Neuroinform (2011) 9:133–142
DOI 10.1007/s12021-011-9114-1

COMMENTARY

Automated Reconstruction of Neural Tissue and the Role of Large-Scale Simulation

James Kozloski

Published online: 3 May 2011

Keywords DIADEM · High-throughput · Neural reconstruction · Neural tissue · Simulation · Compartmental modeling · Hodgkin Huxley model · Structural modeling · Developmental modeling · Neuron growth · Parallel computing · Data decomposition · Neural Tissue Simulator · Blue Gene

The brain implements a myriad of global brain functions to support adaptive behaviors. Despite their seeming innumerability, these emerge from combinations of lower level functions implemented by a relatively small set of brain tissues. Evidence from brain imaging studies shows that spatiotemporal patterns of activations across different brain tissues correlate with brain function (and hence with an organism's behavior). To support a diversity of global functions, gross connections between brain tissues, while structurally static, must undergo modulation. The strength of this modulation can define functional boundaries and interfaces between brain tissues: wherever functional relationships between brain regions are highly modulated, tissue boundaries occur.

Tissue-level functions, while also diverse, are more stereotyped than global brain functions. Similar to spatiotemporal modulation and recombination of tissue activation, variation and recombination of familiar structural elements of the brain (neurons and their connections, synapses) generate tissue-level functions. Unlike other organs' gross morphological specializations of single tissues (e.g., muscle, bone) brain specialization yields distinct tissues derived from stationary statistical combinations of a variety of neuron and synapse types in space, which we define as *microcircuitry*. Measurable, consistent patterning of microcircuitry across a tissue and in different organisms (i.e., *stereotypy*) further defines a tissue's boundaries: wherever patterning changes abruptly, one tissue ends and another begins.

Shepherd defined microcircuits abstractly and independent of neural tissues, based on simple computations they might implement.[1] Defining stereotyped microcircuitry as a stationary combination of neuron and synapse types within a specific tissue restricts strong synaptic plasticity to its boundaries. Where plasticity is strongest, stationary circuit components are recombined to serve underlying tissue-level functions, for example learning and memory.[2] Observations that strong departures from stereotypy in developing vertebrate tissue arise where neural competition dominates supports this view.[3] We therefore define microcircuitry circumscribed by strong plasticity as a *microcircuit*, which is then iterated to create a tissue.

For example, cerebellar tissue derives from a microcircuit iterated millions of times.[4] Boundaries between components occur at highly plastic parallel fiber synapses onto Purkinje cells. Similarly, neocortex derives from a microcircuit with stereotypical properties along its radial

J. Kozloski (✉)
IBM T.J. Watson Research Laboratories,
1101 Kitchawan Rd.,
Yorktown Heights, NY 10598, USA
e-mail: kozloski@us.ibm.com

[1] Shepherd, G. M. (2004) Introduction to synaptic circuits. In *Synaptic Organization of the Brain*. Shepherd G. M. (Ed.) New York:Oxford University Press. 1–38.

[2] Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: from synapses to maps. *Annual Review of Neuroscience*, 21, 149–186.

[3] Lu, J., Tapia, J. C., White, O. L., & Lichtman, J. W. (2009). The interscutularis muscle connectome. *PLoS Biol*, 7(2), e1000032.

[4] Voogd, J., & Glickstein, M. (1998). The anatomy of the cerebellum. *Trends in Cognitive Sciences*, 2(9), 307–313.

axis, iterated many millions of times within the cortical plane.[5] Highly plastic lateral connections between microcircuitry delineate the columnar cortical microcircuit, smaller and distinct from functional cortical columns that are characterized by intrinsic variability in receptive fields and connections.[2]

How do we attack the problem of analyzing the functions of neural tissues and synthesize a theory of global brain function? One option is to first map microcircuits in these tissues then use maps to constrain functional simulations aimed at modeling and explaining function. Long underway,[6] mapping approaches change as new techniques are developed to attack the problem.,[7,8,9]

Typically, approaches study functional connectivity between neurons using physiological recording techniques,[10] or reconstruct and analyze tissue structure at the level of neurons and their connections by determining three-dimensional locations of tissue components and their relationships. [11] The purpose of this commentary is to consider how and the degree to which high throughput reconstruction might transform mapping microcircuits in the brain both in technical execution and in its application to elucidating global brain function.

Solution Requirements: Inputs, High Throughput Reconstruction, and Outputs

High throughput neural tissue reconstruction depends first on treating a tissue to reveal its histological structure.[12] The usability of structural data is determined first by the resolution of the light microscope. Small caliber fibers (for example, axons) found in all microcircuits typically lie near diffraction limits of resolution, such that only experimental fluorescent microscopic techniques promise to achieve sensitivity necessary to resolve sparsely stained tissues.[13] Second, data usability depends on resolution of relationships between fibers in densely stained tissue. This requirement, which we term *relationship determination*, depends on, but is not equivalent to, fiber resolution.

To illustrate relationship determination, consider two fibers that originate and terminate at separate resolvable points. The proximity of their component sections may change, making them impossible to resolve at some intermediate point. When this degradation occurs, relationships between the unresolved fiber components and their relationships to all subsequent components become uncertain. Thus relationship determination remains degraded even when component resolution recovers.

Since neural tissue is dense, the distance between unrelated, stained components frequently falls below the resolution limits of the imaging device. In fact, the number of components spuriously contacted in fully stained tissue would be much greater than the number to which a component is actually related. Even in sparsely stained tissue, uncertainty during tracing will arise and propagate along fibers, compounding as local uncertainties invade larger branches, whole neurons, and ultimately whole tissues and circuits, such that no reconstruction escapes at least some uncertainty about the relationships of each of its components to the remainder of the reconstruction. How then are acceptable reconstructions achieved?

Various techniques exploit tissue imaging dimensions such as staining density and color to achieve better relationship determination. From bright field sparse staining,[14] to genetically varied fluorescence, in which only components of the same neuron fluoresce with the same color,[15] these techniques allow relationships to persist even when resolution fails. Still, because of the difficulty in resolving critical components and the risk of propagating and compounding errors in relationship determination, expert anatomists typically perform reconstructions manually. For high throughput reconstruction, user input must be streamlined to require decisions only at points where resolution and relationship determination are poor. Ultimately, this demands experts be replaced by fiber tracing algorithms capable of drawing upon contextual cues used by the expert. These cues typically derive from two models of neural tissue.

[5] Silberberg, G., Gupta, A., & Markram, H. (2002). Stereotypy in neocortical microcircuits. *Trends in Neurosciences*, 25(5), 227–230.

[6] Douglas, R. J., Martin, K. A. C., & Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Computation*, 1(4), 480–488.

[7] Nikolenko, V., Poskanzer, K. E., & Yuste, R. (2007). Two-photon photostimulation and imaging of neural circuits. *Nature Methods*, 4, 943–950.

[8] Callaway, E. M. (2008). Transneuronal circuit tracing with neurotropic viruses. *Current Opinion in Neurobiology*, 18(6), 617–623.

[9] Micheva, K. D., Busse, B., Weiler, N. C., O'Rourke, N., & Smith, S. J. (2010). Single-synapse analysis of a diverse synapse population: proteomic imaging methods and markers. *Neuron*, 68(4), 639–653.

[10] Gupta, A., Wang, Y., & Markram, H. (2000). Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex. *Science*, 287(5451), 273–278.

[11] Binzegger, T., Douglas, R. J., & Martin, K. A. C. (2004). A quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience*, 24(39), 8441–8453.

[12] Senft S. L. (2011). A brief history of neuronal reconstruction. *Neuroinformatics*. doi: 10.1007/s12021-011-9107-0.

[13] Ji, N., Milkie, D. E., & Betzig, E. (2010). Adaptive optics via pupil segmentation for high-resolution imaging in biological tissue. *Nature Methods*, 7, 141–147.

[14] Rockland, K. S. (2004). Connectional neuroanatomy: the changing scene. *Brain Research*. 1000(1–2), 60–63.

[15] Lichtman, J. W., Livet, J., & Sanes, J. R. (2008). A technicolour approach to the connectome. *Nature Reviews Neuroscience*, 9, 417–422.

First, branched structures to which each fiber component belongs inform relationship determination. Larger structures (typically neurons) provide rich context that makes determining relationships at a single point of overlap easier, and in cases where entire accurate reconstructions from either side of the ambiguity are available, trivial. Ideally, this decision involves only matching neuron subsections based on type, orientation, and branching pattern. Typically however, full reconstructions of subsections are not available, so anatomists engage in a larger search for contextual cues then proofread problem areas.

Second, anatomists draw upon a model of local rules of neuronal growth and development. This model allows certain behaviors to extend neural fibers and create certain structures but not others. The decisions that impose these constraints on tissue emerge from a set of molecular sensing and motility components packed into the specialized tip of a growing fiber, known as the growth cone.[16] By applying this understanding of the dynamics and constraints imposed by growth cone behavior (for example, constraints on extension rate, turning radius, branching frequency, etc.) to decisions about likely fiber trajectories and component relationships, anatomists rule out unlikely reconstructions.

Faced with ambiguity, automated reconstruction algorithms may either proceed with a decision, note the need for future computer-aided proofreading, then guide the user back to the problem location, or avoid a decision altogether and instead allow connections of partial reconstructions at a later time. Guiding users back to locations of uncertainty should reduce overall user intervention, especially if models increase in quality as a reconstruction advances. Even with a sophisticated computer-aided reconstruction interface, it is unlikely that either approach could accelerate reconstructions sufficiently to reach high-throughput levels, given the need for user input.

Model creation may be necessary for fully automated systems, creating greater contextual information for local reconstruction decisions and potentially eliminating user input altogether. For example, solutions presented in the current issue extract local fiber sections by analyzing larger image contexts,[17] construct generic tree structures from local optimization techniques operating over these extracted sections,[18] and employ global optimization techniques to choose from multiple alternative neuronal reconstructions.[19] Additionally, an abstract model based approach considers neuron morphology and imaging techniques in order to disambiguate alternative reconstructions.[20] Finally, growth cone modeling techniques might complement existing tracing and fiber extension techniques such as gradient vector flow, employed in another solution to model axon direction with a deforming and stretching force.[21]

Automated tissue reconstruction proceeds from image segmentation and tracing to the generation of large volumes of fiber component coordinates and relationship information as outputs. We propose that standard outputs should also include measures of confidence. These measures would quantify confidence for each component relationship determination, as well as include alternate determinations and their associated (lower) confidence levels.

Recording these measures of confidence together with each point would have three advantages. First, it would allow subsequent quantitative analysis to explore alternative models of the tissue. Retrospective analysis represents one way in which automated reconstruction could surpass manual reconstruction in usefulness to the field (since uncertainty is not recorded in manual reconstructions). Second, estimates of confidence could allow for computational optimization in which tracing proceeds until an unacceptable level persists, at which point more costly models that supplement local context could be applied. Finally, models of context might be based on different uncertainty criteria and used to differentially inform reconstruction decisions and iterative revision of the reconstruction. For example, the determination of which of two models is most likely could be deferred as each is constructed in parallel until enough context has been uncovered.

Data Produced: Size, Time, and Applications

To estimate data requirements for a complete mapping of fiber components in the brain, assume that a component specifies on average 10 π μm^3 of tissue, derived from an average fiber section length of 10 μm, measured in rat Purkinje cells[22] and an average fiber diameter of 1 μm

[16] Hong, K., & Nishiyama, M. (2010). From guidance signals to movement: signaling molecules governing growth cone turning. *The Neuroscientist*, 16(1), 65–78.

[17] Bas, E., & Erdogmus, D. (2011). Principal curves as skeletons of tubular objects: locally characterizing the structures of axons. *Neuroinformatics*. doi: 10.1007/s12021-011-9105-2.

[18] Chothani, P., Mehta, V., & Stepanyants, A. (2011). Automated tracing of neurites from light microscopy stacks of images. *Neuroinformatics*. doi: 10.1007/s12021-011-9121-2.

[19] Türetken, E., Gonzalez, G., Blum, C., & Fua, P. (2011). Automated reconstruction of dendritic and axonal trees by global optimization with geometric priors. *Neuroinformatics*. doi: 10.1007/s12021-011-9122-1.

[20] Zhao, T., Xie, J., Amat, F., Clack, N., Ahammad, P., Peng, H., Long, F., & Myers, E. (2011). Automated reconstruction of neuronal morphology based on local geometrical and global structural models. *Neuroinformatics*. doi: 10.1007/s12021-011-9120-3.

[21] Wang, Y., Narayanaswamy, A., Tsai, C., & Roysam, B. (2011). A Broadly Applicable 3-D Neuron Tracing Method Based on Open-Curve Snake. *Neuroinformatics*. doi: 10.1007/s12021-011-9110-5.

[22] Berry, M., & Flinn, R. (1984). Vertex analysis of Purkinje cell dendritic trees in the cerebellum of the rat. *Proceedings of the Royal Society. B*, 221(1224), 321–348.

(ignoring the reconstruction of spines). This yields ~32 trillion three-dimensional component coordinates plus radii (x, y, z, and r) for a human-sized (1 liter) brain, and ~32 billion for a rodent-sized (1 milliliter) brain. In addition, recording relationships between components requires at minimum a topology identifier for each component and its parent. While some tissues will require a higher density of components,[23] these averages provide a starting point for further analyses.

Our total data estimate assumes 8 bytes for each floating point coordinate and radius, and 4 bytes for each integer topological identifier, yielding ~1 TB for 1 milliliter of tissue and ~1 PB for 1 liter. Recording confidence estimates for each component would likely double requirements for representing component relationships, depending on the complexity of the metric, but should increase the estimate by less than an order of magnitude. Therefore, data requirements for a single specimen would remain manageable, and could fit within the memory of a medium- to large-sized memory server or cluster.

The serial generation of ~10^{10}–10^{13} components and their relationships would be prohibitive by all estimates. The current processing time of ~1–3 h for data sizes on the order of 10^4 components indicates a serial processing time of approximately one component per second. This translates to processing times of ~10–10,000 centuries for the tissue volumes considered. Clearly a parallel reconstruction algorithm will be required. Fortunately, parallelization of neural tissue reconstruction has already begun for certain image preprocessing steps.[24] Massive parallelization of full algorithms could speed up calculations by a factor of ~10^4, resulting in compute times of ~1 month to 1 century for the volumes considered, depending on the data decomposition and the parallelization approach used, and assuming the entire algorithm can be optimally accelerated on today's largest machines (i.e., petaflop). Therefore, while rodent-sized brains could likely be reconstructed using today's supercomputers, human-sized brains would require exascale (i.e., exaflop) supercomputers (expected this decade), which would deliver a speed up factor of ~10^7.

Data collection times would also be prohibitive if performed serially. Assuming one second of image acquisition (depending on imaging modality) per micron optical section through a 100 μm field of view (i.e., 10,000 μm^2, imaged under a 60×, NA 1.4 oil immersion objective, optically zoomed to sample at the Nyquist limit) imaging times could range from ~3 years to 30 centuries, thus requiring parallelization of image data collection. Unlike parallelization of reconstruction algorithms, which requires communication between parallel tasks, the task of parallelizing data collection requires only the resources to replicate the imaging apparatus approximately 1,000 times and to receive and image the tissue slices from a single specimen. In addition, specializations of imaging devices and collection methods for high-throughput solutions may reduce the scale of parallelization required.[25]

Obviously, the challenges and costs of high-throughput reconstruction are great. Meeting them is worthwhile only if the data can be used for valuable scientific and applied pursuits. Here we describe several potential uses for the data and comment on their value. First, physiological recordings from connected neurons have provided a rich source of information on how microcircuit components function and propagate signals,[10] especially when accompanied by anatomical reconstructions. Analyses show vertebrate circuits correlate function and structure and are at times clearly stereotyped.[26] If large-scale reconstructions of brain tissues become available, functional recordings correlated to key structural observation could serve as annotations to the reconstructions, resulting in an opportunity to accelerate microcircuit tracing through higher-order structure-function correlation.

For example, higher-order structure-function correlation informed the role of Martinotti cell inhibition in the neocortical microcircuit. An initial study correlated functional synaptic connections from layer 5 pyramidal neurons onto layer 5 Martinotti cells with stereotyped spatial patterns of neurons and synapses,[26] and a higher-order correlation suggested the existence of disynaptic loops joining the same neurons. Specifically, structural knowledge of the Martinotti cell's axonal ramifications in layer 1 overlapping with the layer 5 pyramidal cell's apical tuft suggested a region for synapses to complete a loop, which was subsequently confirmed through paired physiological recordings.[27] Similarly, physiological recording guided by suspicious higher-order correlations in a larger structural database and aimed at testing functional connectivity could make microcircuit analysis proceed more efficiently.

[23] Mishchenko, Y., Hu, T., Spacek, J., Mendenhall, J., Harris, K. M., & Chklovskii, D. B. (2010). Ultrastructural analysis of hippocampal neuropil from the connectomics perspective. *Neuron*, 67(6), 1009–1020.

[24] Narayanaswamy, A., Wang, Y., & Roysam, B. (2011). 3-D Image Pre-processing Algorithms for Improved Automated Tracing of Neuronal Arbors. *Neuroinformatics*. doi: 10.1007/s12021-011-9116-z.

[25] Dodt, H.-U., Leischner, U., Schierloh, A., Jährling, N., Mauch, C. P., Deininger, K., Deussing, J. M., Eder, M., Zieglgänsberger, W., & Becker, K. (2007). Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain. *Nature Methods*, 4, 331–336.

[26] Kozloski, J., Hamzei-Sichani, F., & Yuste, R. (2001). Stereotyped position of local synaptic targets in neocortex. *Science*, 293(5531), 868–872

[27] Silberberg, G., & Markram, H. (2007). Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells. *Neuron*, 53(5), 735–746.

Global brain morphologies and connectivity between homologous neural tissues are considered topological equivalents across different vertebrate species.[28] Animal models therefore aid understanding of both human brain structure and fundamental brain processes. To what extent homologous neural tissues themselves and their microcircuitry are topological equivalents is a more difficult question, and requires comparing microcircuits of different vertebrate species quantitatively. Since meaningful comparisons require both a large quantity and consistent quality and format of digitized structural data, automated high-throughput neural tissue reconstruction is needed. Comparative analysis could help identify fundamental circuit components and functions of tissues, and ultimately provide deeper understanding of global brain function and its emergence from a conserved vertebrate brain plan and microcircuitry.

Despite its importance, a definitive set of synaptic connections will not emerge from the techniques described, since structures indicating the existence of a synapse are too small to resolve using any light or fluorescence microscope. Therefore, the goal of tracing microcircuitry in neural tissue with these methods must proceed by statistical means, and specifically by identifying where stereotyped synapses are likely. A *connectome* at the level of microcircuitry that identifies all synaptic connections in a tissue[29] is less likely to emerge than a *juxta-connectome*, which identifies all potential synapses in a tissue based on the apposition of neural fibers in the structural model.

Arguing against this perspective, a recent study examined potential and actual synapses from electron microscopic reconstructions of small regions of tissue from an individual animal,[23] and its results call into question Peter's rule,[30] which states that the number of synapses along a fiber should be proportional to overlap between axonal and dendritic arbors. Variants of the rule explored in the study did however maintain some predictive power. We anticipate that analyses of juxta-connectomes constructed from larger tissues in multiple individuals could yield additional variants on Peter's rule that exploit other statistical regularities to predict where synapses are most likely in the tissue. These analyses, for example, could look for statistically significant correlations of neuron type appositions across different individuals, a prerequisite of stereotyped microcircuitry.

Any description of statistical regularities among fiber identities and appositions in neural tissue will also provide a basis for inferring proximal developmental trajectories of structures within the tissue. Changing statistical relationships among fibers in the juxta-connectome derive from the actions of growth cones, which sensed and responded to the surrounding phenotype of fibers and neurons (and the genes they expressed) during development. These growth cone artifacts permitted further developmental and experience-dependent changes, such as spine extension and retraction[31] and synapse elimination,[32] to ultimately determine connectivity.

To infer developmental trajectories from structure, a model must first approximate the role that fibers and neurons play in secreting molecules and generating field potentials and concentration gradients within tissue. These fields and gradients deform the trajectory of the growth cone in predictable and stereotyped ways.[16] Modeling the interplay between cellular and fiber identity and the effective forces acting upon growth cones to create neuronal and circuit morphology[33] could exploit constraints from complete structural data collected from tissue at various stages of development. Neuron growth simulation could then help uncover the mapping by neuronal growth and development from a compact set of genetic markers to stereotyped microcircuits.

Structural data collected from high-throughput tissue reconstruction may ultimately constrain functional simulations of tissue. We term this coupling between structural models and physiological simulation at the level of branched fibers (using the equations of Hodgkin and Huxley and the numerical methods of compartmental modeling) *neural tissue simulation*. Structure and function simulated in a three-dimensional coordinate system corresponding to real brain tissue[34] together with functional synapse model placement derived from structural data analysis[35] are prerequisites of neural tissue simulation. Structural constraints derived from whole tissue reconstructions and a juxta-connectome would allow parameterization of simulations to better fit functional observations, since the

[28] Nieuwenhuys, R. (1998) Comparative neuroanatomy: place, principles and programme. In *The Central Nervous System of Vertebrates*. Nieuwenhuys, R., Donkelaar H. J., & Nicholson C. (Eds.) Berlin: Springer Verlag. 273–326.

[29] Eisenstein, M. (2009). Neural circuits: Putting neurons on the map. *Nature*, 461, 1149–1152.

[30] Peters, A., & Feldman, M. L. (1976). The projection of the lateral geniculate nucleus to area 17 of the rat cerebral cortex. I. General description. *Journal of Neurocytology*, 5(1), 63–84.

[31] Holtmaat, A., & Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, 10, 647–658.

[32] Lichtman, J. W., & Colman, H. (2000). Synapse elimination and indelible memory. *Neuron*, 25(2), 269–278.

[33] Koene, R. A., Tijms, B., van Hees, P., Postma, F., de Ridder, A., Ramakers, G. J., van Pelt, J., & van Ooyen, A. (2009). NETMORPH: a framework for the stochastic generation of large scale neuronal networks with realistic neuron morphologies. *Neuroinformatics*. 7(3), 195–210.

[34] Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience*, 7, 153–160.

[35] Kozloski, J., Sfyrakis, K., Hill, S., Schürmann, F., Peck, C., & Markram, H. (2008). Identifying, tabulating, and analyzing contacts between branched neuron morphologies. *IBM Journal of Research and Development*, 52(1.2), 43–55.

effects of neuron structure and synapse placement on physiological models are well known and significant.[36] [37]

On DIADEM and Next Steps: Problems and Ways Forward

The DIADEM Challenge recently culminated in a final competition and workshop at the Howard Hughes Medical Institute's Janelia Farm Research Campus. This challenge aimed to identify new approaches to the problem of automated neural tissue reconstruction by inviting teams of researchers in disciplines ranging from neuroscience to computer science to compete for cash prizes developing automated reconstruction algorithms. The approaches and outcomes are reviewed elsewhere in this issue.[17,18,19,20,21,24] Here we summarize what was learned, and briefly outline a path to high-throughput reconstruction and large-scale neural tissue simulation.

In retrospect, the challenge of DIADEM was outliers. As noted in the final evaluation of solutions, clear advances in dealing with the bulk of problems in automated reconstruction created "shock and disbelief" among the organizers. Errors due to rare conditions were not unexpected, given the range of difficult tasks present in resolving structure from raw microscopic images and disambiguating tissue components in dense, cluttered fields of stained fibers. Because errors compound in neural tissue reconstruction, and because finding them using the solutions described in this issue was not always streamlined, accuracy of the reconstructions suffered. In addition, human intervention was not only required, but at times much of the speed up derived from using the algorithm was lost.

The challenge now for competing teams is analyzing, categorizing, and ultimately deploying solutions to errors that escaped their preliminary reconstruction algorithms and rendered each too costly in terms of the need for human intervention and proofreading. Outlier error categories are likely separate problems (for example, axon-axon crossover is likely a separate problem from axon-dendrite crossover, etc.), and each must therefore be addressed within its own context, and with its own solution.

The ability to detect each problem category's context and automatically deploy a tailored solution is therefore yet another area for future research. Certain problems will require improved local image processing, while others a more costly iterative approach that includes creation of models of the larger context of neurons and the tissue in which they are embedded. Ultimately, some problem categories will not be solvable, given that even trained anatomists cannot disambiguate definitively all structures. For solutions that do not achieve some reasonable criterion (such as DIADEM's 20× speed up, which none did) there still exists the possibility that each might possess a subset of solutions to the host of problems presented by automated tissue reconstruction. Drawing on all solutions to create a single application that deploys solutions based on recognition of a problem context is another possible course. To facilitate this, categorizing failure conditions, and evaluating solutions separately for each condition would be a valuable collaborative undertaking.

In developing post-DIADEM solutions to the problem of high-throughput tissue reconstruction, another consideration is *scaling up* vs. *scaling out*. Scaling up is required if more sophisticated, and therefore more computationally costly, algorithms are developed that attempt to address local problems in the reconstruction with serial computation. Run time remains constant if the local computing power scales up to match the increased computing cost of the solution. Scaling up also would allow existing algorithms to run faster. Because the main cause of slowdown in DIADEM solutions was errors and the need for human intervention, scaling up would not likely have permitted applications to achieve DIADEM's criterion 20× speed up at this stage. The computational cost of resolving these errors automatically will likely demand scaling up for future solutions.

Alternatively, scaling out would execute a local reconstruction algorithm on a single node of a parallel system then increase the throughput of the application by deploying more nodes running the same algorithm on similar image volumes. As the amount of tissue reconstructed in parallel grows, the time to reconstruct it remains constant, provided the parallel algorithm balances computation with communication on and between separate computational nodes. Because of inter-process communication, scaling out is not a reasonable approach to managing increasing complexity in the local reconstruction itself, except for those approaches that exploit stochastic search algorithms, where the sampling step may be distributed across a parallel architecture. Scaling out would not have helped a solution to meet DIADEM's 20× speed up criterion, since errors predominated and image volumes were small. High-throughput reconstructions on the scale of the tissue volumes considered will certainly require scaling out to ensure image data can be processed in reasonable compute times.

Data decomposition, or the placement of data on a parallel architecture, is a central problem in parallel computing, and as high-throughput tissue reconstruction scales out to exploit

[36] Krichmar, J. L., Nasuto, S. J., Scorcioni, R., Washington, S. D., & Ascoli, G. A. (2002). Effects of dendritic morphology on CA3 pyramidal cell electrophysiology: a simulation study. *Brain Research*, 941(1–2), 11–28.

[37] Ascoli, G. A., Atkeson, J. C. (2005). Incorporating anatomically realistic cellular-level connectivity in neural network models of the rat hippocampus. *Biosystems*. 79:1–3, 173–181.

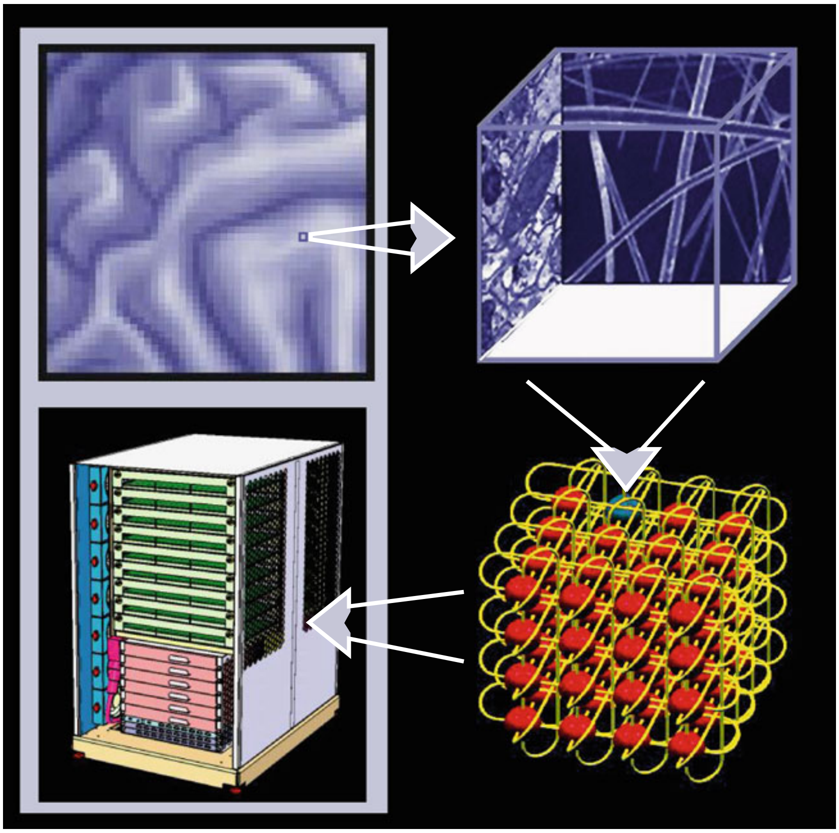

Fig. 1 Data decomposition of neural tissue simulation for ultra-scalability. We exploited the volume-filling nature of neural tissue (*upper left*) to formulate a simulation in which all communication is local. Here a volume of simulated tissue (*upper right*) is represented as consisting of fibers (*back face*) and fiber cross sections (*left face*). Because fiber density, length, and diameter are usually well known neuroanatomical parameters, they can be used to reasonably estimate all computation and communication costs for a given volume size, across simulations of any scale. Note that whole neurons are not necessarily contained in any volume, and long range connectivity of any fiber (beyond nearest neighbor volumes) is not represented within any volume. This decomposition is ideal for the IBM Blue Gene series architecture (*lower left*), which is characterized (*lower right*) by a torus network of connections joining all nodes of the machine to which a tissue volume is assigned

supercomputing-scale machines (thousands of nodes), the appropriate decomposition for image data will need to be determined. More important than balancing data across nodes of a parallel machine, an appropriate decomposition must ensure that computational loads are balanced, and that communication between nodes is minimized. These requirements often lead to algorithms that do not employ obvious data decompositions or the most popular or intuitive data abstraction.

In the domain of neural tissue simulation, we have explored an alternative to the standard neuron decomposition of simulation data,[38] and have instead placed data and the calculation of compartments, channels, and synapses from within bounded tissue volumes onto each of the nodes of a Blue Gene/P supercomputer, creating the ultra-scalable Neural Tissue Simulator (Fig. 1). An in depth treatment of the Neural Tissue Simulator will be taken up in a subsequent publication. Briefly, because Blue Gene/P's nodes are connected in a torus network topology, our volume decomposition's nearest-neighbor communication is highly efficient, resulting in constant simulation rates (1 processor-second per simulated-neuron-millisecond) as a machine was loaded with a constant average 250 neurons per node across machine sizes ranging from 64 to 4,096 nodes.

Because neurons have diffuse, long-range, and largely unknown connection patterns in the brain, neuron decomposition makes balancing load and predicting communication patterns in a large neural tissue simulation difficult. In our novel design, however, we can balance the computational load across nodes without prior knowledge of connectivity or neuron morphology by weighting each simulation coordinate with the local computational costs associated with it (ie., compartment, channel, and synapse calculation costs) (Fig. 2), then slicing the tissue into balanced volumes. Communication is easily predictable due to the known density of fibers cut at each volume interface.

[38] Migliore, M., Cannia, C., Lytton, W. W., Markram, H., & Hines, M. L. (2006). Parallel network simulations with NEURON. *Joural of Computational Neuroscience*, 21(2), 119–129.

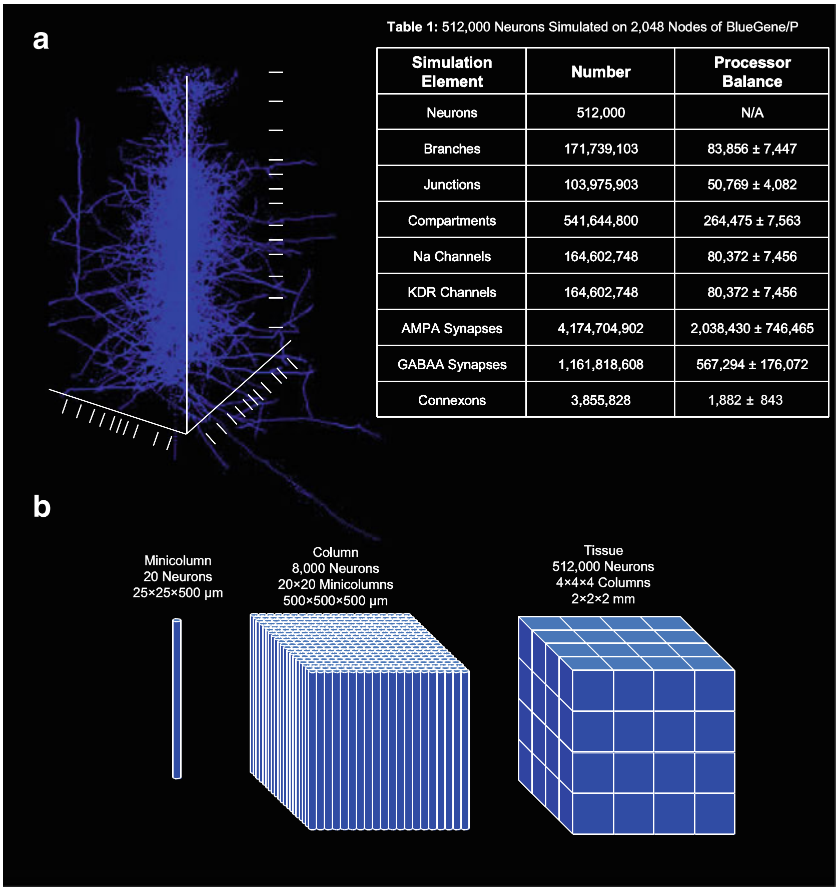

Table 1: 512,000 Neurons Simulated on 2,048 Nodes of BlueGene/P

Simulation Element	Number	Processor Balance
Neurons	512,000	N/A
Branches	171,739,103	83,856 ± 7,447
Junctions	103,975,903	50,769 ± 4,082
Compartments	541,644,800	264,475 ± 7,563
Na Channels	164,602,748	80,372 ± 7,456
KDR Channels	164,602,748	80,372 ± 7,456
AMPA Synapses	4,174,704,902	2,038,430 ± 746,465
GABAA Synapses	1,161,818,608	567,294 ± 176,072
Connexons	3,855,828	1,882 ± 843

Fig. 2 Processor balancing and test simulation of large neural tissue using cortical neurons. **a**. Using an adaptive slicing technique (illustrated) which generates volumes of nearly equal computational load (see text), we created a simulation of 512,000 neurons. Table 1, inset: Simulation components and their effective distribution across the machine's nodes ± standard deviation, following the use of this technique on the entire tissue. (Simulations already performed of larger scales will be illustrated in a subsequent publication.) **b**. The simulation was based on neuron morphologies taken from the Markram lab (downloaded from Neuromorpho.org), each assigned to appropriate layers in a simulated minicolumn, and rotated randomly around the cortical axis for each neuron in each minicolum. Mincolumns were composed into columns, and the tissue then scaled outward in all three dimensions (i.e., violating gross cortical morphology). We added columns in this way to preserve a cubic structure of the tissue and generate scaling data appropriate for large tissue simulations that grow similarly, and encompass more than a single brain structure

Likewise, in the case of tissue reconstruction, data decomposition may be based on image volumes, where each processor operates over data from a set of contiguous sections within a field of view (100×100 μm). For the range of tissue volumes considered, we estimate 100 million to 100 billion of these image sections would be required. At ~1 MB per image, between 100 TB and 100 PB of machine memory or storage would be required. While the former is feasible for today's supercomputers (comprising ~10^5 nodes, gigabytes of memory per node), for the latter, storing images in main memory would require larger machines (millions of nodes) or larger per node memory sizes (~1 TB) than exist today.

Algorithms operating over image volumes could balance load according to data provided computational cost is proportional, and would incur communication costs whenever local reconstructions required contextual information about a neighboring node's reconstruction. Variable slicing of image volumes could mitigate imbalance when reconstruction costs are irregular (for example when fiber density varies). Depending on the degree to which algorithms use larger contextual models to improve reconstruction performance, communication may occur only once following local reconstruction (for cases using local context only), or repeatedly, as iterative model-building proceeds. Though not impossible, it is difficult to imagine a solution for which the most obvious and intuitive image volume decomposition will not be optimal.

Scaling out solutions to machines of tens of thousands of nodes in order to achieve high-throughput tissue reconstruction would require on the order of ~100 PB of storage or machine memory for the largest tissue considered. While machine sizes will likely grow orders of magnitude larger than today's, per node memory requirements for a solution within main memory will likely continue to exceed what is

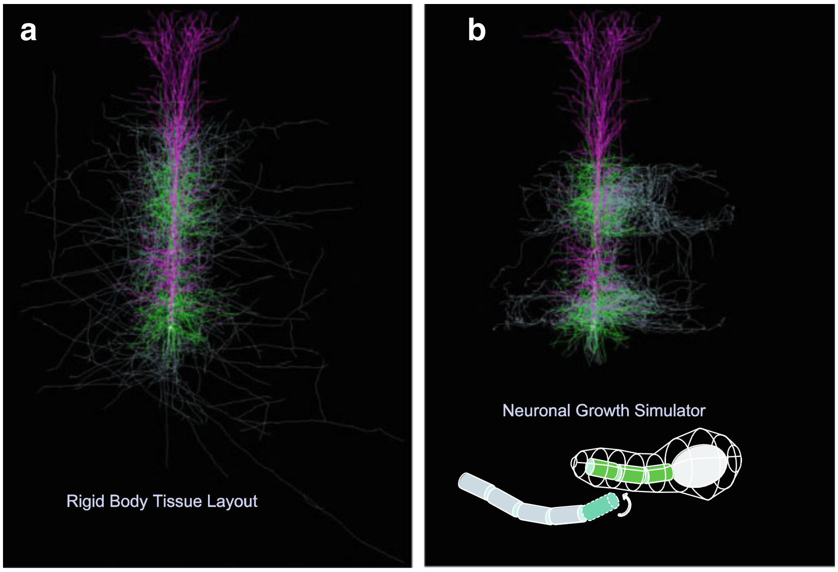

Fig. 3 Neuron growth simulation demonstrates parameterizable deformation of axon growth trajectories. **a**. A single minicolumn from the simulation described in Fig. 2 is represented, with axons in white. **b**. Simulating neuron growth in a single column of 8,000 neurons on 1,024 nodes of Blue Gene/P, neurons grew for 24 h such that each axon tip encountered strong attractive forces from cell bodies and basal dendrites of neighboring neurons (bottom, represented by an attraction shell and a rotating tip). This resulted in denser, more stratified axonal ramifications that contacted dendrites more frequently (*top*)

possible. Tissue image data could easily remain on disk, but I/O bandwidths (at ~100 MB/s for today's disks) and the size of the data (filling ~1 billion 100 MB disks) would limit compute times for solutions requiring frequent disk accesses in largely random patterns as they traverse a local image volume repeatedly.

An intermediate solution that could serve the needs of high-throughput tissue reconstruction for fast computation and data access at a reasonable cost/performance ratio is *Blue Gene Active Storage*.[39] Here, a large amount of storage class memory (such as phase-change or flash memory) is placed very close to each machine node, allowing high capacity (on the order of 100 GB-1 TB per node in the coming years) of local storage together with high I/O bandwidths (on the order of 1 GB/s per node) at a more reasonable cost. It will be interesting to see how new architectures such as Active Storage may contribute to satisfying the requirements of high-throughput tissue reconstruction.

In considering drivers for future research in high-throughput tissue reconstruction, a specific target application for the data generated could address several interrelated needs. Beyond the obvious need to maintain a compelling vision of a potential application for the huge amounts of data generated, three other practical needs emerge. First, the problem of checking and proofreading data will become unmanageable after data begins to exceed what can be reasonably proofread by an expert anatomist.[23] Second, the need for alternative metrics of quality beyond comparisons to manually reconstructed data will be required when no such manual reconstructions exist. Finally, automatically identifying problem areas in a reconstruction will benefit if calculations use data from local regions of the reconstruction for a purpose that can be validated separately from the structural model itself. With these needs in mind, we propose large-scale neural tissue simulation as a reasonable target application for high-throughput reconstruction, and argue that tighter coupling between research activities in both areas could be advantageous.

Neural tissue simulation is a compelling application of high-throughput reconstruction data foremost because it is *data driven*. Data driven refers to its incorporating all available data from real tissue before additional user-defined parameters are introduced. Because these data derive first from structural constraints, which identify the location of tissue elements (such as ion channel types, kinetics, and densities) and second functional constraints, which give rise to physiological dynamics, a direct pipeline from high-throughput reconstructions to these models could immediately be exploited by the field.

Parameterizing simulations after these constraints are imposed then aims to fit simulation outputs to functional observations from the tissue. Our own work with the Neural Tissue Simulator and its tissue volume decomposition produced simulations of a neocortex-derived tissue of up to 1,024,000 neurons, comprising ~1,000 compartments

[39] Fitch, B., Rayshubskiy, A., Ward, T. J. C., Pitman, M., Metzler, B., Schick, H. J., Krill, B., Morjan, P., & Germain, R. S. (2010). Blue Gene Active Storage. Presentation to The National Science and Technology Council's High-End Computing File System and I/O Workshop, Arlington, VA, http://tinyurl.com/BGAS-HECFSIO2010.

each, with a total of ~10 billion conductance-based synapses (AMPA and GABAA) and gap junctions, using neuron morphologies from the Markram lab[40] downloaded from http://NeuroMorpho.org[41] (Fig. 2). Parameterization of simulations like this remains a challenge for us and the field in general. For example, specific structural relationships between neurons are artificially generated when neuron reconstructions from different tissues are embedded into a single tissue model.[35] Because of a lack of whole tissue reconstructions, these artificial relationships may create errors in the placement and identity of synapses in the microcircuit. High-throughput tissue reconstruction could provide the data needed to constrain neural tissue simulations more accurately.

Neural tissue simulations ultimately generate data that model the functional outputs of the tissue, and therefore any validation of the simulation with data collected from real tissue will likely be a validation of tissue physiology and function. Identifying errors in the reconstruction through this validation procedure may be possible, and would provide one means for feeding information from large-scale simulations back to the algorithm responsible for reconstructing the tissue. In this scenario, a failure to replicate microcircuit behavior in some local region of the tissue may highlight the need for revision of the structural model by an expert anatomist or further iteration of a reconstruction algorithm over this location. Such coupling would then provide a means for functional simulations to shape the structural model.

Alternatively, simulations could generate structural data based on constraints imposed by data collected from whole tissue reconstructions. In this scenario, whole tissue reconstructions may be error prone, but sufficient in identifying regularities in tissue at each stages of development to provide parameters for simulation of neuron growth. We have implemented within the Neural Tissue Simulator a neuron growth algorithm that abstracts growth cone interactions with the tissue milieu as a set of short and long range forces applied to growing fiber ends by the surrounding simulation elements (Fig. 3). Tissues could result from such simulations that are consistent with all data collected from reconstructions, though identical to none. Such simulated tissues may also be less subject to outlier errors and local noise in any one reconstruction, given that growth constraints could be imposed uniformly and without exceptions across the tissues during simulated development. The validity of a simulated structure could then be tested by functional simulations constrained by the simulated structure.

In closing, the goals of high-throughput reconstruction and large-scale neural tissue simulation should now be recognized as overlapping and synergistic. Both aim to create models of neural tissue that can be used for the purpose of prediction. Predictive structural models derived from high-throughput reconstructions would identify stereotyped organizing principles and relationships in neural tissues and microcircuits that allow the generalization of structure across individuals, at different stages of development, and across different vertebrate species, ultimately including human. Neural tissue simulation aims to aggregate both these structural and functional data from real neural tissues then use them to constrain physiological and developmental models of dynamics in the tissue. Ultimately, predictive models of structure and function could inform each other, and coupling the validation of structural models generated from high-throughput reconstructions and functional models generated from neural tissue simulation would advance each, and our understanding of brain function, more rapidly than either effort could in isolation.

Information Sharing Statement

All data presented in this paper were derived from the publicly accessible database found at www.neuromorpho.org. The Neural Tissue Simulator software described is experimental. IBM would like to create an active user community. Please contact the author if you are interested in using the tool.

Acknowledgments We acknowledge the crucial contributions of John Wagner, IBM Research, in development of the Neural Tissue Simulator and the numerical methods it employs and Mike Pitman, IBM Research, in development of the neuron growth algorithm and the numerical methods it employs. We also thank Charles Peck and Blake Fitch, IBM Research, for critical input on the manuscript.

[40] Wang, Y., Gupta, A., Toledo-Rodriguez, M., Wu, C. Z., & Markram, H. (2002). Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. *Cerebral Cortex*, 12(4), 395–410.

[41] Ascoli, G. A., Donohue, D. E., Halavi, M. (2007) NeuroMorpho. Org: a central resource for neuronal morphologies. *Journal of Neuroscience*., 27(35), 9247–9251.