NEUROSCIENCE

Brain research challenges supercomputing

Big data obtained from unraveling human brain structure raise processing demands

By **Katrin Amunts** *and* **Thomas Lippert**

The adult human brain contains ~86 billion neurons (*1*). Zooming into its cellular and subcellular details to reveal different aspects of neuronal connectivity is a key area of research. However, to link the different spatial scales from the synaptic level (at nanometer range) through single neurons and glial cells (at the micrometer level) to the whole organ is most challenging. Recently, the connectome of *Caenorhabditis elegans*, with its 302 neurons, has been characterized, and a complete structural-functional model has been proposed (*2*). A comparable level of detail of the human brain connectome is still a long way off. As such, decoding the human connectome, the mechanisms of signal transduction, and relationships to brain function are linked to exponentially growing challenges in advanced computational and storage technologies, which in turn may lead to creative solutions beyond neuroscience.

Challenges of scanning time, storage technology, and data processing do not start with studies of the human brain but are important considerations when studying brains of vertebrates and even invertebrates (see the figure). For example, the reconstruction of the synapse connectivity of an adult fruit fly brain, with ~100,000 neurons, culminated in ~21 million camera images and a dataset of 106 terabytes (*3*). Recently, a fragment of the human cerebral cortex, a volume of ~1 mm³, was reconstructed in three dimensions, corresponding to a data volume of ~1.4 petabytes (*4*). Although 1 mm³ is a tiny portion of the cerebral cortex and represents ~0.00007% of the total brain volume, a high-speed multibeam electron microscope was required to collect these data with a scan time of 326 days. Such research is conducted with the motivation that insights into the detailed architecture of the brain may reveal a new understanding of cortical networks and provide new quantitative data that describe tissue properties, with implications for brain activity.

This type of analysis of human brain tissue is an important complement to stud-

Forschungszentrum Julich, Institut fur Neurowissenschaften und Medizin, Juelich, Germany. Email: k.amunts@fz-juelich.de; th.lippert@fz-juelich.de

ies of brains from other species because there are considerable differences between them, particularly connectivity. For example, the volume of white matter of the cerebral cortex, which contains axons that support long-range connectivity, increases faster during mammalian evolution than gray matter (which contains neuronal cell bodies) (*5*). The ratio of long to short connections has consequences for electrophysiological properties and, thus, for brain function (*6*). However, studying the full extent of axons and their synapses, which may be several centimeters from their cell bodies, represents a much more substantial problem in human brains than in those of rodents or invertebrates.

Magnetic resonance imaging (MRI) (e.g., diffusion MRI) is capable of addressing long-range connections and revealing connectivity patterns, including functional connectivity in the living human brain, but image voxels are in the millimeter range. This is far above the spatial resolution necessary to detect single axons and therefore is prone to lead to ambiguities inherent in tract reconstruction (*7*). Moreover, brain connectivity includes phenomena such as collateralization (one axon may target multiple areas), convergence, reciprocity, or spatially separate dendritic arbors, which are considered key in understanding connectivity (*8*) but are not accessible with MRI.

Consequently, there is a discrepancy between macro- and microconnectome scales (*9*). Different approaches have been proposed to bridge the two worlds: For example, a combination of diffusion MRI and high-resolution optical imaging of fluorescently labeled neurofilaments in brain tissue cleared using the CLARITY technique has recently been explored in the same tissue block to allow more robust multimodal MRI-CLARITY comparisons (*10*). In another study, a tissue block from the human hippocampus was imaged with anatomical and diffusion MRI, three-dimensional polarized light imaging [3D-PLI; (*11*)], and two-photon fluorescence microscopy to bring together information of the fiber architecture of the hippocampus at different spatial scales and to represent it in a common reference space, the BigBrain model (www.ebrains.eu). These data show the perforant pathway, which does not represent a uniform tract but rather a multicomponent system that originates in

the entorhinal cortex with many thin axon fiber bundles (in the range of 20 μm) and projects to the cornu ammonis and subiculum (*12*). 3D-PLI is a microscopic technique that allows decoding of the fine architecture of the perforant pathway, which has a central role in learning and memory, as well as in the pathogenesis of Alzheimer's disease. Methodologically, the combination of different techniques helps to cross-validate findings. Practically, it paves the way to combine high-resolution region-of-interest analysis with whole-brain studies.

Atlas methods can, in a systematic manner, serve as a bridge of macro- and microscales by integrating data from ultra-high-resolution studies of single cells and their connections to small regions of interest into a common microscopical reference space. BigBrain represents such an anatomical model at 20-μm isotropic resolution, which is slightly higher than the resolution required to see details of cell morphology. It is based on 7404 histological, cell body–stained sections; the original dataset is ~1 terabyte (*13*). BigBrain provides a template to integrate findings of cortical layers and even sublayers with data on whole-brain structure, connectivity, and function. Data integration in an atlas combines the advantages of a bottom-up approach that starts with cellular characteristics and connections to build up to the macroscopic scale with the advantages of a top-down perspective that starts with behavioral function and underlying large-scale networks to better understand the properties of components, that is, cells and their circuits. Both approaches are required to provide mutual information and to constrain predictions.

It is a gigantic challenge to process and analyze entire human brains at cellular resolution. At present, it is not within reach to go down to individual axons at the whole–human brain level. However, what would it mean computationally to track axons at the whole-brain level, assuming for simplicity that an axon connects two neurons? To compute the course of fibers at the whole-brain level in the millimeter range based on diffusion MRI of postmortem brains would mean to optimize 4×10^5 spin directions. This will require ~130 megabytes of storage and 1 day of computation time on a desktop computer. The analysis of axons at the level of 60-μm isotropic resolution for a whole human brain

using 3D-PLI (*11*) would require 8.3 terabytes of storage and several ~~years~~ on existing supercomputers to optimize 2×10^{10} spins. Such large datasets also create substantial challenges for the visualization of data. Paraview, for example, based on the open-source software VTK (www.paraview.org), can use parallel graphics processing units (GPUs) and has been applied to render and visualize 3D-PLI data. Going further and optimizing fiber orientations obtained with 3D-PLI at 1.3-μm in-plane resolution (i.e., single axons) with 10^{13} spins would result in a storage demand of 3.2 petabytes and ~~decades~~ of computation. This is not possible with current petascale technology but can be accomplished with future exascale computing power, that is, computers capable of executing 10^{18} floating point operations per second (i.e., 1 exaflop). Handling such large datasets, however, creates massive computational demands at the level of input-output. More efficient input-output procedures and algorithms are emerging, which should help, but the computational challenges are still extraordinarily high.

The computational requirements also increase further when cellular and axonal information is linked to other dimensions of brain organization, for example, the molecular architecture, which is regionally specific and key for signal transduction. Similarly, requirements further increase when adding temporal changes to simulate cell activity, plasticity, or network function, to name a few examples. When developing more biologically realistic brain models, it is relevant to consider not only the regional segregation of the brain but also the laminar and sublaminar features with their specific cell numbers, anatomical and physiological properties, and distributions (*14*). Likewise, computer simulations of the human brain that are based on more realistic models that account for ana-tomical and physiological constraints must be scaled accordingly.
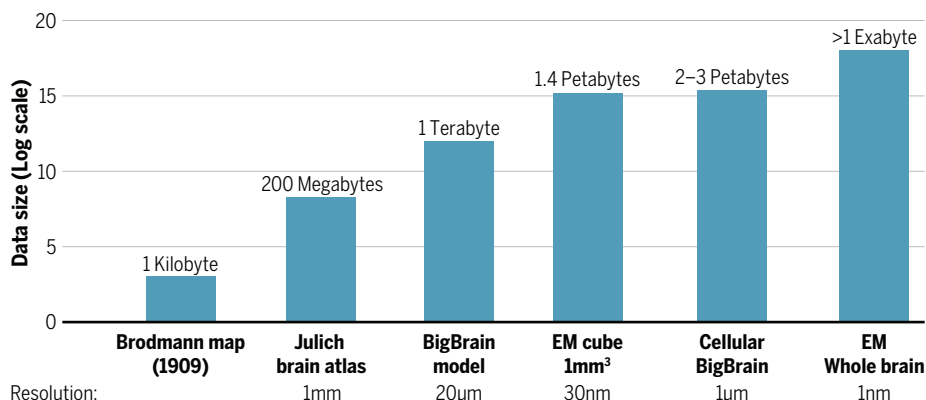
These endeavors set requirements for high-performance computing technology. For example, neuroscientific use cases become dependent on scalable workflows, starting from the extraction of multimodal datasets stored in data repositories to their preprocessing, simulation, visualization, and analysis, using machine and deep learning even more. The computation of the components of such workflows will benefit from modular and interactive concepts of future supercomputing, such as those developed in the European Deep Projects. In a modular supercomputing architecture, specialized computing modules are integrated to form a deeply connected hardware architecture and can be used coherently with each module that is best suited for a particular component of the workflow (*15*). This includes modules as diverse as central processing unit (CPU) clusters, GPU boosters, field-programmable gate array modules for data analytics, neuro-morphic systems, extreme storage modules, and, in the future, quantum computers and annealers. In addition, there is a growing need for neuroscience workflows that require interactive supercomputing—for example, to visualize intermediate results and correct parameter settings before proceeding—and that rely on the interactive use of software such as Paraview.

The big data challenge in neuroscience demands technical solutions for storing data (in the terabyte to petabyte range) that are simultaneously available to the broader community through the cloud. This requires transparent data transfer from object-oriented cloud storage systems to exabyte-scale parallel file systems of the supercomputer. The Allen Brain Institute provides a platform for neuroscientists and makes available a large number of image stacks reflecting multiple aspects of brain organization from different species— including *Drosophila*, zebrafish, mouse, and human—as well as tools for visualization, uploading, and downloading (BigNeuron; www.alleninstitute.org/bigneuron/data/). The Human BioMolecular Atlas Program aims to develop an open and global platform to map healthy cells in the whole human body, an endeavor that is accompanied by considerable big data and computational challenges. In Europe, the Human Brain Project has developed EBRAINS to provide a large spectrum of research instruments, data, and related services—with a focus on human, rat, and mouse brains—that are linked to each other and is enabling cloud access to interactive supercomputing, web-based visualization and analysis, and high-end simulation and data services through Fenix. This platform was jointly built by neuroscientists and developers based on research needs such as running large-scale simulations or 3D-reconstructions and analyzing brain models. Fenix was designed as a general infrastructure-as-a-service platform, including Europe-wide authentication services to enable access for many research communities. Sharing data, methods, and research instruments in the neuroscience community and federating services across different communities can leverage the immense associated methodological and resource investments, which in turn is fundamental to develop new diagnostic tools and therapies in brain medicine. Undoubtedly, cloud-based supercomputing and distributed, collaborative research platforms will play an increasing role in neuroscientific research to better understand brain complexity. ■

Big data from human brain research

The size of data is exponentially growing according to the resolution of data on the human brain connectome and organization. The largest datasets are estimated sizes because such data do not exist for the human brain.

GRAPHIC: K. FRANLIN / *SCIENCE*

REFERENCES AND NOTES

1. S. Herculano-Houzel, *Proc. Natl. Acad. Sci. U.S.A.* **109** (Suppl 1), 10661 (2012).
2. C. A. Brittin, S. J. Cook, D. H. Hall, S. W. Emmons, N. Cohen, *Nature* **591**, 105 (2021).
3. Z. Zheng *et al.*, *Cell* **174**, 730 (2018).
4. A. Shapson-Coe *et al.*, *bioRxiv* 10.1101/2021.05.29.446289 (2021).
5. B. Mota *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15253 (2019).
6. A. Stepanyants, L. M. Martinez, A. S. Ferecskó, Z. F. Kisvárday, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3555 (2009).
7. K. H. Maier-Hein *et al.*, *Nat. Commun.* **8**, 1349 (2017).
8. K. S. Rockland, *Brain Struct. Funct.* **225**, 1327 (2020).
9. D. C. Van Essen *et al.*, *Neuroimage* **80**, 62 (2013).
10. C. Leuze *et al.*, *Neuroimage* **228**, 117692 (2021).
11. M. Axer *et al.*, *Front. Neuroinform.* **5**, 1 (2011).
12. M. M. Zeineh *et al.*, *Cereb. Cortex* **27**, 1779 (2017).
13. K. Amunts *et al.*, *Science* **340**, 1472 (2013).
14. G. T. Einevoll *et al.*, *Neuron* **102**, 735 (2019).
15. E. Suarez *et al.*, in *Contemporary High Performance Computing: From Petascale Toward Exascale*, Volume Three, J. S. Vetter, Ed. (CRC Press, 2019), pp. 223–251.