Computational neuroscience

https://doi.org/10.1038/s43588-024-00743-z

On the path toward brain-scale simulations

Felix Wang & James B. Aimone

Today's high-performance computing systems are nearing an ability to simulate the human brain at scale. This presents a new challenge: going forward, will the bigger challenge be the brain's size or its complexity?

How far are we from simulating the human brain? Ever since computers have been invented, researchers have wondered about instilling them with artificial intelligence. At the same time, full-scale brain simulations provide us with the potential to discover deep understanding in the fields of neuroscience and health[1] and are increasingly enabled by the availability of whole-organism connectomes (meaning, the map of neuron connectivity)[2]. Especially with the rapidly growing societal costs of mental health, the use of advanced simulations is a valuable tool to explore brain function and its dysfunction[3,4]. Nevertheless, full-scale brain simulations are a challenging task, and one of the main bottlenecks boils down to their exponential computational cost. Writing in *Nature Computational Science*, Wenlian Lu and colleagues present an approach that has pushed the limit on the scale and complexity of brain simulations that can be performed[5].

Beyond just its sheer size, simulating the human brain is a complex challenge due to its intricate structure and connectivity. Using a simple back-of-the-envelope estimate, the brain contains approximately 100 billion neurons, each capable of connecting with up to 10,000 other neurons. Conservative estimates place the number of connections between neurons in the brain — called synapses — at about 1 quadrillion, or 10^{15}. Reaching toward large-scale simulations, the key limiting factor lies in how to efficiently model these synapses. On modern parallel supercomputing systems, while it may be straightforward enough to compute the model updates for each neuron independently, the across-model communication costs imposed by the synapses quickly eclipse that of neurons, even before incorporating synaptic stochasticity (each synapse is an independent source of randomness) and plasticity (the presumed locus of learning in the brain).

Another substantial challenge in brain-scale simulation is the vast diversity of timescales involved within the model. Neuron dynamics for transmitting and processing spiking signals typically occur on the order of milliseconds, but the processes governing neural plasticity (that is, learning and adaptation), typically unfold over seconds, or over months in the case of neurogenesis or synaptic restructuring[6]. Similarly, the timescales of interest for simulations vary as well: while there are some neurological conditions, such as epilepsy, which exhibit aberrant dynamics observable in short time scales, many neurological conditions also arise over weeks and months[6]. For this reason, we ideally need to be able to simulate the brain at speeds much faster than real-time.

In their study, Lu and colleagues have tackled some of these challenges to increase the scale and complexity of brain simulations[5] (Fig. 1).

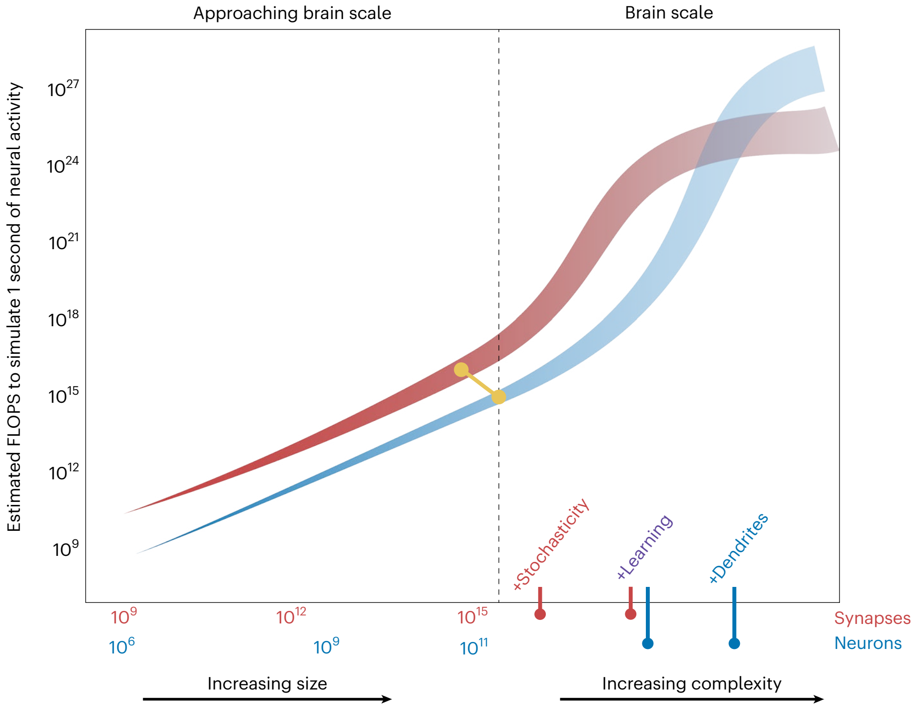

Fig. 1 | Estimation of computational costs of full-brain simulations. Lu and colleagues show that full-brain scales are achievable with today's high-performance computing (HPC) systems. Future brain simulations will push these limits by including stochastic synapses, continual learning, and anatomically realistic neurons, potentially greatly exceeding today's conventional HPC capabilities. The scale achieved by the work by Lu and colleagues is shown by a yellow marker. The dashed line represents the estimated scale of the human brain. The red curve represents the estimated cost of synapses, and the blue curve represents the estimated costs of simulating neurons. FLOPS, floating-point operations per second.

By utilizing over 14,000 graphics processing units (GPUs) over 3,500 compute nodes, they were able to simulate a full 86 billion neurons and 47.8 trillion synapses in an approximated brain model. While there have been other simulations that have reached ultra-large scales (>10^{13} synapses)[7,8], this is the first study that achieved this scale of simulation while both incorporating measured neurobiological constraints and targeting a functional cognitive task.

The use of GPUs is particularly promising because they are able to process necessary computations in parallel, which helps to accelerate these large-scale workloads[9]. However, the use of GPUs for biological simulations has historically been challenging due to the diverse types of neurons and the complex connectivity of the brain, which tends to be locally dense and globally sparse. When simulations get large enough to span multiple compute nodes, the additional complexity of having to shuffle data to and from GPUs as well as between compute nodes makes the design of simulation software much more difficult.

Although the authors used simplified, homogeneous neuron models in their work, they proposed solutions to address the critical communication problem. This was accomplished through a two-level scheme to reduce the total amount of communication that could take place between compute nodes. Within a local group of GPUs, they allowed for direct connections. Between groups of GPUs, they assigned a bridge node that was responsible for forwarding data. This substantially cut down on the total number of overlapping communication pathways in their multi-node system, thus reducing network congestion. Between a GPU and its compute node, they also separated the computation and communication into three main tasks — sending, computing, and receiving — which could be partially overlapped. This additionally improved efficiency by minimizing the time spent waiting on data transfer where possible.

As a result of these efforts, the authors were able to achieve impressive simulation times for a brain-scale model compared to similar simulation efforts. Using these simulations for computational neuroscience experiments, the authors were also able to fit their spiking neural simulation to data observed through functional MRI (magnetic resonance imaging), which is a method for measuring brain activity through blood oxygenation levels. These experiments compared the resting and active states of the brain, and the responses to a visual evaluation task. Although these experiments are only able to paint the brain in broad strokes, they provide an important neuroscience result in advancing how we may learn about the brain through simulation as an important computational milestone.

The methods developed by Lu and colleagues provide valuable technical insights for realizing efficient brain-scale simulations, as well as a sense of the considerable computational resource needs. It is notable that this simulation, which captures roughly one tenth of the synapses of the brain, used roughly one tenth the floating-point operations per second (FLOPS) of an exascale machine, suggesting that today's largest high-performance computing systems are in the ballpark of what is necessary for modeling every synapse. For this reason, we expect to see some growing tension between improving our models and our simulators. Increasing simulation fidelity to capture more biological detail and diversity will require refinements to the neuron, synapse, and whole-brain connectivity models, greatly increasing computational cost at a constant scale. We can only speculate as to how this complexity will increase computational costs, but it is possible that the inclusion of learning[10] and spatial considerations such as dendrites (the tree-like arborization of neurons)[11] will eventually make neurons, not synapses, the dominant cost. To reduce simulation costs accordingly, we may have to look into advancing computational techniques and supporting hardware systems that are less reliant on homogeneity, perhaps shifting toward neuromorphic technologies for acceleration over GPUs[12]. Returning to their applications in neuroscience and health, ideally these efforts would lead us to a more complete understanding of the brain by expanding the scope and detail of questions we may ask about the brain through simulation.

Felix Wang & James B. Aimone ✉
Neural Exploration and Research Laboratory, Sandia National Laboratories, Albuquerque, NM, USA.
✉e-mail: jbaimon@sandia.gov

Published online: 19 December 2024

References

1. Aimone, J. B. et al. *Front. Neuroinform.* https://doi.org/10.3389/fninf.2023.1157418 (2023).
2. Shiu, P. K. et al. *Nature* **634**, 210–219 (2024).
3. Jirsa, V. et al. *Lancet Neurol.* **22**, 443–454 (2023).
4. Sanz Leon, P. et al. *Front. Neuroinform.* https://doi.org/10.3389/fninf.2013.00010 (2013).
5. Lu, W. et al. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-024-00731-3 (2024).
6. Aimone, J. B. & Weick, J. P. *Front. Comput. Neurosci.* https://doi.org/10.3389/fncom.2013.00150 (2013).
7. Igarashi, J., Yamaura, H. & Yamazaki, T. *Front. Neuroinform.* https://doi.org/10.3389/fninf.2019.00071 (2019).
8. Ananthanarayanan, R., Esser, S. K., Simon, H. D., & Modha, D. S. The cat is out of the bag: cortical simulations with 10^9 neurons, 10^{13} synapses. In *Proc. Conf. on High Performance Computing Networking, Storage and Analysis*, 63 (ACM, 2009).
9. Knight, J. C. & Nowotny, T. *Front. Neurosci.* https://doi.org/10.3389/fnins.2018.00941 (2018).
10. Bono, J. & Clopath, C. *Nat. Commun.* **8**, 706 (2017).
11. Ben-Shalom, R. et al. *J. Neurosci. Methods* **366**, 109400 (2022).
12. Wang, F. et al. *Neuromorphic Comput. Eng.* **4**, 024002 (2024).

Acknowledgements

We thank L. Aimone for copyediting assistance. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the US Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the US Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

We acknowledge support from the DOE Office of Science ASCR Collaborative Research in Computational Neuroscience (CRCNS) and the DOE NNSA Advanced Simulation and Computing program.

Competing interests

The authors declare no competing interests.