specimen was imaged at extremely high resolution using serial electron microscopy. (More accurately, several specimens were imaged partially in order to achieve a dataset equivalent to one full body coverage.) Consequently, synapses were manually found in the images and associated with the corresponding axons and dendrites, and axons and dendrites were traced to the respective cell bodies through multiple EM images, also manually. Total count of synapses between different pairs of neurons was thus tabulated, and the table describing these counts was compiled as the neural wiring diagram of *C. elegans* (now available from *wormatlas.org*). We used this actual wiring diagram as the ground truth for a simulated neural connectivity reconstruction experiment, see Section 2.6. The wiring diagram of *C. elegans* contains $N \approx 300$ neurons connected via a grand total of about 6000 synapses in 2500 distinct neural connections.

We simulated a neural connectivity reconstruction experiment with stochastically expressed synaptic marker GRASP (stochastic GRASP) using from $K=500$ to $K=10,000$ animals, and inspected obtained reconstructions both in matrix form and as a scatter plot of the reconstructed vs. actual connection weights (Fig. 4). Quantitatively, we characterized the reconstructions using the correlation coefficient between the reconstructed and actual connection weights, r^2. $r^2=0$ would correspond, naturally, to no correlation between the reconstructed and true connection weights, and $r^2=1$ would correspond to a reconstruction that was perfect or exact. Reconstructions with $r^2=0.5$, as we observed in particular, already were generally good enough to provide a practically meaningful estimate of the neural connectivity matrix, Fig. 4.

In Fig. 5 we show r^2 for the stochastic GRASP reconstructions plotted vs. the number of measurements K. As can be seen from this figure, perfect reconstructions were obtained already with $\approx 10,000$ measurements. This is in excellent agreement with the CS theory and far below $\approx 90,000$ measurements that can naively be expected to be necessary to completely characterize the connectivity matrix of 300x300 neurons in *C. elegans* at least once. Practically meaningful reconstructions could be obtained with $\approx 5,000$ measurements. These results were obtained for both exclusive and non-exclusive constructs, Fig. 2A and 2B. Compressive Sensing was critical here as, e.g., a naïve solution using a more conventional L_2 regularization (i.e.,

$$\min \|C\|_{l_2} = \min \sum \sum \left| C_{ij} \right|^2 \text{ s.t. } n(k) = \sum \sum a_i^{(k)} b_j^{(k)} C_{ij} \text{) resulted in practically useless}$$

reconstructions with the sample of this size.

These results indicate that complete reconstructions of the wiring diagram in *C. elegans* using stochastic GRASP can be feasible and in principle can be performed with quite modest experimental effort, relying only on already existing technologies such as Cre/Lox or Flp/Frt recombinase systems, GRASP, and low-end light microscopy.

3.2. Impact of biological variability on the wiring diagram reconstructions using stochastic GRASP

It may be expected that the connectivity matrix in different animals of the same species may vary from animal to animal. In this case, in each stochastic GRASP experiment a slightly different connectivity matrix will be probed. Although it is not yet known to what degree real neural circuits should be expected to vary (in fact directly opposite opinions exist on this topic among neuroscientists), we inspected potential impact of such variability on the neural connectivity reconstructions using stochastic GRASP, see Section 2.6. Results of this study are shown in Fig. 6. In every case, we found that the impact of biological variability on the connectivity reconstructions was insubstantial.

3.3. Impact of added noise on the wiring diagram reconstructions using stochastic GRASP

It can be expected that a perfect measurement of the total puncta count or size may not be possible to obtain under realistic conditions. Such noise added during the process of collecting the measurements themselves may obviously distort the results of the connectivity reconstructions. We inspected potential impact of such noise on the reconstructions using stochastic GRASP, see Section 2.6. Results of this study are shown in Fig. 7. Reconstructions were found to be stable under small additions of noise, i.e., small amounts of noise caused only proportionately small deviations in the reconstruction result (Candes et al., 2006; Candes and Wakin, 2008). Yet, sensitivity to noise was substantial, and only up to 3-4% of added noise could be tolerated for the minimal $K \approx 10,000$. Robustness improved with the number of measurements K, so that higher levels of noise could be potentially mitigated by collecting datasets with a larger number of measurements. Although we did not extend this study to larger K, because such an extension would be very computationally expensive, the specifications for particular target reconstruction accuracy and specified noise levels can be computed straightforwardly using the methods in Sections 2.5 and 2.6.

3.4. Impact of misidentifications in detected GRASP expression patterns on the wiring diagram reconstructions using stochastic GRASP

It can be expected that a perfect measurement of GRASP expression patterns, *a* and *b*, may not be possible to obtain under realistic conditions, so that a certain amount of errors in the detected expression patterns should be anticipated. Clearly, such errors can result in deviations of the reconstructed connectivity matrix from the truth. We inspected potential impact of such errors on the reconstructions using stochastic GRASP in greater detail, see section 2.6. Results of this study are shown in Fig. 8. Although reconstructions were found to be stable again, sensitivity to errors in the expression patterns was substantial, and only up to 5-6% of misidentified neurons in the detected expression patterns (i.e. 15-20 misidentified neurons per 300 in *C. elegans*) could be tolerated for the minimal $K \approx 10,000$. As before, reconstructions with larger number of measurements K were more robust to errors, so that higher error levels could be tolerated with datasets of larger size. Although we did not extend this study to K greater than 10,000, again

because that would be very expensive computationally, the specifications for particular target reconstruction accuracy and specified error levels can be computed using the methods in Sections 2.5 and 2.6.

4. Discussion

We describe a new approach for reconstructions of neural connectivity that utilizes two-component fluorescent synaptic markers (Feinberg et al., 2008; Mishchenko, 2010) together with a recombinase system for stochastic gene expression (Livet et al., 2007; Lichtman et al., 2008; Luo et al., 2008) to randomly sample synaptic connectivity in a neural circuit. Complete connectivity matrix is reconstructed from produced fluorescent measurements using the methods of *Compressive Sensing* (Candes and Romberg, 2005; Candes et al., 2006; Donoho, 2006; Candes and Wakin, 2008).

We use real neural wiring data for *C. elegans*, available from EM (White et al., 1986), to show that described approach in *C. elegans* is feasible and can be executed with already existing technologies. In *C. elegans*, we find that 5,000-10,000 measurements performed with stochastically expressed synaptic marker GRASP (Feinberg et al., 2008) will suffice to successfully recover the complete wiring diagram. Given small size (100x100x1000 µm) and fast development (2-3 days) of *C. elegans*, 10,000 animals can be incubated on a single Petri-dish in the span of several days. Modern *C. elegans* phenotype screens, in fact, routinely work with the populations that large. Fluorescent measurements can be obtained by performing 3D-scans of the specimen bodies, while GRASP expression patterns can be obtained at the same time using associated nuclei-targeted fluorescence. Computer algorithms such as (Long et al., 2008) can be used to automatically determine GRASP expression patterns in *C. elegans* from produced imaging data. Although, according to our calculations, the identification error rates reported in (Long et al., 2008) are still quite high for this purpose (i.e., 95% correct identifications and 5% errors), significant improvement from this first attempt in the computerized cell identification can be naturally expected in the near future. Assuming that the imaging data at resolution ≈0.5 µm/pixel can be acquired with a confocal microscope at the speed of at least 10MHz, it should be possible to perform such a full scan of one specimen's body in 1 minute or less using a specialized imaging setup (Kerr, Personal communication). 10MHz acquisition rate can be achieved, e.g., with a 1000x1000 pixel CCD camera mounted on a confocal microscope and taking images at the frequency of 10 frames per second. The body of one *C. elegans* specimen at the resolution of 0.5 µm/pixel contains approximately 80 million pixels. Using a CCD camera with 1000x1000 pixels and by taking one image of the part of the longitudinal section of *C. elegans* of 1000x200 pixels every 1/10 of a second, 80 million pixels can be acquired in $[80*10^6$ pixels]/$[2*10^5$ pixels/frame]/[10 frame/second] = 40 seconds. We may also note that the entire body of the specimen may not need to be imaged, since the neural connections will typically be localized only to small regions in the specimen's body. Then, the total imaging time for the entire sample of 10,000 animals can be estimated in only 1-7 days. Subsequent reconstruction of the connectivity matrix from such data is computationally straightforward and can be performed in several hours of computations on a laptop computer, as was shown here.

Several remarkable features of such reconstruction approach should be explicitly mentioned. Complete connectivity can be recovered using what can be considered a grossly incomplete and non-specific dataset, with the number of available measurements $K \sim N_{sp} \log N << N^2$, where

N_{sp} is the number of connected neural pairs in the circuit and N is the total number of neurons (Candes and Romberg, 2005). Clearly, should we have known the identity of all connected neural pairs in the circuit, we could use $K \sim N_{sp}$ measurements to determine the connectivity matrix, e.g., by using pair-wise cell recordings. With Compressive Sensing we can achieve similar performance bound without any prior knowledge about the identity of the connected neural pairs. Reconstructions with smaller than minimally sufficient number of measurements K will produce the best possible approximations to true connectivity matrix with $\approx K$ terms (Romberg, 2008). Described approach does not involve tracing of neurons, recordings from individual neurons, genetically targeting individual neurons or small groups, or other explicit association of synaptic connection with remotely located neurons, which is the main difficulty of many other connectomics paradigms. Synapses are associated with the relevant neurons a-posteriori, and detailed connectivity matrix is obtained even when all neural "targeting" patterns are broad and nonspecific (e.g., stochastic).

We can compare our approach with alternative existing connectomics paradigms, although such comparison is admittedly complicated by the diversity of existing approaches, making it difficult to bring them to a common denominator, and the fact that none of the existing approaches is yet capable of detailed reconstructions of connectivity even in a system such as *C. elegans* (with the exception of serial electron microscopy). Here, we compare two such paradigms that in our opinion have the best prospects for complete neural connectivity reconstructions in the near future – serial electron microscopy (EM) (White et al., 1986; Briggman and Denk, 2006; Smith, 2007; Mishchenko, 2009) and optically assisted pair-wise cell recordings (PCR) (Bureau et al., 2004; Baker et al., 2005; Petreanu et al., 2007). We have left out a number of promising techniques for estimating neural connectivity from functional correlations between neurons, e.g., using calcium imaging (Broome et al., 2006; Jones et al., 2007; Pillow et al., 2008). Although these now potentially allow reconstructions of circuits with hundreds and thousands of neurons (Stevenson et al., 2009; Mishchenko et al., 2010a), only an effective connectivity matrix is produced with such an analysis, whose relationship to the physical circuit structure is yet to be clearly elaborated.

EM reconstruction paradigm comprises: a) directly imaging a block of neural tissue with serial electron microscopy; b) finding synapses in the images and associating them with corresponding axons and dendrites; c) tracing associated axons and dendrites through multiple images to corresponding cell bodies; d) recording a connection between a pair of neurons. The main advantage of EM reconstructions is their ability to assess the physical structure of a neural circuit directly, as individual synapses, axons, and dendrites can be all directly seen in EM images. Furthermore, details of neural morphology such as arbor shape, branching structure, synapses bunching, etc., can be extracted from EM images. Such data can be used to directly model physical processes and signal transmission inside axons, dendrites, and neurons as well as to study sub-cellular organization of neurons and micro-organization of neuropil.

The main disadvantage of EM approach is very low speed with which the data can be acquired and extreme difficulty of subsequent image analysis. Analysis of EM data is characterized by very high labor-intensity – manual reconstructions in one *C. elegans* specimen in (White et al., 1986) took over 10 years to complete. Modern automation techniques can reduce this time, but has to rely on complex and brittle image-understanding algorithms, making reliable scaling of

such approaches difficult (Jurrus et al., 2006; Jain et al., 2007; Macke et al., 2008; Helmstaedter et al., 2009; Mishchenko, 2009; Mishchenko et al., 2010b). In (Mishchenko, 2009; Mishchenko et al., 2010b), in particular, a complete automation approach was presented and reconstruction speed of about 5 μm^3/man-hour in dense neuropil in rat hippocampus was demonstrated, estimated to correspond to at least a 10-fold improvement over the purely manual analysis. Such automation in *C. elegans* potentially could allow another complete connectome reconstruction in 1-2 years of work by a single operator, with subsequent speedup possible by parallelizing. Of course, this estimate should be only viewed as a rough, order of magnitude estimate, since many different factors will play role in the actual performance including simpler longitudinal organization of axons in *C. elegans* (which may help reconstruction), their smaller size (which may make reconstruction more difficult), etc. Also note that (Mishchenko, 2009; Mishchenko et al., 2010b) deals with full volume neural tissue reconstructions in which the contours of all neurons are completely recovered. Computer assisted manual skeleton-based reconstructions such as recently reported in neuroscience conferences by W. Denk potentially may allow yet faster analysis of EM data.

PCR reconstruction paradigm comprises: a) stimulating different neurons individually electrically or optically; b) observing sub-threshold electrical responses in a selected neuron using an electrode patch. While this approach is responsible for the bulk of the information available today about neurons and their circuits, such measurements are also notoriously difficult to perform. Normally, only as few as 10-100 neural pairs can be patched and tested a day. In *C. elegans*, in particular, this implies that a complete scan of the connectivity matrix would take anywhere from 3 to 30 years to complete, although we should also note that patching neurons in *C. elegans* is admittedly more difficult due to their extremely small size. With optical stimulation and optical readout techniques such as ChR2 and voltage sensitive dyes, however, this time can be dramatically reduced, making complete scans possible in the span from several minutes to hours. We will further revisit this option a little later in this section.

We inspected performance of EM and PCR reconstructions in terms of the degree of the *reconstruction effort* involved (Figure 9A and C) and noise and contributed errors (Figure 9B and D). See Section 2.7. for the details of these calculations. In EM, reconstruction effort refers to the fraction of the neural tissue volume imaged in respect to the total volume of the neural circuit, and in PCR this corresponds to the fraction of all neural pairs recorded from. Errors in EM reconstructions are typically introduced via point errors in the traces of thin axons, e.g., such as confusing and mixing similar nearby axons or "losing" axons due to local mistakes in interpretation of images or defects in the photographs (Mishchenko, 2009; Mishchenko et al., 2010b). Because each such error affects a large number of synapses located downstream on the affected axon (e.g., causing all such synapses to be lost or mis-assigned to a different neuron), EM reconstructions can be very sensitive to such small errors. In Fig. 9B, in particular, the quality of EM reconstructions in *C. elegans* vs. the rate of such errors is shown. The error rate is characterized by the probability of an error in the trace of an axon on its interval between two consecutive synapses, *f*. For instance, if EM reconstruction contains on average one error in a trace per 1000-2000 serial EM sections, i.e., 50-100 μm under typical conditions, the chance of an error on the 5-10 μm axonal segment comprising the typical distance between two consecutive axonal boutons in mammals is $f \approx 10\%$. When tracing an axon from one synaptic connection to the next, then, there will be a 10% chance to make a point error affecting that reconstruction. From Fig. 9B, we observe that the largest error rate that can be tolerated in *C.*

elegans in such settings is only about 10%, or one error per 50-100 µm. Interestingly, an estimate of EM error rates in (Mishchenko, 2009; Mishchenko et al., 2010b) indicates that these may be in fact as large as $f \approx 5-10\%$. At the same time, also note that the above error rate bound may depend on the size of the reconstructed neural circuit and may become progressively smaller as the circuit size becomes larger than that in *C. elegans*.

In PCR, errors typically correspond to failures to detect a connection when recording from a pair of neurons. This can occur, e.g., due to failure of the target neuron to get stimulated or failure to observe a weak post-synaptic sub-threshold response in the recorded neuron, etc. Such errors are mathematically represented as point deletion errors in the connectivity matrix and can be characterized by their probability per one tested neural pair, *f*. From Fig. 9D we observe that PCR is quite robust to such errors, and error levels of up to 40-60% can be easily tolerated.

While we show that the proposed approach can be successful already today in *C. elegans*, it is important to think about possible strategies for its applications in larger systems, e.g., such as *Drosophila*. In particular, in *Drosophila* with $N \approx 10^5$ neurons and on average 100 synapses per neuron the neural circuit complexity is $N_{sp} \approx 10^7$ connections, and an experiment such as we described may appear prohibitive.

We must note that this problem is not confined to our method per se. Even though reconstruction of the circuit with 2500 connections in one *C. elegans* specimen had been completed with EM (White et al., 1986), and connectivity studies involving ≈ 1000 neural connections had been conducted in the past with other techniques (Ikegaya et al., 2005; Song et al., 2005; Broome et al., 2006; Pillow et al., 2008; Mishchenko et al., 2010a), scaling of any of these techniques to a circuit with $\approx 10^7$ connections is highly nontrivial to say the least.

Notwithstanding, three general directions can be proposed that can allow applying our strategy to a circuit as large as that in *Drosophila*. First, while we assumed that only a single measurement can be obtained from one stochastic GRASP animal, this does not need to be the case in principle. Whereas synaptic marker GRASP utilizes chemical reconstitution of a particular protein across synaptic cleft, it is bound to be essentially a single-color marker. However, recently the author had shown that mere spatial proximity of two different wavelength pre- and post-synaptic fluorochromes can allow detecting synapses without the need for such chemical reconstitution (Mishchenko, 2010). In this case, fluorochromes can be multiplexed onto synapses in large numbers, allowing for up to 2^{N_c} different synaptic labels for N_c distinct color fluorochromes [similar in essence to the idea employed in the Brainbow mouse (Livet et al., 2007; Lichtman et al., 2008)]. This case can be accommodated by our framework without any modifications - the number of measurements obtained from one animal then will be 2^{N_c} instead of 1. This will allow accelerating data acquisition dramatically and may allow reconstruction of the neural circuits as large as that in *Drosophila* by imaging, e.g., 10 synaptic fluorochromes in 1000 animals.

Second, one may consider formulating connectivity reconstructions in terms of certain neural populations rather than individual neurons (Luo et al., 2008). As the role of individual neurons vs. neural classes in the animal behavior is still being debated, such approach may prove to be not only less practically involving but also more behaviorally relevant. The connectivity matrix in that case would describe couplings between different neural populations rather than individual neurons, and can be "sampled" using the same general strategy that we described here. In

particular, in *Drosophila* libraries of Gal4 lines provide particularly convenient tool for accessing as well as cataloguing such neural populations. Then, the connectivity matrix can be accessed in a series of experiments targeting synaptic markers to different neural populations using different Gal4 lines and their combinations, e.g., using the methods from (Luo et al., 2008). The expression patterns *a* and *b* in that case will be known a-priory from the identity of the Gal4 lines used in each experiment and will not need to be measured. The connectivity matrix in terms of the *smallest resolved groups of neurons* for the set of used Gal4 lines, as defined below, will be produced even if none of the neurons were targeted specifically in any of the experiments. An approximation to the true connectivity matrix given incomplete available data can be obtained and then continuously improved as more data is obtained.

Third, our strategy need not be constrained to use of anatomical synaptic markers but can be employed also with other modalities that may ultimately allow faster neural connectivity reconstructions. Of these, pair-wise recordings using optical stimulation [e.g. see (Bureau et al., 2004; Petreanu et al., 2007)] and optical observation of the membrane potential with voltage sensitive dyes [e.g. see (Baker et al., 2005)] appear to be the most promising. Post-synaptic sub-threshold responses after activation of a group of pre-synaptic neurons, at least in the first approximation, can be represented as linear sums of the connectivity matrix elements corresponding to these for individual neurons, thus, allowing application of the same formalism that we developed here. Specifically, we can represent post-synaptic sub-threshold response in neuron *i*, observed after stimulation of a group of neurons $b_j(t)$ during trial *t*, as follows,

$$V_i(t) = \sum_j C_{ij} b_j(t).$$

(6)

Here, $V_i(t)$ is a measure of the integrated evoked post-synaptic response (iEPSP) in the imaged neuron, and C_{ij} is the connectivity matrix describing the iEPSP strength for all different post- and pre-synaptic neurons in the circuit. As is easy to see, Eq. (6) is identical to Eq. (1) and can be analyzed using the same approach. Note that the ability to observe sub-threshold changes in the membrane potential is critical in Eq. (6) and, thus, the use of *voltage-sensitive dyes* and not *calcium sensitive dyes* is required.

Our method calls for identification of neurons expressing the synaptic marker during each experiment. Thus, another important question that can be asked is whether our strategy can be applied in the organisms where neurons are not individually identifiable. While it is known how to identify neurons in *C. elegans* and some other organisms such as *leech*, in general methods for and even possibility of identification of individual neurons in larger organisms are subject of significant debate. It is important to ask, therefore, how our approach can be applied in such settings where neurons cannot be individually identified.

This question, again, is not specific to our approach per se. A degree of "identifiability" is required for results of any neural connectivity experiment to be comparable or generalizable across different animals. Really, if individual neurons in an organism cannot be identified, this also means that it is impossible to co-relate neurons in different animals. In turn, this means that neural connectivity reconstructions obtained in one animal cannot be compared with that obtained in another and, even more generally, any similar experimental results in one animal cannot be compared or generalized to that in another. "Identifiability" in some form should

always be available and, indeed, it is always present in neuroscience experiments. In *C. elegans*, specifically, neurons can be identified individually; in *Drosophila* genetic lineages and Gal-4 lines are used to identify neurons and their populations; in higher organisms morphological classes and anatomical landmarks such as cortical layers, neural nuclei, and functional areas are used to characterize connectivity. If any such system is available in a target organism, then the framework developed in this work can be immediately applied in the context of such a system.

Specifically, consider a case where one has a large neural circuit and a system of different morphological classes for classification of its neurons. In each experiment, then, one can count neurons expressing the synaptic marker in such different morphological classes. Such counts can be used to replace the indicator functions $a_i^{(k)}$ and $b_j^{(k)}$ in our formalism, and the connectivity matrix C_{ij} can be subsequently calculated from such measurements following exactly the same procedure as described in Section 2.5. In this case, clearly, i and j will refer to the different morphological classes and not individual neurons, and C_{ij} will describe the connectivity in the neural circuit as it exists between such classes. Similarly, if one has an atlas of anatomical landmarks such as a map of brain areas or neural nuclei, one can count how many neurons expressed the synaptic marker in different areas and use such counts in place of the indicator functions $a_i^{(k)}$ and $b_j^{(k)}$ to recover the connectivity matrix in terms of such anatomical landmarks.

Libraries of genetic classes, such as Gal4 lines or lineages in *Drosophila*, have recently emerged as a powerful tool for classifying neural structures in large neural systems (Phelps and Brand, 1998; Luo et al., 2008). Such libraries can provide the means for identifying and cataloguing neurons even when morphology and location cannot be used for that reliably. Having a library consisting of certain genetic lines *A*, *B*, *C,* etc., one can associate with each neuron a pattern of 1's and 0's describing whether that neuron is present in the expression patterns of lines *A*, *B*, *C*, etc. For example, pattern "110…" can correspond to a neuron that is present in the expression patterns of lines *A* and *B* but not *C*, etc. Different such patterns will constitute an identification system for the neural structures, whereas different neurons will be identified by the combinations of the genetic lines containing them. One does not need the ability to detect or identify these neurons directly anatomically: as long as such neurons can be accessed in reproducible manner by manipulating the Gal4 lines, experiments with such groups of neurons can be conducted and connectivity between them as well as their properties can be reliably determined. Note that neurons that are present in all the same lines will be indistinguishable and equivalent for such an identification system, and will constitute what we shall call the *smallest resolved group of neurons,* or the equivalence class in mathematical terms, in such an identification system (see Table 1). By introducing synaptic markers into the neural circuit using such different genetic lines and their combinations and performing the measurements and the calculations that we described above the reconstruction program can be completed successfully. Detailed connectome in terms of the smallest resolved groups of neurons for the set of Gal-4 lines will be recovered.

One may argue that such "class-wise" reconstructions that we thus described fall far short from the ideal neuron-wise connectome that EM can potentially deliver. However, class-wise reconstructions already contain substantial amount of significant information and can be of substantial interest to researchers. Class-wise connectomes in fact had been already widely used in neuroscience research in the systems where individual neurons cannot be identified, including

large EM studies of the optic lobe organization in *Drosophila* (Meinertzhagen and Sorra, 2001), studies of the organization of the cortical micro-circuits in the mouse (Shepherd and Svoboda, 2005), etc. Where neurons in principle may not have explicit "individual identities", class-wise connectomes may not only prove to be significantly more practically accessible, but also may provide a more physiologically adequate viewpoint on the organization of the neural circuits (Luo et al., 2008).

5. Conclusion

We describe a new neural connectivity reconstruction paradigm that utilizes random or pseudo-random sampling of neural connectivity with anatomical fluorescent synaptic markers localized genetically or otherwise to different parts of a neural circuit. Synaptic connectivity matrix is recovered statistically from a collection of fluorescent measurements obtained with such a probe. While conventional neural connectivity reconstruction paradigm focuses on finding individual synapses in neuropil and explicitly relating these with remotely positioned neurons, we propose here to collect large samples of simple fluorescent measurements of connectivity and combine these using minimal model assumptions to determine detailed configuration of the probed neural circuit.

The key contribution of this paper is to recognize that physical connectivity in a neural circuit can be recovered using a sample of simple observations produced with anatomical fluorescent synaptic markers such as GRASP. We observe that a certain type of fluorescent measurements that can be obtained with such markers can be mathematically represented as linear sums over the elements of the synaptic connectivity matrix. This allows us to propose an algorithm for accurate and efficient reconstruction of such connectivity matrix from such observations. *Compressive Sensing* (CS), in particular, is especially attractive framework for such an analysis since neural connectivity is expected to be sparse: in *C. elegans*, e.g., the total number of connections is ≈2500 out of the total possible ≈90,000 and sparseness is 0.03; in *Drosophila* the number of connections per neuron is ≈100 out of the total possible ≈100,000 and sparseness is 10^{-3}, etc. CS algorithms are mathematically tractable and provably nearly optimal for reconstructions of sparse signals from ensembles of linear measurements.

The method that we proposed potentially can allow high speed, high detail connectome reconstructions in many model organisms in the future. The advantages of the proposed approach are: the method relies on relatively fast and easy to obtain fluorescent light microscopy measurements; the measurements are simple – the total count or the total fluorescence strength of labeled synapses and the patterns of co-expressed nuclei tags; analysis of the data is computationally straightforward and uses well understood statistical methodologies; tracing of neural projections or any other explicit association of synapses with neurons over macroscopic scales is not required; highly detailed connectivity maps can be produced without targeting individual neurons or their small groups. Practically meaningful strategies are available for extensions to larger organisms such as *Drosophila* including the use of multiplexed fluorescent synaptic markers (Mishchenko, 2010), use of libraries of genetic lines such as Gal4 lines (Luo et al., 2008), and use of different modalities such as voltage sensitive dyes (Baker et al., 2005), as described in the Discussion. The approach can yield the average connectivity matrix for an entire population of animals at once as well as the variability in the connectivity matrix, which can be

calculated, e.g., using bootstrapping – producing reconstructions from different parts of the same dataset.

Among the disadvantages of the present approach is its relatively high sensitivity to noise. In particular, stringent conditions on the data quality should be maintained if one wishes to recover useable reconstructions with the near-minimal set of measurements. The method yields the average connectivity matrix and, thus, can only be used to study general or stereotypical neural connectivity or gross structures and not the connectivity in *individual* animals, as opposed to the techniques of electron microscopy and pair-wise cell recordings. The discussion that we presented here is theoretical and no actual implementation or experimental data are shown. Although we consulted to the fullest degree the latest literature and experimentalists to realistically assess the approach's feasibility and possibility of its practical implementation (Livet et al., 2007; Micheva and Smith, 2007; Feinberg et al., 2008; Long et al., 2008; Bargmann, Personal communication; Kerr, Personal communication), the impact of a number of different experimental factors remains unknown and cannot be assessed now due to lack of experimental data. For example, it is not yet known quantitatively how specific synaptic marker labelings can be, how effective GRASP can be in inhibitory vs. excitatory synapses, how well pan-neuronal expression can be achieved and how uniform such expression can be, etc. Unfortunately, answering these questions is currently beyond the limited capabilities of the author, who does not have access to experimental facilities or resources needed to carry out extensive experimental program associated with answering these questions. While recognizing the many uncertainties that remain, the author believes that the data available in the literature today support a very favorable view on the prospects of the practical implementations and utilization of the described approach (Livet et al., 2007; Micheva and Smith, 2007; Feinberg et al., 2008; Long et al., 2008; Bargmann, Personal communication; Kerr, Personal communication).

This work opens many exciting possibilities for new quantitative data-rich studies of neural connectivity in large neural circuits. In *C. elegans*, in particular, described approach in principle allows reconstructions of complete wiring diagram in the span of several days using existing genetic and off-the-shelf fluorescent light microscopy tools and straightforward data processing. For comparison, the alternative approach of serial electron microscopy would require at least 1-2 years of imaging and analysis to obtain reconstruction of a single neural circuit of comparable size, employing sophisticated and expensive high-end imaging equipment and complex and fragile image understanding algorithms. Described approach can be also employed in larger organisms such as *Drosophila* using different connectivity probes and/or genetic targeting techniques.

Acknowledgements

References

Baker BJ, Kosmidis EK, Vucinic D, Falk CX, Cohen LB, Djurisic M, Zecevic D (2005) Imaging brain activity with voltage- and calcium-sensitive dyes. Cellular and Molecular Neurobiology 25:245-282.

Bargmann CI (Personal communication).

Bohland J et al. (2009) A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. arXiv:0901.4598.

Boyd S, Vandenberghe L (2004) Convex Optimization: Oxfor University Press.

Bregman LM (1965) The method of successive projection for finding a common point of convex sets. Soviet Math Dokl 6:688-692.

Briggman KL, Denk W (2006) Towards neural circuit reconstruction with volume electron microscopy techniques. Curr Opin Neurobio 16:562.

Broome BM, Jayaraman V, Laurent G (2006) Encoding and decoding of overlapping odor sequences. Neuron 51:467-482.

Bureau I, Shepherd GM, Svoboda K (2004) Precise development of functional and anatomical columns in the neocortex. Neuron 42:789-801.

Candes EJ, Romberg J (2005) Practical signal recovery from random projections. In: SPIE Conference on Wavelet Applications in Signal and Image Processing XI.

Candes EJ, Wakin M (2008) An introduction to compressive sampling. IEEE Signal Processing Magazine 25:21-30.

Candes EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52:489-509.

Donoho D (2006) Compressed sensing. IEEE Trans on Information Theory 52:1289-1306.

Feinberg EH, Vanhoven MK, Bendesky A, Wang G, Fetter RD, Shen K, Bargmann CI (2008) GFP Reconstitution Across Synaptic Partners (GRASP) defines cell contacts and synapses in living nervous systems. Neuron 57:353-363.

Gustafsson MGL (2000) Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. J Microsc 198:82.

Gustafsson MGL (2005) Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. PNAS:13081.

Hagmann P, Kurant M, Gigandet X, Thiran P, Wedeen V, Meuli R, Thiran J-P (2007) Mapping Human Whole-Brain Structural Networks with Diffusion MRI. PLoS ONE 2:e597.

Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey C, Wedeen V, Sporns O (2008) Mapping the Structural Core of Human Cerebral Cortex. PLoS Biology 6:e159.

Helmstaedter M, Briggman KL, Denk W (2009) 3D structural imaging of the brain with photons and electrons. Curr Opin Neurobio:Epub.

Ikegaya Y, Le Bon-Jego M, Yuste R (2005) Large-scale imaging of cortical network activity with calcium indicators. Neuroscience Research 52:132-138.

Jain V, Murray J, Roth F, Turaga S, Zhigulin V, Briggman K, Helmstaedter M, Denk W, Seung H (2007) Supervised learning of image restoration with convolutional networks. . In: International Conference on Computer Vision, pp 1-8.

Jones LM, Fontanini A, Sadacca BF, Miller P, Katz DB (2007) Natural stimuli evoke dynamic sequences of states in cortical ensembles. Proceedings of the National Academy of SciencesJournal of Neuroscience 104:18772-18777.

Jurrus E, Tasdizen T, Koshevoy P, Fletcher P, Hardy M, Chien C, Denk W, Whitaker R (2006) Axon tracking in serial block-free scanning electron microscopy. In: MICCAI 2006 workshop. Copenhagen.

Kerr R (Personal communication).

Lichtman JW, Livet J, Sanes JR (2008) A technicolour approach to the connectome. Nat Rev Neurosci Epub ahead of print.

Livet J, Weissman TA, Kang H, Draft RW, Lu J, Bennis RA, Sanes JR, Lichtman JW (2007) Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature 450:56-62.

Long F, Peng H, Liu X, Kim S, Myers E (2008) Automatic recognition of cells (ARC) for 3D images of C. Elegans. In: Lecture Notes in Computer Science: Research in Computational Molecular Biology, pp 128-139: Springer Berlin

Lu J, Fiala JC, Lichtman JW (2009a) Semi-automated reconstruction of neural processes from large numbers of fluorescence images. PLoS ONE 4:e5655.

Lu J, Tapia J, White O, Lichtman JW (2009b) The interscutularis muscle connectome. PLoS Biology 7:e1000108.

Luo L, Callaway EM, Svoboda K (2008) Genetic dissection of neural circuits. Neuron 57:634-660.

Macke J, Maack N, Gupta R, Denk W, Scholkopf B, Borst A (2008) Contour-propagation algorithms for semi-automated reconstruction of neural processes. J Neurosci Methods 167:349-357.

Meinertzhagen IA, Sorra KE (2001) Synaptic organisation in the fly's optic lamina: few cells, many synapses and divergent microcircuits. Progress Brain Research 131:53-69.

Micheva KD, Smith SJ (2007) Array tomography: a new tool for imaging the molecular architecture and ultrastructure of neural circuits. Neuron 55:25-36.

Mishchenko Y (2009) Automation of 3D reconstruction of neural tissue from large volume of conventional Serial Section Transmission Electron Micrographs. Journal of Neuroscience Methods 176:276-289.

Mishchenko Y (2010) On optical detection of densely labeled synapses in neuropil and mapping connectivity with combinatorially multiplexed fluorescent synaptic markers. PLoS ONE 5:e8853.

Mishchenko Y, Vogelstein J, Paninski L (2010a) A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. Annals of Applied Statistics To appear

Mishchenko Y, Hu T, Spacek J, Mendenhall J, Harris KM, Chklovskii D (2010b) Ultrastructural analysis of hippocampal neuropil from the connectomics perspective. Neuron 67:1009-1020.

Petreanu L, Huber D, Sobczyk A, Svoboda K (2007) Channelrhodopsin-2-assisted circuit mapping of long-range callosal projections. Nature neuroscience 10:663-668.

Phelps CB, Brand AH (1998) Ectopic gene expression in Drosophila using GAL4 system. Methods 4:367-379.

Pillow J, Shlens J, Paninski L, Sher A, Litke A, Chichilnisky EJ, Simoncelli E (2008) Spatiotemporal correlations and visual signaling in a complete neuronal population. Nature 454:995-999.

Romberg J (2008) Imaging via compressive sampling. IEEE Signal Processing Magazine 25:14-20.

Sarkar M, Magliery TJ (2008) Re-engineering a split-GFP reassembly screen to examine RING-domain interactions between BARD1 and BRCA1 mutants observed in cancer patients. Mol Biosyst 4:599-605.

Shepherd GM, Svoboda K (2005) Laminar and columnar organization of ascending excitatory projections to layer 2/3 pyramidal neurons in rat barrel cortex. Journal of Neuroscience 25:6037-6046.

Smith SJ (2007) Circuit reconstruction tools today. Curr Opin Neurobiol 17:601-608.

Song S, Sjostrom PJ, Reiql M, Nelson S, Chklovskii DB (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biology 3:e68.

Sporns O, Tononi G, Kotter R (2005) The Human Connectome: A Structural Description of the Human Brain. PLoS Computational Biology 1:e42.

Stevenson IH, Rebesco JM, Hastopoulos NG, Haga Z, Miller LE, Kording KP (2009) Bayesian inference of functional connectivity and network structure from spikes. IEEE Trans Neural Systems and Rehab 17:203-213.

Svoboda K (Personal Communication) Cajal 2.0 Cajal 2.0 optical mapping of mouse brain. In.

Vanderbei RJ (2001) Linear Programming: Foundations and Extensions. New York: Springer-Verlag.

White J, Southgate E, Thomson JN, Brenner S (1986) The structure of the nervous system of the nematode caenorhabditis elegans. . Phil Trans Royal Soc London Series B, Biological Sciences 314:1-340.

Wright S (1997) Prima-dual interior-point methods: SIAM.

Figures Legends and Tables Legends

Figure 1: Schematic description of GRASP (Feinberg et al., 2008), two-component co-localization synaptic marker (Mishchenko, 2010), and the Cre/Lox system for stochastic expression of genes (Livet et al., 2007). A) Two fragments of a split-GFP (sA and sB) are expressed separately at post-synaptic (left) and pre-synaptic (center) sites of different neurons. Separately, split-GFP fragments do not produce fluorescence. At the location of synapses split-GFP can recombine into a functional GFP molecule and produce fluorescence, thus, fluorescently tagging respective synapses (right). B) Two different wavelength fluorochromes (sA and sB) are used to tag separately post-synaptic (left) and pre-synaptic (center) surfaces of different neurons. Co-localization of fluorescence from such fluorochromes can be used to detect and identify synapses with high accuracy (right). C) Schematic description of the Cre/Lox system for stochastic expression of genes. A cassette with two genetic sequences for two different wavelength fluorochromes (red and green, left) in mutually inverted orientations is introduced into the genome and flanked with inversely-oriented loxP-sites. When Cre is introduced into cells, recombinase reacts with loxP-sites in such a way that orientation of the flanked sequence is flipped at certain rate. When Cre is removed, the sequence may be found in either original or reversed orientation (right). When fluorochrome sequences in the cassette are transcribed, only those in the direct orientation can be produced successfully. Thus, different cells produce either green or red fluorochromes at random.

Figure 2: Schematic description of the action of the stochastic synaptic marker, combining a a two-component fluorescent synaptic marker (Figure 1A-B) with a recombinase system for stochastic gene expression (Figure 1C) . A) In the first implementation, a cassette with two inversely oriented sequences encoding pre- and post-synaptic marker fragments (sA and sB) and associated nuclei-bound fluorescent proteins (nA and nB) is introduced into genome and flanked with inversely-oriented loxP-sites. Effect of Cre is to randomize the orientation of such cassette in different neurons leading to two recombination outcomes: (i) post-synaptic fragment and associated nuclei XFP are expressed, tagging post-synaptic sites and nuclei of such neurons, and (ii) pre-synaptic fragment and associated nuclei XFP are expressed. A promoter (p) can be used to additionally restrict cassette expression to a defined population of neurons. B) In the second

implementation, a cassette with two loxP-flanked sequences for pre- and post-synaptic marker fragments is introduced into genome. Effect of Cre is to randomize orientation of both sequences leading to four recombination outcomes: (i) pre- and post-synaptic fragments are expressed together in a neuron, (ii) post-synaptic fragment is expressed only, (iii) pre-synaptic fragment is expressed only, and (iv) neither of the fragments is expressed. C) When two neurons form a synapse while expressing the synaptic marker fragments complementary a fluorescent punctum is formed. Nuclei of the cells expressing the synaptic marker are labeled at the same time with the associated nuclei XFP and also can be observed.

Figure 3: Schematic description of the neural connectivity reconstructions using stochastically expressed synaptic markers. A) In different animals random neurons in a circuit express the synaptic marker. Three random expression patterns (phenotypes) are shown for illustration. In each pattern all synapses formed by "affected" neurons produce same-color fluorescent puncta that can be observed with a light microscope. An overall measure such as the total count or the combined size of all labeled puncta is produced. B) For reconstruction of the connectivity matrix it is sufficient to collect the above overall measurements, n, along with the pre- and post-synaptic marker fragments expression patterns, a and b, for a sample of such stochastic phenotypes. C) The sample (n,a,b) can be mathematically represented as a sample of linear measurements over the connectivity matrix. Specifically, n can be related to certain sums over the connectivity matrix elements that correspond to the intersection of the rows and columns associated with the neurons expressing the pre- and post-synaptic marker fragments. We recognize in these settings an instance of Compressive Sensing problem (Donoho, 2006; Candes and Wakin, 2008) that allows tractable recovery of the synaptic connectivity matrix from a sample of such data.

Figure 4: Result of a hypothetical neural connectivity reconstruction experiment in *C. elegans* simulated using real wiring diagram for that animal available from electron microscopy (White et al., 1986). Shown are the scatter plots of the reconstructed connectivity weights versus that for the true connectivity weights. Reconstructions with 4000-6000 measurements can be already practically meaningful and the reconstruction with 10,000 measurements is exact.

Figure 5: Result of a hypothetical neural connectivity reconstruction experiment in *C. elegans* simulated using real wiring diagram available from electron microscopy (White et al., 1986). Shown is the correlation coefficient square, r^2, for the true and reconstructed connectivity weights as a function of the number of measurements K. Reconstructions from 4000-6000 measurements are already practically meaningful and the reconstruction from 10,000 measurements is exact. No significant difference in the performance is observed when using exclusive (Fig. 2A) or non-exclusive (Fig. 2B) constructs. These results allow us to estimate that complete wiring diagram in *C.Elegans* could be re-obtained using described approach in the span of 1-7 days. Result of a similar reconstruction using simpler L_2 regularized algorithm is shown to emphasize dramatic improvement conferred by Compressive Sensing algorithms.

Figure 6: Impact of biological variability on the reconstruction of neural connectivity using stochastic GRASP. Biological variability describes possible variation in the connection matrix from one animal to another. Different degrees of variability from 0 (no variability) to 100% (connection weights in different animals are purely random with a Poisson statistics) are shown. In all cases the impact of biological variability is found to be insignificant.
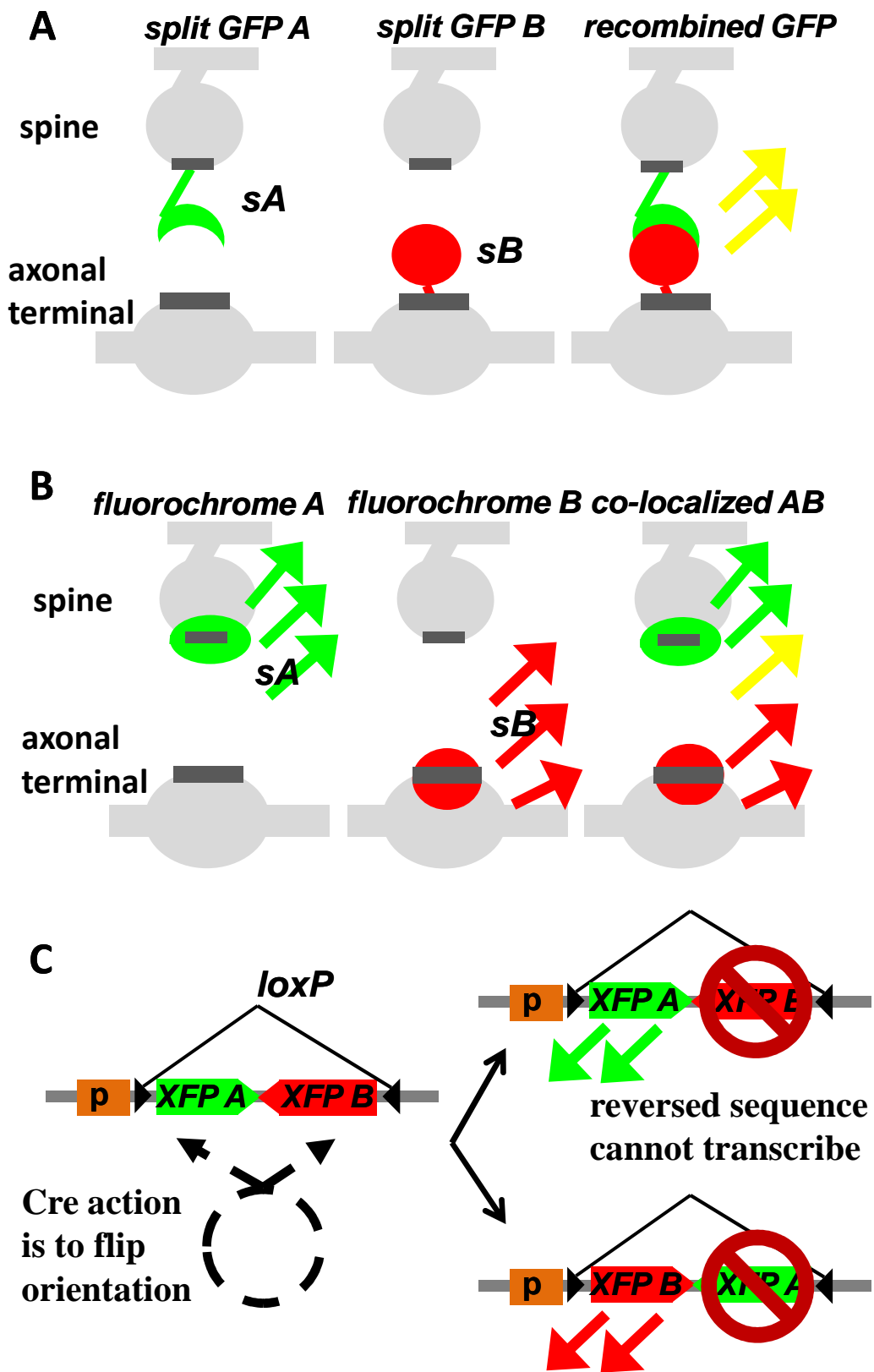
Figure 7: Impact of added noise in the measurements on the reconstruction of neural connectivity using stochastic GRASP. Although the reconstructions are stable (i.e., they do not break down in the presence of small amounts of noise), the sensitivity is substantial and only up to 4-5% of added noise can be tolerated at the minimal $K{\approx}10{,}000$. Larger sample size can be used to mitigate the impact of higher levels of added noise.
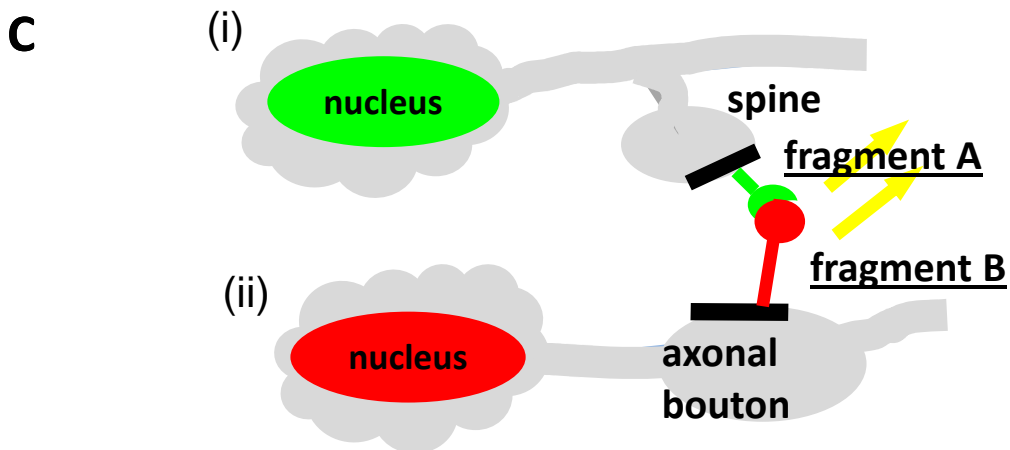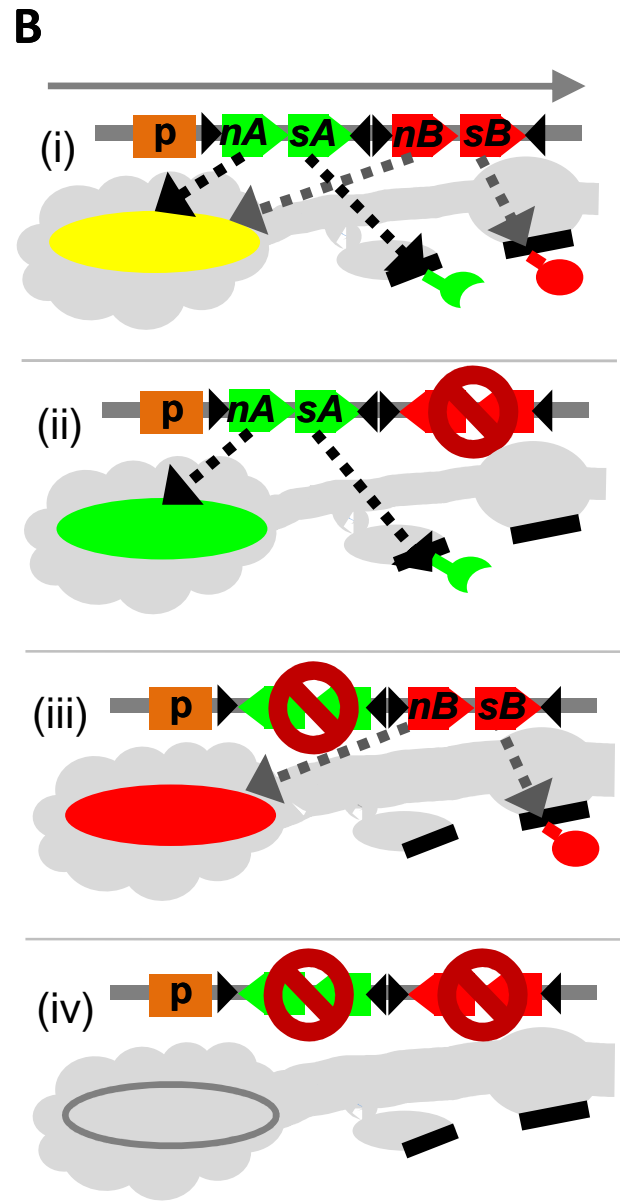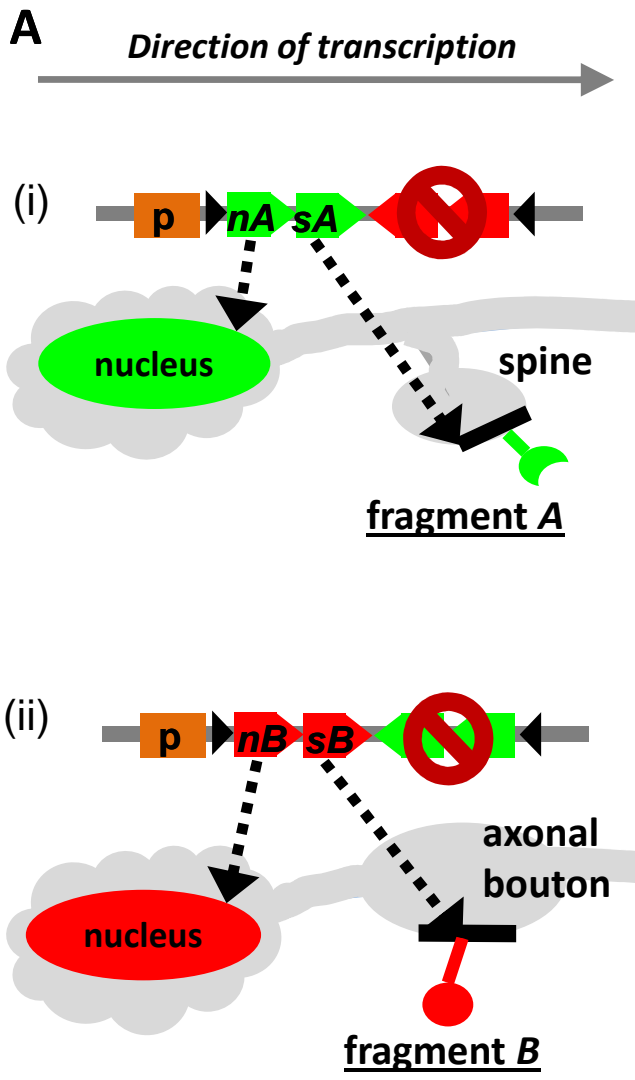
Figure 8: Impact of errors in detected GRASP expression patterns on the reconstruction of neural connectivity using stochastic GRASP. Although the reconstructions are stable, the sensitivity is again high with only up to 5-6% of errors that can be tolerated at the minimal $K{\approx}10{,}000$. Larger sample size can be used to mitigate the impact of such errors at higher rates.

Figure 9: Comparison with two different prospective connectomics paradigms, namely, serial electron microscopy (EM, panels A and B) and en-mass pair-wise cell recordings (PCR, panels C and D). Reconstructions are characterized by the reconstruction effort, E, where $E{=}1$ corresponds to the effort sufficient to obtain a single complete reconstruction (i.e. image one entire circuit). In *C. elegans*, we estimate that $E{=}1$ currently corresponds to $\approx$1-2 man-years of work for EM and $\approx$3-30 man-years of work for PCR. A) Quality of EM reconstructions as a function of the reconstruction effort in the ideal case (i.e., no noise, compare Figure 5). B) Quality of EM reconstructions as a function of the reconstruction effort in the presence of errors (compare Figure 7 and 8). Errors are characterized by the probability of one point error in a trace of an axon on its interval between two subsequent synapses. Due to long reach of such point errors EM reconstructions are vulnerable to their small amounts. C) Quality of PCR reconstructions as a function of the reconstruction effort in the ideal case. D) Quality of PCR reconstructions as a function of the reconstruction effort in the presence of errors. Errors are characterized by the probability of an error in the detection of a connection between two random cells. PCR reconstructions are found to be robust to such point errors and can successfully tolerate their large amounts.

Table 1: Libraries of genetic lines such as Gal4 lines in *Drosophila* can provide a natural classification and identification system for neural circuits. Each neuron can be identified by the set of genetic lines in which it appears, i.e. for each neuron one can associate a pattern of 1's and 0's corresponding to the sets of genetic lines containing it. E.g., a pattern 110… identifies a neuron that is present in lines *A* and *B* but not line *C*, etc. Neurons that are present always together in the same genetic lines comprise groups of indistinguishable, equivalent neurons with respect to such identification system. Such groups can be accessed and manipulated via associated combinations of genetic lines and the methods from (Luo et al., 2008). This allows conducting experiments involving such neurons and studying their properties even if it is not possible to locate and identify such neurons anatomically.
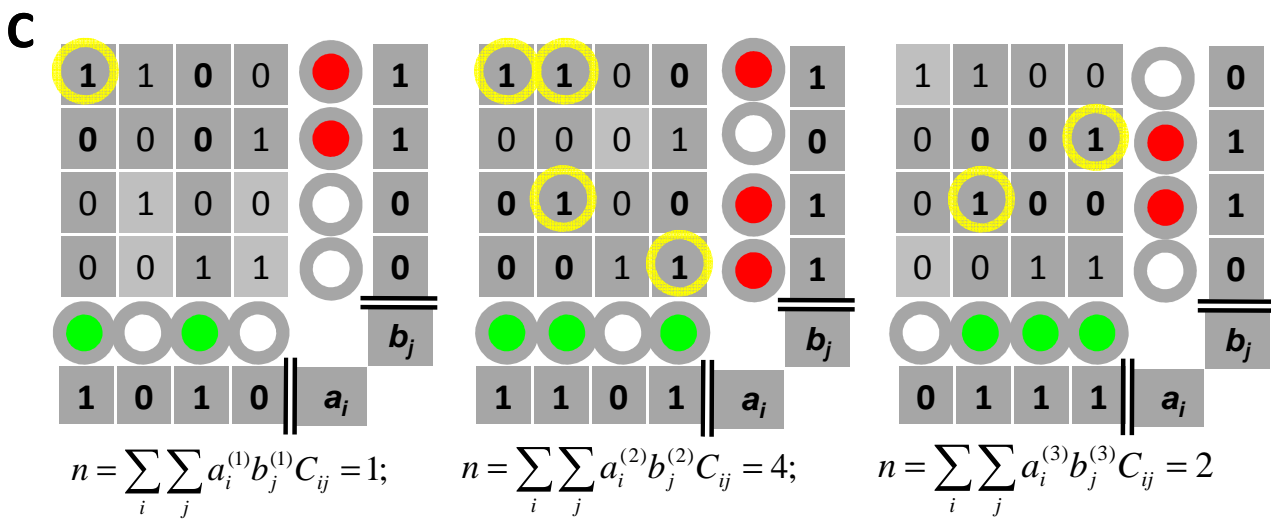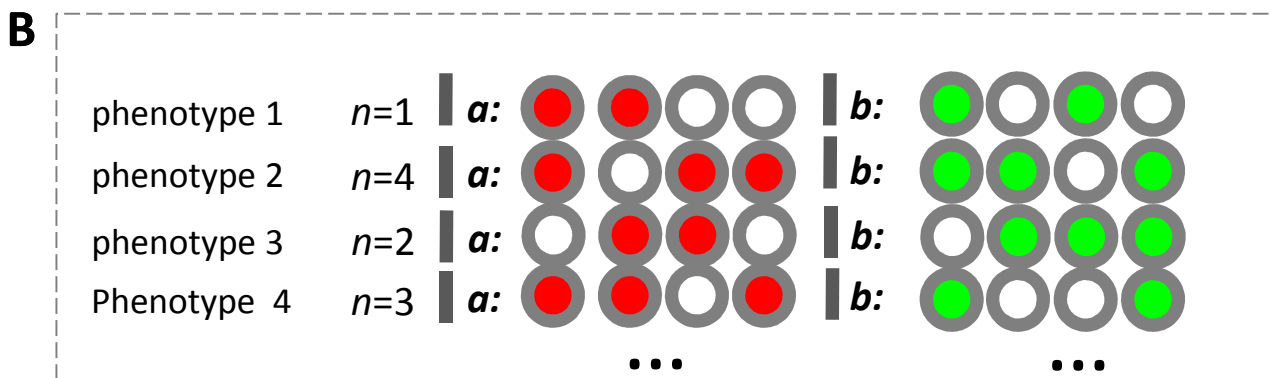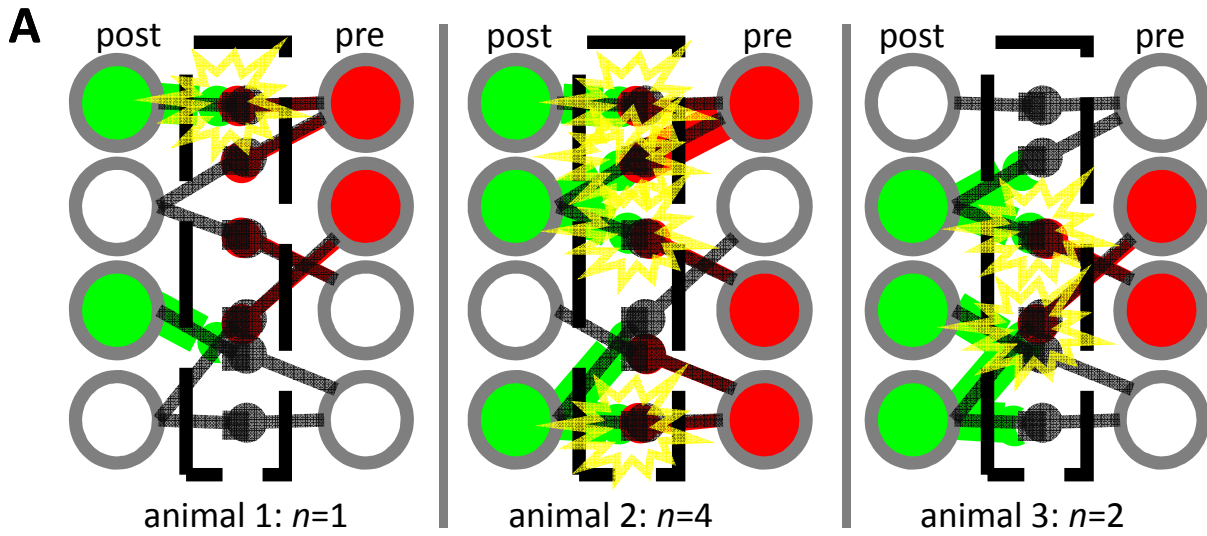
Mishchenko Figure 1

A

split GFP A   split GFP B   recombined GFP

spine

sA

axonal
terminal

sB

B

fluorochrome A   fluorochrome B   co-localized AB

spine

sA

axonal
terminal

sB

C

loxP

p   XFP A   XFP B

Cre action
is to flip
orientation

p   XFP A   XFP B

reversed sequence
cannot transcribe

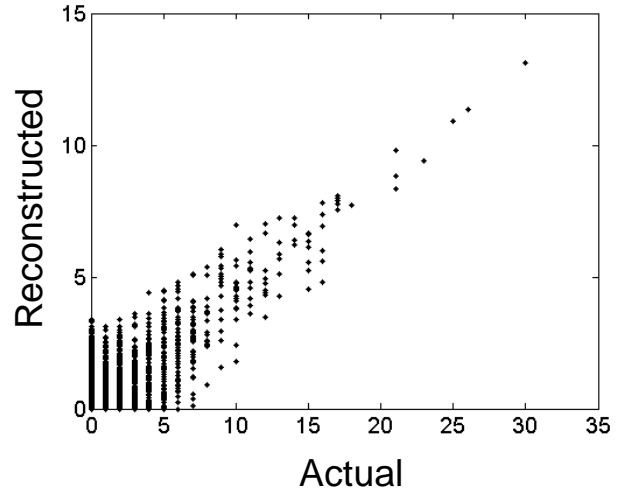p   XFP B   XFP A

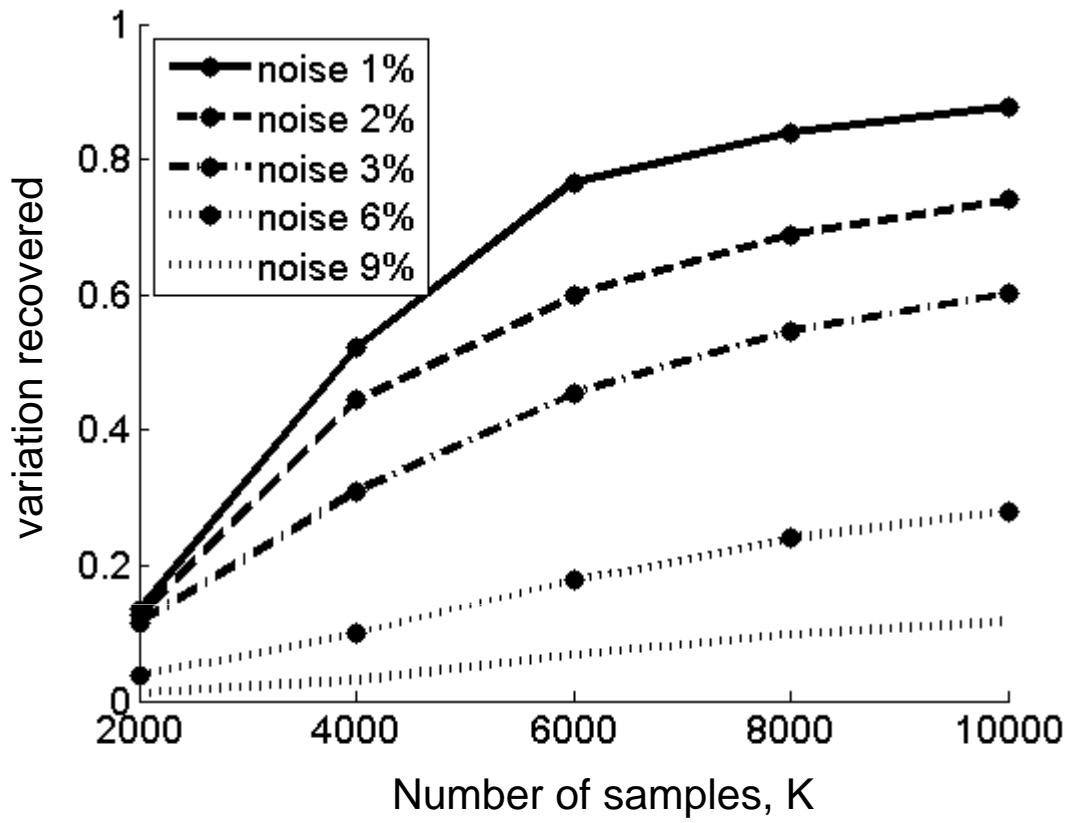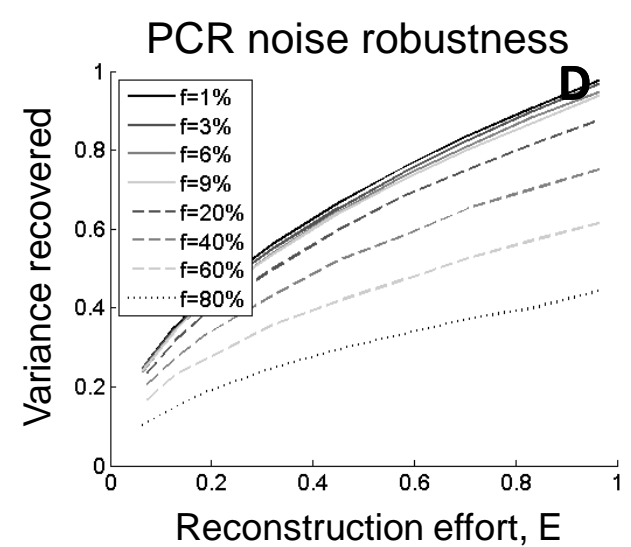Mishchenko Figure 2

Mishchenko Figure 3

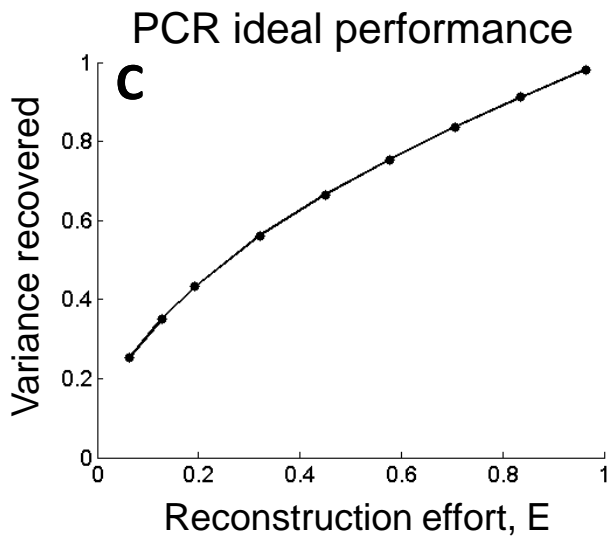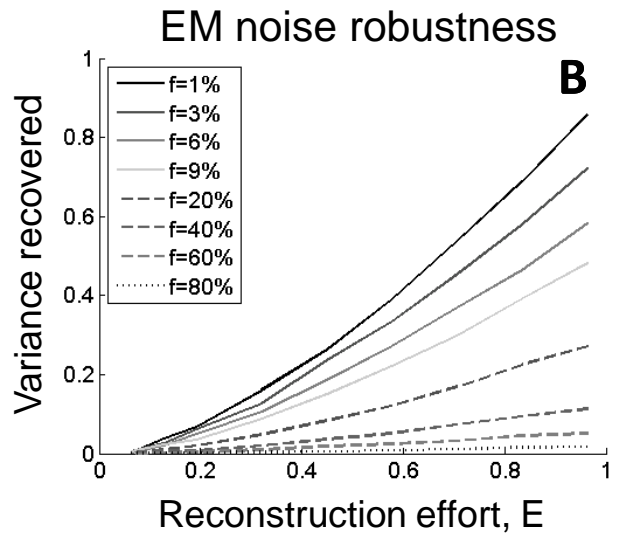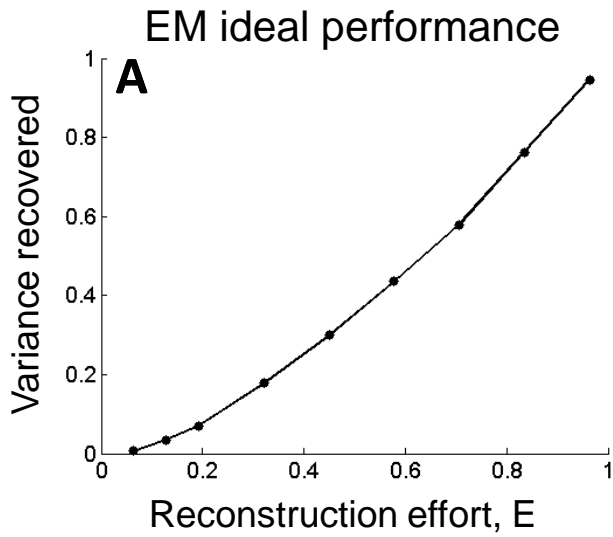# Mishchenko Figure 4
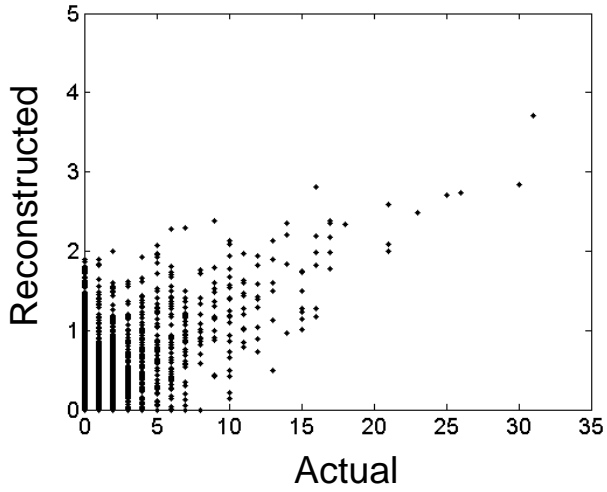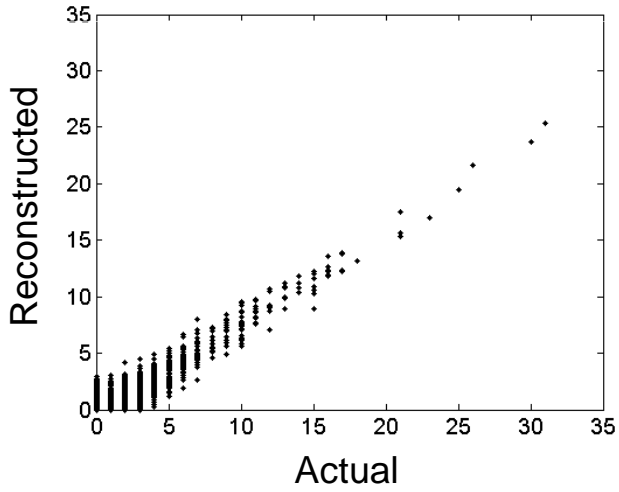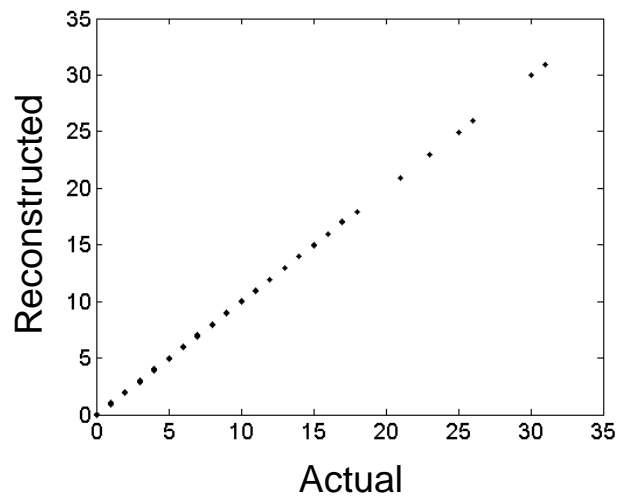
Mishchenko Figure 5
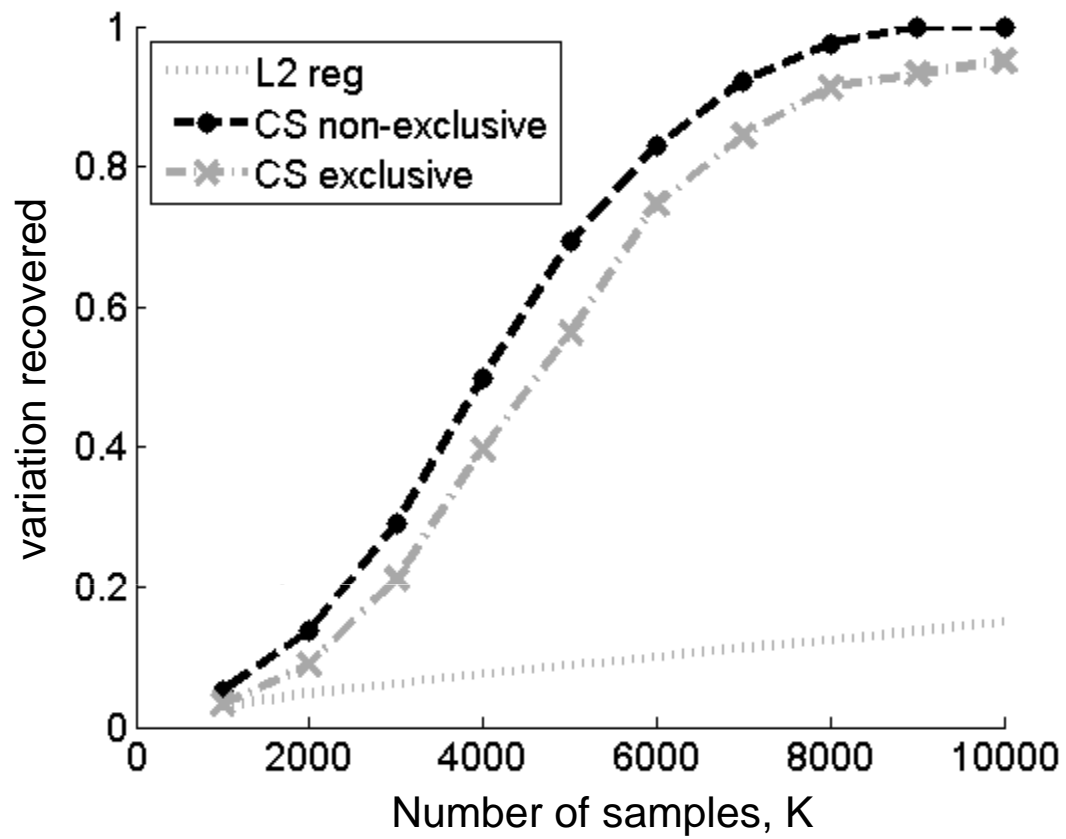
Mishchenko Figure 6
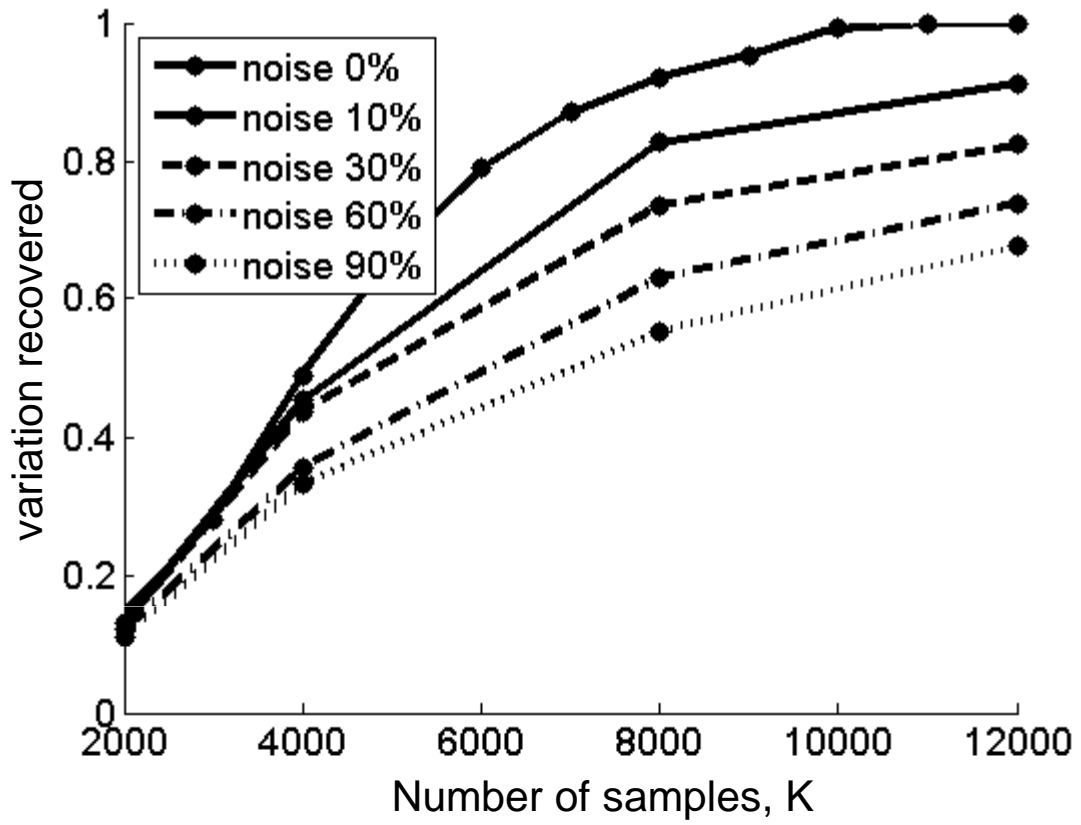
Mishchenko Figure 7

Mishchenko Figure 8
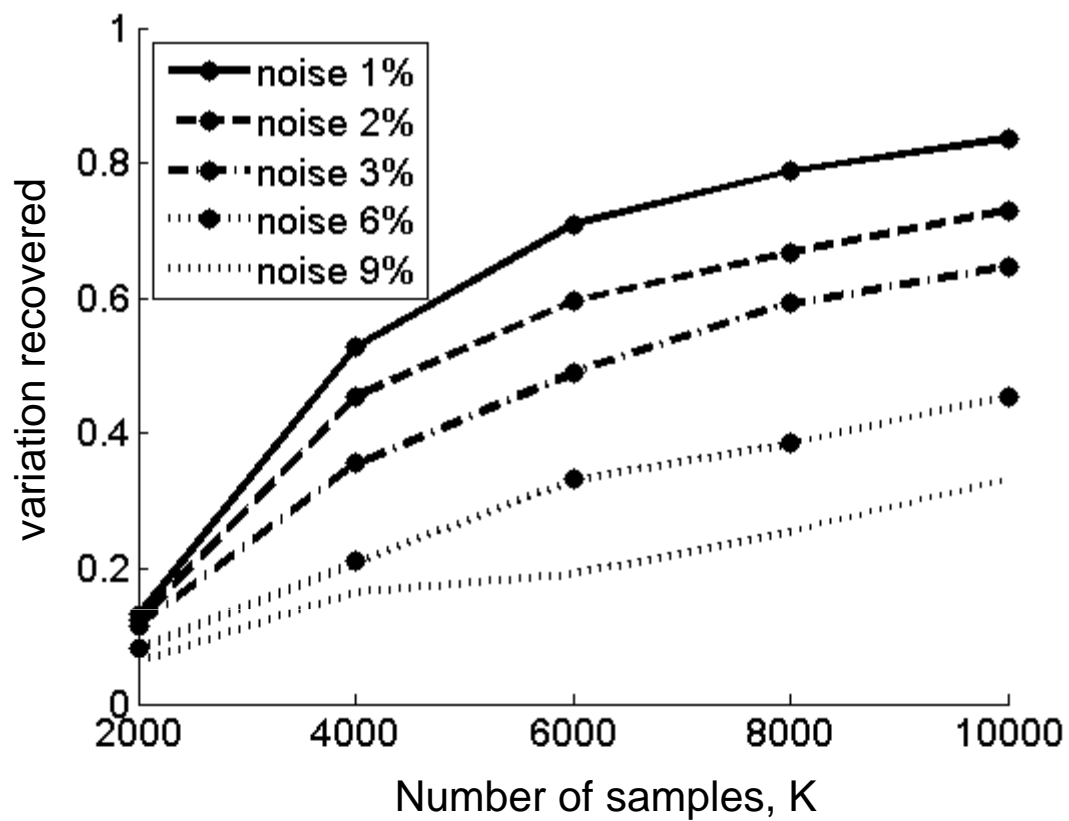
Mishchenko Figure 9

Mishchenko Table 1

Groups of neurons	Present in expression pattern of…			Code
	line A	*line B*	*line C*	
ABC…	yes	yes	yes	111…
BC…	no	yes	yes	011…
AC…	yes	no	yes	101…
AB…	yes	yes	no	110…
A…	yes	no	no	100…
B…	no	yes	no	010…
C…	no	no	yes	001…
…	no	no	no	000…